Statistics and Data Analysis

for the Physicist

Chandra Prakash

lunar-photon.github.io

Preface

Who made these notes

These notes were made by Chandra Prakash (chandra.pp@alumni.iitg.ac.in; the latest version lives at lunar-photon.github.io) together with Claude, an AI model made by Anthropic. The goals, the vision and the way every idea is framed are Chandra's: which questions matter, which pictures make them clear, and how the whole subject hangs together for a physicist. The drafting of the text and the writing of the code were done with Claude's help, one section at a time, with Chandra reading, correcting and redirecting. It was a joint project and an ambitious one: to take a physicist who has never properly learned statistics all the way from coin tosses to Monte Carlo simulations of the microwave sky, Markov chain samplers, Fisher forecasts and the topology of random fields, and to have every step run as a Python script on the reader's own machine.

Goals and vision

What the notes are for. The reader we have in mind is a phenomenologist working in particle physics, cosmology or gravitational waves. You are at home with calculus, linear algebra, Fourier analysis, quantum mechanics and statistical mechanics. You use likelihoods, error bars and MCMC chains in your work, often through code written by someone else, and you would like to know what each tool really does and why it exists. You also want to *implement* it: to write the fifty lines of Python that do the job, see them work on data, and then trust the standard library because you know what it is doing inside.

The vision. Three convictions shape every chapter.

1. *Every idea opens with a story.* Before any definition we meet a concrete situation that makes the idea necessary, and a physical picture of what the idea will do for us. The formula comes after we know what question it answers.

2. *Probability runs in two directions.* Prediction goes forward: given a hypothesis, how often does each outcome occur? Inference goes backward: given the outcome we saw, which hypotheses could lie behind it, and how much does each deserve belief now? Bayes' theorem is the bridge. Almost every confusion in statistics is a confusion between these two directions.

3. *One pipeline organises the whole subject.* A physical model produces a field or a data set. We choose what information in it matters, compress the data into a summary S, work out how S fluctuates from one realisation of the noise to the next (its sampling distribution $p(S \mid \boldsymbol{\theta})$), and only then infer the physics (Figure P.1). The book walks along this pipeline, from probability distributions to Bayes, Monte Carlo, MCMC and the Fisher matrix, and

ends where the summaries become most interesting: topological summaries of random fields.

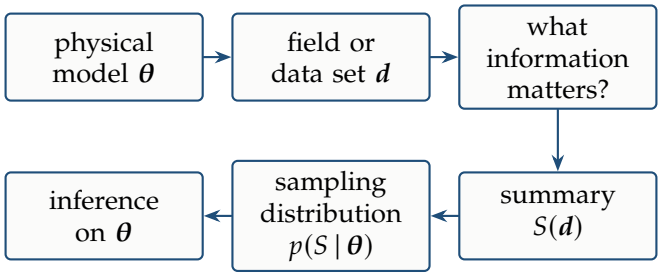


Figure P.1: The pipeline that organises the book. The top row is the physics and the observation; the bottom row is the statistics. Each chapter serves one or two of these steps, and Chapter 0 places every chapter on its step.

What you will be able to do at the end. You will be able to look at a physics measurement and say which of the two directions each part of the analysis runs in. You will be able to write down a likelihood for Poisson counts in a collider bin, for a Gaussian random field on the sphere and for a gravitational-wave detector with coloured noise, and explain why these are the same idea. You will be able to write a Monte Carlo simulation that checks an analytic result, and to recognise when the same simulation is no longer a check but the only instrument available. You will be able to code a Metropolis–Hastings sampler from scratch and diagnose it, compute a Fisher forecast and know when it can be trusted, and treat a topological summary of a field (its Betti numbers, its Euler characteristic, its persistence diagram) as one more random observable, with its own covariance and likelihood.

How the book is organised

Chapter 0 is the map: the two directions of probability, the list of questions the book answers, the pipeline, the menu of summaries, and how to run the code. Chapter T0 is the small part of Python that every script in the book uses, and how to tell whether a number a script prints can be trusted.

The first part builds the statistics. Chapter 1 and Chapter 2 build probability and the distributions a physicist meets. Chapter 3 turns data into estimates and introduces the likelihood; Chapter 4 asks whether a signal is really there and how to set a limit when it is not; Chapter 5 is Bayesian inference.

The second part collects the physics the statistics is used on, one toolkit per field, each a map of the objects and equations that the later chapters need: cosmology and weak lensing (Chapter T1), astrophysics and gravitational waves (Chapter T2), the morphology of the microwave sky (Chapter T3), and collider particle physics (Chapter T4). The last toolkit is statistical: Chapter T5 sets the estimators side by side, adds the robust, shrinkage, quadratic, iterative and recursive ones, and says which problem calls for which. They can be read in order or consulted when a chapter points to them.

The third part turns the statistics into instruments. Chapter 6 is Monte Carlo and the pseudo-random numbers behind it. Chapter 7 makes the summary concrete for fields ($\xi(r)$, $P(k)$, C_ℓ) and Chapter 8 is a Monte Carlo research problem on the microwave sky. Chapter 9 and Chapter 10 turn sampling distributions into parameter constraints with MCMC and the Fisher matrix, and Chapter 11 applies all of it to galaxy surveys and the local Universe.

The last part goes beyond two-point statistics and onto real data. Chapter 12 is topological data analysis of random fields, Chapter 13 runs four complete analyses on public data (the Planck sky, its topology, SDSS galaxies and a Higgs search), and Chapter 14 pulls the threads together: given a new problem, which idea to reach for. Every chapter opens with a roadmap and closes with a summary table.

Bayes' theorem in one picture, told with one box of each kind

The text of each chapter carries the whole argument, and it reads correctly from start to finish if every box is skipped. Boxes are a sparse sidebar: something to find again, or to skip on a first reading. Rather than describe them, we show one real box of each kind. All of them tell one story, the story that also opens Chapter 0: Bayes' theorem as information shrinking the sample space.

Story: the coin from the drawer

A drawer holds two coins that look identical. One is fair. The other is bent and lands heads 80% of the time. A friend picks one at random, tosses it ten times, and reports 8 heads. Which coin was it?

Stories give the concrete situation that makes an idea necessary. In the chapters they are ordinary prose, the first paragraph of a section.

The story contains both directions at once. If we knew which coin it was, we could predict how often it gives 8 heads. We do not know the coin; we know the 8 heads. So the question runs backward, from the outcome to the coin.

The central framing

A question about chance can face either way. Facing forward, we fix the situation and ask which outcomes it can give and how often each one turns up. Facing backward, the outcome is already in hand, and we ask which of the situations that might have been true could have given it, and how much each of them deserves to be believed now that we have seen it.
For the coin, prediction asks $\mathrm{P}(E \mid H)$: *given this coin, how likely was this outcome?* Inference asks $\mathrm{P}(H \mid E)$: *given this outcome, which way could it have happened, and how probable is each way?* The two numbers are different, and Bayes' theorem converts one into the other.

The central framing states the question a chapter answers.

To make the conversion we need one definition. We will look it up many times, so it is worth having in a box.

Working definition P.1 (conditional probability)

For events A and B with $\mathrm{P}(A) > 0$, the **conditional probability** of B given A is

$$\mathrm{P}(B \mid A) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(A)}.$$

It is the probability of B once we know that A happened [we keep only the outcomes in which A is true] and renormalise [divide by $\mathrm{P}(A)$, so the kept outcomes again have total probability 1]. Operationally: in a long run of repetitions, throw away every run in which A failed and count the fraction of the survivors in which B happened.

Working definitions keep the precise sentence and unpack every term a forgetful reader may need, in green brackets.

Now the reasoning, in words, before any number.

The reasoning

The backward question can always be said in four plain steps before any equation is written:

1. What did we actually see?
2. Which explanations could have given it?
3. If each explanation were true, how probable would that sight have been?
4. How do those answers change the weight each explanation had before?

For the coin: (1) E = "8 heads in 10 tosses"; (2) B (bent coin) and F (fair coin), each picked with probability $\frac{1}{2}$; (3) the forward question, answered by the binomial distribution; (4) keep only the paths that end in E, and share the probability among them in proportion to how much each contributes.

Reasoning boxes state a question in words before any symbol appears.

Step 3 is ordinary prediction: if this coin were the one, how often would it give 8 heads? For n tosses of a coin with heads probability p, the probability of exactly k heads is $\binom{n}{k}p^k(1-p)^{n-k}$. The factor $p^k(1-p)^{n-k}$ is the probability of one particular sequence with k heads, and $\binom{n}{k}$ counts how many such sequences there are. With $n = 10$ and $k = 8$, the fair coin gives $\mathrm{P}(E \mid F) = \binom{10}{8}/2^{10} = 45/1024 = 0.04395$ and the bent coin gives $\mathrm{P}(E \mid B) = 45 \times 0.8^8 \times 0.2^2 = 0.302$. Step 4 is Bayes' theorem, and it has a picture.

Physical picture: information shrinks the sample space

Before the evidence, the whole sample space is relevant. Draw it as a rectangle of area one and cut it into two strips, one for each coin, each as wide as its prior, $\mathrm{P}(B) = \mathrm{P}(F) = \frac{1}{2}$. Inside each strip, mark the part in which 8 heads happens. It fills the fraction $\mathrm{P}(E \mid B) = 0.302$ of the bent-coin strip and only $\mathrm{P}(E \mid F) = 0.04395$ of the fair-coin strip, so the two pieces have areas $\mathrm{P}(E \cap B)$ and $\mathrm{P}(E \cap F)$. Together they are the event E, one box straddling the line between the strips. The evidence tells us that we are somewhere inside that box, so everything outside it is ruled out: the sample space has shrunk to the box. The posterior $\mathrm{P}(B \mid E)$ is the share of the box that the bent-coin piece fills (Figure P.2).

Intuition boxes give the physical or visual picture.

The picture turns into a calculation with no new idea in it.

Worked example P.1: the posterior for the coin

With $\mathrm{P}(B) = \mathrm{P}(F) = \frac{1}{2}$,

$$\begin{aligned}\mathrm{P}(B \mid E) &\overset{(1)}{=} \frac{\mathrm{P}(E \cap B)}{\mathrm{P}(E \cap B) + \mathrm{P}(E \cap F)} \\ &\overset{(2)}{=} \frac{\mathrm{P}(E \mid B)\,\mathrm{P}(B)}{\mathrm{P}(E \mid B)\,\mathrm{P}(B) + \mathrm{P}(E \mid F)\,\mathrm{P}(F)} \\ &\overset{(3)}{=} \frac{\mathrm{P}(E \mid B)}{\mathrm{P}(E \mid B) + \mathrm{P}(E \mid F)} = \frac{0.302}{0.302 + 0.04395} = 0.873.\end{aligned}$$

(1) Conditional probability, $\mathrm{P}(B \mid E) = \mathrm{P}(E \cap B)/\mathrm{P}(E)$, with $\mathrm{P}(E)$ split into the two disjoint ways E can happen (the two green pieces of the box).

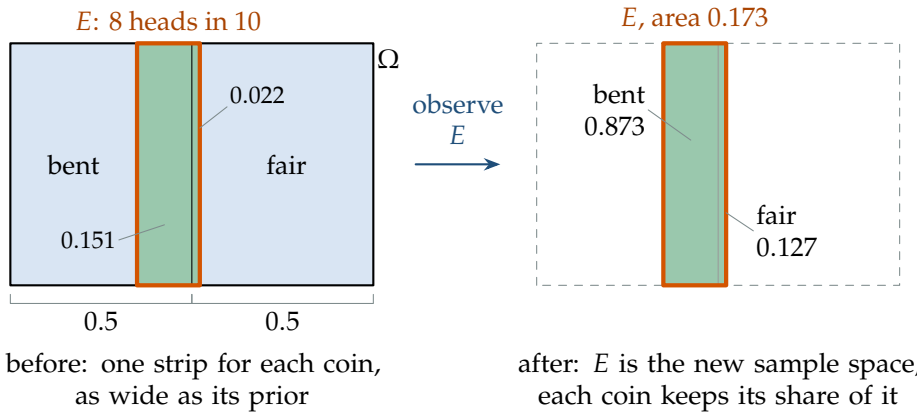


Figure P.2: Bayes' theorem as information shrinking the sample space, drawn to scale. Left: the rectangle Ω has area 1, and since the coin is either bent or fair, the two hypotheses fill it in two strips of width $\frac{1}{2}$. The event E, "8 heads in 10" (orange outline), covers the fraction 0.302 of the bent-coin strip, because that is how often the bent coin gives 8 heads, and only 0.04395 of the fair-coin strip. Each green piece therefore has area prior times likelihood. Right: hearing "8 heads" discards everything outside the orange box, and each coin keeps the share of the box that its piece fills: 0.873 for the bent coin.

(2) The product rule $\mathrm{P}(E \cap H) = \mathrm{P}(E \mid H)\,\mathrm{P}(H)$ for each piece: its share of the strip times the strip's width.
(3) The equal priors $\frac{1}{2}$ cancel between numerator and denominator.

The 8 heads moved the probability of the bent coin from 0.5 to 0.873.

Examples are solved to the last line. Every non-trivial equals sign is labelled and explained right after the chain.

One detail of the example is easy to overlook: the prior widths of the two strips cancelled only because they were equal. With unequal priors they stay in the answer.

Takeaway

In the picture, Bayes' theorem needs no extra assumption: it falls out of the picture. There is one patch, the overlap $H \cap E$, and two ways to measure it. As a share of the H strip it is $\mathrm{P}(E \mid H)$; as a share of the box E it is $\mathrm{P}(H \mid E)$. Bayes' theorem is the statement that both shares describe the same area, $\mathrm{P}(E \mid H)\,\mathrm{P}(H) = \mathrm{P}(H \mid E)\,\mathrm{P}(E)$.
For the coin, with $H = B$: a forward probability is the *share of a strip* that the evidence covers; a joint probability is an *area*; the posterior is a *ratio of areas* once the evidence has cut the sample space down to the box. Move the line between the strips and both pieces change width, so the same data give a different posterior.

Takeaways say in a sentence or two what a computation taught us.

The most common error in reading such results is to swap the two directions, and it has a famous statistical form.

Watch out: a p-value is not the probability that the null is true

Suppose we test "the coin is fair" by asking how often a fair coin gives 8 or more heads in 10 tosses. That number, $(45 + 10 + 1)/1024 \approx 0.055$, is a forward probability, $\mathrm{P}(\text{this or more extreme} \mid F)$. It is not the probability that the coin is fair, $\mathrm{P}(F \mid E)$. That we just found to be $1 - 0.873$ with equal priors, and it depends on the prior as well. Chapter 4 makes this precise.

Pitfalls flag a genuine trap.

Every step of the story is a few lines of code. The script behind the numbers above computes the forward probabilities from the binomial formula, checks them by simulating 20000 ten-toss experiments with each coin, and then runs Bayes on the observed outcome. The backward step is four lines.

Listing P.1: The backward direction: keep the paths that end at the observed k, then normalise (`code/ch00/01_coin_two_directions.py`)

```
42 # ---- backward direction: Bayes on the observed column of the tree ----
43 num_bias = pmf_bias * PRIOR_BIAS                  # path "biased, then k heads"
44 num_fair = pmf_fair * (1 - PRIOR_BIAS)            # path "fair, then k heads"
45 post_bias = num_bias / (num_bias + num_fair)      # normalise over the paths that end at k
```

Line 43 is the area of the bent-coin piece of the box in Figure P.2, line 44 the area of the fair-coin piece, and line 45 divides the first by their sum. The arrays run over every possible k at once, so the script produces the posterior for all eleven outcomes in one go. For $k = 8$ it prints 0.873; for $k = 5$ it gives 0.09696. Every number quoted in the text is written to a file by the script that computed it and read back into LaTeX. So the text and the code cannot drift apart (Chapter 0 describes the machinery).

The simulation in the same script is a small instance of an idea that runs through the book: a Monte Carlo experiment can check a formula, or it can replace one.

How this is used in real research: verification versus research

Here the simulation *verifies* a known answer. The fraction of simulated fair-coin experiments with 8 heads is 0.04105, against the exact 0.04395. The difference is about two Monte Carlo standard errors (±0.001449 each), an ordinary fluctuation that shrinks as $1/\sqrt{N}$ when more experiments N are simulated. In research the same machinery is often the only instrument. The covariance of a power-spectrum estimator on a masked, noisy microwave sky has no closed form, and thousands of simulated skies are how it is measured (Chapter 8).

These boxes separate a numerical check of known theory from the same tool used where no analytic answer exists.

When an idea has several tools, a small table says which to use when.

Taxonomy: the coin, asked four ways

question in words	object	chapter
how often does a fair coin give 8 heads?	$P(k \mid p)$	Chapter 2
what is our best guess of p?	estimator $\hat{p} = k/n$	Chapter 3
is 8 heads surprising for a fair coin?	p-value	Chapter 4
how probable is each value of p?	posterior $p(p \mid k)$	Chapter 5

Taxonomies list which tool answers which question.

The picture is not ours alone; the literature draws it in several ways.

Reading the source (Wasserman p. ix, Fig. 1)

Wasserman opens with one diagram: a data generating process and the observed data, joined by an arrow labelled "probability" going forward and one labelled "inference" coming back [Wasserman p. ix]. Kruschke describes the backward arrow as a reallocation of credibility across possibilities [Kruschke pp. 16–22]. In the chapters, such links are short citations at the point of use, with one "Reading the literature" list per chapter.

Source boxes map the notes onto the books, and flag their misprints.

The labelled-equality convention

Every calculation is written as a chain. Each non-trivial equals sign carries a small label, and the labels are explained, in order, right after the chain. For the odds form of the coin problem:

$$\frac{\mathrm{P}(B \mid E)}{\mathrm{P}(F \mid E)} \overset{(1)}{=} \frac{\mathrm{P}(E \mid B)\ \mathrm{P}(B)/\mathrm{P}(E)}{\mathrm{P}(E \mid F)\ \mathrm{P}(F)/\mathrm{P}(E)} \overset{(2)}{=} \frac{\mathrm{P}(E \mid B)}{\mathrm{P}(E \mid F)} \cdot \frac{\mathrm{P}(B)}{\mathrm{P}(F)} = 6.872 \times 1.$$

(1) Bayes' theorem for each coin.
(2) $\mathrm{P}(E)$ cancels. Posterior odds are the likelihood ratio times the prior odds.

Odds of 6.872 to 1 are a probability $6.872/(1 + 6.872) = 0.873$, as before.

The code

Every figure made from data and every simulated number comes from a short script in `code/chNN/`, runnable from the Notes directory with `python3 code/chNN/NN_name.py`. Randomness is seeded from the script's name, so rerunning a script reproduces its figures bit for bit. Each method is first implemented from scratch, so the mechanism is visible, and then checked against a standard library (`emcee`, `getdist`, `healpy`, `gudhi` and others). Chapter 0 ends with a tour of the layout and the shared helpers.

Sources

- L. Wasserman, *All of Statistics* (Springer 2004), followed chapter by chapter and cited as [Wasserman p. …] with printed page numbers.
- J. K. Kruschke, *Doing Bayesian Data Analysis* (2nd ed.), the model of clarity, cited as [Kruschke …].
- G. Cowan, *Statistical Data Analysis*; D. S. Sivia, *Data Analysis: A Bayesian Tutorial*; G. Casella and R. L. Berger, *Statistical Inference*; B. Lambert, *A Student's Guide to Bayesian Statistics*; M. P. Hobson et al., *Bayesian Methods in Cosmology*; and *Seeing Theory*, each cited by name where it is clearest.
- Research papers on cosmological inference, pseudo-C_ℓ estimation, gravitational-wave forecasts and the topology of random fields, cited in the usual way and collected in the bibliography.

Recap: the dictionary of this preface

prediction	$\mathrm{P}(E \mid H)$, the share of the H strip covered by E
joint probability	$\mathrm{P}(E \cap H) = \mathrm{P}(E \mid H)\,\mathrm{P}(H)$, an area
evidence	$\mathrm{P}(E)$, the total surviving area
inference	$\mathrm{P}(H \mid E)$, a ratio of surviving areas
odds form	posterior odds = likelihood ratio × prior odds
the pipeline	model → data → summary → $p(S \mid \boldsymbol{\theta})$ → inference

A recap closes every part.

Contents

Preface ii

I Probability and Inference 1

0 The Map: Two Directions of Probability and the Pipeline from Data to Physics 2

Where this chapter is going . . . 2

0.1 Two directions of probability . . . 2

0.2 The questions the book answers . . . 9

0.3 The pipeline: from a physical model to a statement about physics . . . 14

0.4 Which summary? From amplitude to topology . . . 20

0.5 One idea, several detectors . . . 24

0.6 How to run the code . . . 29

Chapter summary . . . 35

T0 Python for This Book 36

Where this chapter is going . . . 36

T0.1 How much Python does the book use? A census . . . 36

T0.2 Running a script and reading what it prints . . . 40

T0.3 Plain Python: values, containers, loops and functions . . . 45

T0.4 NumPy: computing with whole arrays . . . 54

T0.5 Random numbers, linear algebra and Fourier transforms . . . 61

T0.6 SciPy: distributions, integrals, roots and minima . . . 66

T0.7 Matplotlib, files and folders . . . 70

T0.8 Seen a few times: rare features and two black boxes . . . 74

T0.9 Seven traps that catch everyone once . . . 78

T0.10 Is this number real? Numerical noise and how to check . . . 83

T0.11 When sixteen digits are not enough: mpmath . . . 96

T0.12 How long will it run? Counting operations . . . 101

T0.13 The whole toolkit on one page . . . 106

Chapter summary . . . 108

1 Probability: The Language of Uncertain Events 110

1.1 Sample spaces and events: probability needs a list of possibilities 111
1.2 The axioms of probability, and what they buy us 118
1.3 Equally likely outcomes: probability as counting 127
1.4 Conditional probability: shrinking the sample space 133
1.5 Probability trees and the law of total probability 141
1.6 Independence: when learning B tells us nothing about A 145
1.7 Bayes' theorem: reading the tree backwards 151
1.8 Random variables: turning outcomes into numbers 165
1.9 Mass, density and the cumulative distribution 170
1.10 Several variables at once: joint, marginal and conditional distributions 178
1.11 Functions of random variables: change of variables and Jacobians 188
1.12 Expectation and variance: the long-run average and the spread 202
1.13 Covariance and correlation: how two variables move together 213
1.14 Moments, generating functions and characteristic functions 222
Chapter summary 230

2 Probability Distributions from First Principles 233
Where this chapter is going 233
2.1 One trial: the Bernoulli distribution 234
2.2 Many successive trials: the binomial distribution 238
2.3 Histogram bins: the multinomial distribution 245
2.4 Rare events in many trials: the Poisson distribution 249
2.5 The Poisson process: constant rate, no memory 254
2.6 Waiting times: the exponential and gamma distributions 262
2.7 Discrete waiting times and extra spread: geometric and negative binomial 268
2.8 When the Poisson assumptions fail, and where the Poisson goes for large λ 274
2.9 The Gaussian, first way: the shape of a binomial with many trials 285
2.10 The Gaussian, second way: the central limit theorem 293
2.11 The Gaussian, third way: the least committal assignment 305
2.12 Many correlated Gaussians: the multivariate normal 314
2.13 Sums of squared Gaussians: the χ^2 distribution 320
2.14 When the error bar is itself estimated: Student's t and the F ratio 328
2.15 The rest of the zoo, each with the machine that makes it 334
2.16 Which distribution when 344
Chapter summary 351

3 Estimators, Error Propagation and the Likelihood 354
3.1 One number from many readings: estimators and their sampling distribution . 356
3.2 Aiming and scattering: bias, variance and mean squared error 363
3.3 Why error bars shrink like $1/\sqrt{N}$: the law of large numbers and the central limit theorem 371

3.4 Estimating the scatter itself: the sample variance and its error . . . 378
3.5 Estimating covariance and correlation matrices, and how noisy they are . . . 384
3.6 Propagating errors through a formula . . . 391
3.7 Combining measurements of different quality: the weighted mean . . . 398
3.8 The likelihood: data are the model plus noise . . . 411
3.9 Maximum likelihood: pick the value that needs the least unusual noise . . . 419
3.10 Gaussian noise: least squares and the χ^2 . . . 424
3.11 Counting experiments: the Poisson likelihood of a histogram . . . 431
3.12 The width of the likelihood is the error bar . . . 437
3.13 What makes maximum likelihood good . . . 443
3.14 Nuisance parameters and the profile likelihood . . . 448
3.15 The bootstrap: the sampling distribution from one data set . . . 453
3.16 From a model to an estimator: how to find the right one . . . 458
3.17 A worked problem: finding the estimator of the Hubble constant . . . 468
Chapter summary . . . 484

4 Hypothesis Testing: Asking the Null **487**
4.1 A bump in a histogram: new particle or luck? . . . 487
4.2 Which way is "more extreme"? Sides, and the stopping rule . . . 493
4.3 Two ways to be wrong: size, power and how much data we need . . . 498
4.4 The p-value is a random variable . . . 503
4.5 Three workhorse tests: Wald, Student's t and permutation . . . 508
4.6 The best test: the Neyman–Pearson lemma . . . 512
4.7 Likelihood-ratio tests and Wilks' theorem . . . 517
4.8 Goodness of fit: Pearson's χ^2 and Kolmogorov–Smirnov . . . 523
4.9 From p to $Z\sigma$, and why a discovery needs 5σ . . . 527
4.10 From a test to an interval: the values we cannot reject . . . 531
4.11 The Neyman construction: a belt of acceptance regions . . . 538
4.12 Coverage is a property of the procedure . . . 548
4.13 Upper limits near a physical boundary . . . 553
4.14 Many tests at once and the look-elsewhere effect . . . 560
4.15 A search as a likelihood: templates and nuisance parameters . . . 568
4.16 Discovery and exclusion: q_0, $\tilde{q}_\mu$, CL_s, Brazil band . . . 575
4.17 Confidence intervals and credible intervals . . . 583
Chapter summary . . . 589

5 Bayes' Theorem and Bayesian Inference **591**
5.1 One joint table, read in two directions . . . 591
5.2 Information shrinks the sample space . . . 597
5.3 Reallocating credibility across possibilities . . . 604
5.4 Odds and the likelihood ratio: evidence as a multiplier . . . 610

5.5 Sequential updating: yesterday's posterior is today's prior . . . 614
5.6 The denominator: the sum over all paths . . . 623
5.7 From a list of hypotheses to a continuous parameter . . . 627
5.8 Conjugate priors: posteriors on the back of an envelope . . . 633
5.9 Summarising a posterior: best values and credible intervals . . . 639
5.10 The Gaussian approximation: a peak and its curvature . . . 643
5.11 Nuisance parameters: integrating out what we do not care about . . . 647
5.12 Choosing a prior: subjective, flat and Jeffreys . . . 652
5.13 Forgetting the prior: when the data win and when they cannot . . . 659
5.14 Sampling the posterior: a histogram instead of an integral . . . 664
5.15 Which model produced the data? . . . 670
5.16 Occam's razor from the evidence . . . 675
5.17 Reading a Bayes factor . . . 682
5.18 Nested models, shortcuts and model averaging . . . 689
5.19 Why we need sampling . . . 695
Chapter summary . . . 704

II Physics Toolkits **706**

T1 Cosmology for the Data Analyst **707**
Where this chapter is going . . . 707
T1.1 Six numbers in, statistics out . . . 708
T1.2 The objects the statistics uses . . . 709
T1.3 Experiment 1: how each parameter moves the CMB spectra . . . 714
T1.4 Experiment 2: the matter power spectrum and the BAO bump . . . 717
T1.5 Weak lensing: the matter field written on galaxy shapes . . . 719
T1.6 The objects a weak-lensing analysis uses . . . 723
T1.7 Experiment 1: what cosmic shear measures, S_8 . . . 731
T1.8 Experiment 2: signal and shape noise in a Stage-III survey . . . 737
T1.9 Experiment 3: the two biases of a reconstructed lensing map . . . 741
T1.10 Polarization: sticks on the sky, and two kinds of pattern . . . 750
T1.11 When power moves from E to B . . . 760
T1.12 Where B modes come from . . . 767
T1.13 Cosmic birefringence, the instrument and the Galaxy . . . 773
Chapter summary . . . 786

T2 Astrophysics for the Data Analyst **788**
Where this chapter is going . . . 788
T2.1 What a gravitational wave is . . . 789
T2.2 Why radiation starts at the quadrupole . . . 795

T2.3 How a detector turns a wave into one number . . . 801
T2.4 Why a compact binary chirps . . . 808
T2.5 The chirp in the frequency domain . . . 813
T2.6 Corrections to the phase: post-Newtonian orders and the ppE form . . . 818
T2.7 What LISA measures, and what hides the signal . . . 823
T2.8 The symbols, where they come from, and where they are used . . . 829
T2.9 Dark matter around black holes . . . 831
T2.10 Environmental forces and the power law in the phase . . . 839
T2.11 Type Ia supernovae: candles that need a correction . . . 851
T2.12 The distance ladder: why H_0 needs an anchor . . . 855
T2.13 Selection effects: what a flux limit does to a sample . . . 863
T2.14 Galaxies as biased tracers in redshift space . . . 868
T2.15 Baryon acoustic oscillations: a standard ruler . . . 875
T2.16 Photometric redshifts and their error distributions . . . 883
Chapter summary . . . 888

T3 CMB Morphology and Topology **891**
Where this chapter is going . . . 891
T3.1 The shape of a temperature map . . . 892
T3.2 What a skewed sky and a lensed sky change . . . 897
T3.3 Polarization patterns and their singularities . . . 901
T3.4 Lensing moves the pattern but keeps its topology . . . 906
T3.5 What *Planck* found . . . 909
T3.6 The topology of the cosmic web . . . 912
Chapter summary . . . 914

T4 Particle Physics for the Data Analyst **917**
Where this chapter is going . . . 917
T4.1 What a collider does, and what an event is . . . 918
T4.2 Cross section, luminosity and the expected number of events . . . 923
T4.3 Invariant mass: why a decaying particle makes a peak . . . 928
T4.4 Signal and background: where the templates come from . . . 935
T4.5 Systematic uncertainties: nuisance parameters and their calibrations . . . 941
T4.6 The Higgs boson search of 2012 . . . 946
T4.7 The open data: what is in the files . . . 951
T4.8 The symbols of a collider search . . . 955
Chapter summary . . . 957

T5 Choosing an Estimator **959**
T5.1 The questions that choose an estimator . . . 959
T5.2 Linear estimators and the Gauss–Markov theorem . . . 965

T5.3 Bayes estimators under repeated experiments . . . 972
T5.4 Many parameters at once: shrinkage and Stein's paradox . . . 978
T5.5 Wild readings: robust estimators . . . 983
T5.6 When the parameter lives in the covariance: quadratic estimators . . . 990
T5.7 Iteration: solving the equation that defines an estimator . . . 997
T5.8 Estimators that correct their own bias . . . 1014
T5.9 Estimators that update with every reading . . . 1021
T5.10A field guide: which estimator, when . . . 1027
Chapter summary . . . 1033
Reading the literature . . . 1034

III Simulation, Fields and Forecasts 1035

6 Monte Carlo: Drawing Samples on a Deterministic Machine 1036
6.1 Random numbers from a machine that cannot be random . . . 1036
6.2 Linear congruential generators . . . 1042
6.3 Random numbers fall mainly in the planes . . . 1047
6.4 Modern generators: a bigger state and a better output function . . . 1052
6.5 Seeds, streams and reproducible parallel simulations . . . 1057
6.6 Testing a generator: randomness as a null hypothesis . . . 1061
6.7 When randomness is not the point: quasi-random numbers . . . 1065
6.8 Uniform numbers into any distribution: the inverse CDF . . . 1070
6.9 Gaussian numbers: Box–Muller and the Cholesky factor . . . 1075
6.10 Rejection sampling: throwing points under a curve . . . 1080
6.11 Monte Carlo integration and the $1/\sqrt{N}$ law . . . 1085
6.12 Importance sampling, and why g needs thicker tails than f . . . 1091
6.13 A histogram of samples approximates the density . . . 1097
6.14 Unfolding: undoing the detector with a response matrix from simulation . . . 1100
6.15 Shrinking σ: variance reduction . . . 1113
6.16 Which sampler when . . . 1117
Chapter summary . . . 1118

7 Random Fields and the Angular Power Spectrum 1121
Where this chapter is going . . . 1121
7.1 A theory predicts the statistics of a field . . . 1121
7.2 Gaussian random fields: when two points tell the whole story . . . 1126
7.3 Fourier modes and the power spectrum . . . 1130
7.4 Simulating a Gaussian random field with the FFT . . . 1135
7.5 One universe: ergodicity and sample variance . . . 1140
7.6 Smoothing: beams, pixels and window functions . . . 1144

7.7 What the power spectrum throws away . . . 1150
7.8 The Fourier modes of the sphere . . . 1153
7.9 No preferred direction: why the $a_{\ell m}$ are uncorrelated . . . 1158
7.10 One sky, $2\ell + 1$ numbers: the estimator $\widehat{C}_\ell$ and cosmic variance . . . 1166
7.11 Where C_ℓ and the skies come from: CAMB and HEALPix . . . 1170
7.12 Cosmic variance by Monte Carlo . . . 1178
7.13 What C_ℓ does not see . . . 1184
Chapter summary . . . 1187

8 Monte Carlo as a Research Instrument: Observing the CMB 1189
8.1 The research problem . . . 1189
8.2 What the instrument does to the sky . . . 1193
8.3 The noise bias and how to remove it . . . 1199
8.4 How much the estimate scatters . . . 1203
8.5 The Monte Carlo machine . . . 1209
8.6 What the simulations show . . . 1214
8.7 Reproducing the literature . . . 1220
8.8 How it is done in research . . . 1232
8.9 A sky with holes in it . . . 1235
8.10 The mode-coupling matrix . . . 1243
8.11 Undoing the mask . . . 1250
8.12 The scatter of a cut-sky spectrum . . . 1258
8.13 How many simulations? . . . 1268
8.14 Reproducing the cut-sky literature . . . 1272
8.15 How it is done in research . . . 1280
Chapter summary . . . 1289

9 Markov Chains and MCMC: The Biased Random Walker 1291
9.1 When the motion samples for us: the oscillator . . . 1292
9.2 A walker with a memory of one step . . . 1297
9.3 Who can reach whom: classifying the states . . . 1306
9.4 Equilibrium: stationary distributions and detailed balance . . . 1313
9.5 Forgetting where you started . . . 1319
9.6 One long walk is a sample: the ergodic theorem . . . 1326
9.7 The walker that never settles: free diffusion . . . 1333
9.8 From the drunkard to the biased walker . . . 1339
9.9 Why the rule is $\min(1, R)$: detailed balance . . . 1343
9.10 The sampler in a dozen lines, and two textbook runs . . . 1350
9.11 How big should the steps be? . . . 1355
9.12 Reading a chain: burn-in, autocorrelation and effective sample size . . . 1362
9.13 Many chains from far apart: the Gelman–Rubin statistic . . . 1367

9.14 Forgetting the start is not forgetting the prior . . . 1371
9.15 Back to statistical mechanics: sampling the canonical ensemble . . . 1375
9.16 Relatives of the biased walker: Gibbs, Hamiltonian and ensemble samplers . . . 1382
9.17 A straight line: the chain against the exact answer . . . 1389
9.18 A decaying source: curved posteriors and nuisance parameters . . . 1397
9.19 The likelihood of a CMB power spectrum . . . 1403
9.20 A model spectrum fast enough for a chain . . . 1413
9.21 Cosmological parameters from one simulated sky . . . 1418
9.22 Reproducing Planck 2018: the amplitude and tilt of the primordial spectrum . . . 1423
9.23 How it is done in research . . . 1430
Chapter summary . . . 1437

10 The Fisher Matrix: Forecasting What Data Can Teach 1440
10.1 How sharp will the likelihood be? The Fisher matrix . . . 1441
10.2 Nobody beats the inverse Fisher matrix: the Cramér–Rao bound . . . 1447
10.3 Gaussian data: the Fisher matrix in closed form . . . 1453
10.4 Reading a Fisher matrix . . . 1457
10.5 Derivatives from a code that cannot differentiate . . . 1465
10.6 A CMB forecast for ΛCDM . . . 1469
10.7 How it is done in research . . . 1484
10.8 Noise alone, or a signal as well? . . . 1488
10.9 The Fisher matrix of a chirp . . . 1502
10.10 Environments as power laws in the phase . . . 1508
10.11 One forecast for every environment . . . 1511
10.12 Reproducing the published forecasts . . . 1517
10.13 When the Fisher ellipse lies . . . 1526
10.14 How it is done in research . . . 1531
10.15 How small a twist can the sky reveal? . . . 1534
10.16 Two angles and one sum . . . 1541
10.17 Simulating the measurement . . . 1546
10.18 How a birefringence measurement is made in research . . . 1552
Chapter summary . . . 1554

11 Large-Scale Structure as Data 1557
Where this chapter is going . . . 1557
11.1 Galaxies as a random sample of a smooth density field . . . 1558
11.2 Estimating the correlation function from pair counts . . . 1566
11.3 The power spectrum of a survey: the FKP estimator . . . 1574
11.4 The acoustic peak: a ruler in the galaxy distribution . . . 1579
11.5 Galaxy motions: clustering in redshift space . . . 1588
11.6 How surveys do it, and what our shortcuts cost . . . 1599

11.7 Two universes with one power spectrum . . . 1603
11.8 How many haloes of each mass . . . 1607
11.9 Inside and around a halo . . . 1615
11.10 The power spectrum built from haloes . . . 1620
11.11 Voids . . . 1625
11.12 The average void: profile, outflow and shape . . . 1632
11.13 Haloes and voids in research . . . 1639
11.14 Hubble's diagram, refitted . . . 1642
11.15 Peculiar velocities: what moves the galaxies . . . 1648
11.16 The local flow: falling towards Virgo . . . 1653
11.17 How well can a local measurement know H_0? . . . 1658
11.18 Where is a galaxy whose distance we measured? Malmquist bias . . . 1664
11.19 Cosmic chronometers: the expansion rate from the ages of galaxies . . . 1668
11.20 One constant, many routes . . . 1674
11.21 How the local Universe is measured in research . . . 1679
Chapter summary . . . 1687

IV Beyond Two-Point Statistics, and Real Data 1690

12 Topological Data Analysis: The Shape of a Field as Data 1691
12.1 What survives a nuisance transformation? . . . 1692
12.2 The topology of a field: excursion sets and what we count in them . . . 1695
12.3 Scanning the threshold . . . 1700
12.4 Area, boundary length and Euler characteristic . . . 1704
12.5 The expected Euler characteristic of a Gaussian field . . . 1707
12.6 Testing the formula on simulated maps . . . 1713
12.7 Reproducing the classic results . . . 1716
12.8 How it is done in research . . . 1725
12.9 One threshold is not enough: births, deaths and persistence . . . 1728
12.10 Small changes of the field, small changes of the diagram . . . 1734
12.11 Computing persistence: union–find and the elder rule . . . 1737
12.12 Point clouds: shapes grown from points . . . 1742
12.13 The cosmic web: clusters, tunnels and voids . . . 1746
12.14 Polarization singularities: a charge that a remapping cannot change . . . 1749
12.15 Reproducing the literature . . . 1753
12.16 How it is done in research . . . 1761
12.17 Weak-lensing maps to practise on: Gaussian and lognormal convergence . . . 1763
12.18 From galaxy shapes to a mass map . . . 1768
12.19 The two-point analysis: band powers, their errors, and the banana . . . 1776
12.20 Beyond two points: peaks, shapes and persistence of the mass map . . . 1782

12.21How it is done in research, and what our simplifications cost 1794
12.22A topological summary is a random number 1797
12.23Two fields with one power spectrum 1801
12.24The sampling distribution by simulation 1806
12.25Is this map Gaussian? A test with a simulated null 1811
12.26How well does one map measure ϵ? A Fisher forecast 1815
12.27From one map to a posterior for ϵ 1818
12.28The same experiment on the CMB sky 1821
12.29When there is no likelihood: inference from the simulator alone 1827
12.30Reproducing the literature: topology and f_{NL} 1830
12.31How it is done in research 1835
Chapter summary 1839

13 Capstone: The Analysis on Real Data **1841**
13.1 One sky, and the questions we put to it 1841
13.2 The mask: which sky to trust, and how to let go of the rest 1846
13.3 From the map to bandpowers, one correction at a time 1850
13.4 How much would the bandpowers scatter? Simulated skies 1854
13.5 Our spectrum against *Planck*'s 1859
13.6 Is anything else in the map? The checks 1862
13.7 Two primordial numbers from the real sky 1865
13.8 The full analysis, and what the laptop version gives up 1869
13.9 The shape of the one sky we have 1871
13.10From the archive to a working map 1875
13.11The area, boundary and Euler characteristic of the real sky 1879
13.12Thirty-six numbers, one verdict 1881
13.13Pieces, holes and their lifetimes on the real sky 1886
13.14Side by side with *Planck* 2018 1889
13.15A real galaxy catalogue: CMASS in the southern sky 1893
13.16What the survey could not see: the weights 1897
13.17Counting pairs: the correlation function of CMASS South 1902
13.18Error bars from the data alone: the jackknife 1905
13.19The acoustic peak of CMASS South, and the BOSS distance 1908
13.20The survey window, and what the full analysis adds 1915
13.21A Higgs search on collider open data 1919
13.22Shapes: the Higgs line and the background function 1923
13.23Is there a peak? The discovery test 1928
13.24Anywhere in the range: checking the formulas and looking elsewhere 1934
13.25How strong, and where? Measuring μ and m_H 1937
13.26Where can a Higgs boson be excluded? 1940

13.27Reproducing the 2012 diphoton result 1943
13.28The full analysis, and what our shortcuts cost 1948
13.29A small twist in the polarization of the sky 1949
13.30Polarization spectra through a mask 1954
13.31The estimator, tested on skies where we know the angle 1958
13.32What the SMICA sky says 1963
13.33The published angle, and what one map can measure 1966
13.34The B modes that must be there: lensing 1969
13.35The full analysis, and what the laptop version gives up 1972
Reading the literature 1974
Chapter summary 1974

14 Synthesis: Which Idea, When **1978**
Where this chapter is going 1978
14.1 The question picks the object 1978
14.2 Which distribution, which method 1982
14.3 The pipeline, and who built each arrow 1987
14.4 Four instruments, one likelihood 1990
14.5 The traps that kept coming back 1994
14.6 The words, and the routes through them 2001
Chapter summary 2006

Bibliography **2008**

Part I

Probability and Inference

Chapter 0

The Map: Two Directions of Probability and the Pipeline from Data to Physics

Where this chapter is going

Suppose we know the bias of a coin. We can predict how often a run of tosses will produce a given number of heads. Once we observe the count, we can ask what it tells us about the bias. Bayes' theorem connects these two directions of reasoning. In a physics experiment we also have to decide which features of the data carry information about the quantity we want to measure. We capture those features in a summary and work out how much it would fluctuate if we repeated the experiment. That sequence, from a physical model through a measurement and back to the physics, organises the book. We introduce it here with small calculations, then describe how to find and run the Python code that accompanies each chapter.

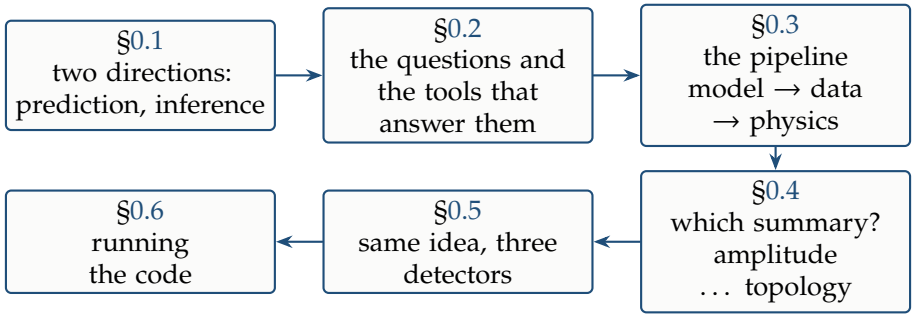


Figure 0.1: The route through the ideas. The top row sets up the logic: the two directions of probability, the list of questions they generate, and the pipeline that strings the questions together. The bottom row makes it concrete: what we choose to measure from a field, how the same logic looks in three different experiments, and how to reproduce every number in the book.

0.1 Two directions of probability

A friend runs a magic shop. In one drawer she keeps coins of two kinds, mixed together in equal numbers: honest coins, and trick coins weighted so that they land heads 80% of the time. The two kinds look identical. She takes one coin out at random and we toss it ten times. It lands heads 8 times.

Two different questions can be asked about this coin, and they must not be confused. The first could be asked *before* the tosses: "if this coin is honest, how likely is it to give 8 heads in 10 tosses?" The second is the one we actually want *after* the tosses: "we have seen 8 heads; which kind of coin is it, and how sure can we be?" Each question asks for a conditional probability,

and Bayes' theorem turns an answer to the first into an answer to the second. Every later calculation in the book, from a p-value to an MCMC chain over six cosmological parameters, answers one of these two questions.

Forward and inverse problems. Physics already knows this distinction. Given a Hamiltonian and an initial state, working out where the particle goes is the *forward* problem, and it has one answer. Given the hits in a tracking detector, working out which track, momentum and vertex produced them is the *inverse* problem. Several tracks could have produced similar hits, so we must weigh them against each other. Probability in the forward direction is like integrating the equations of motion. Probability in the backward direction, called inference, is like track reconstruction: we hold the outcome and reason back to its cause.

Prediction versus inference. Probability answers two kinds of question.

Prediction	**Inference**
Start with a hypothesis H.	Start with an observed event E.
Ask what outcomes it can produce, and how often.	Ask which hypotheses might lie behind it, and how probable each is now.
Use $\mathrm{P}(E \mid H)$.	Want $\mathrm{P}(H \mid E)$.
Forward direction: $H \longrightarrow E$.	Reverse direction: $E \longrightarrow H$.

These are not two rival interpretations of probability. They are two different questions asked of the same probabilities: the probabilities of every combination of hypothesis and outcome.

In the backward direction, ask two questions in words before writing any symbol. Which hypotheses could have produced what we saw? How probable was what we saw under each of them? Chapter 5 makes this a working habit.

A **probability tree** shows both directions at once. The first set of branches chooses the hypothesis (which coin). The second set chooses the outcome (8 heads or not). Each branch carries the probability of taking it, *given that we have reached the node it leaves from*. Multiplying the numbers along a path gives the probability of the whole path.

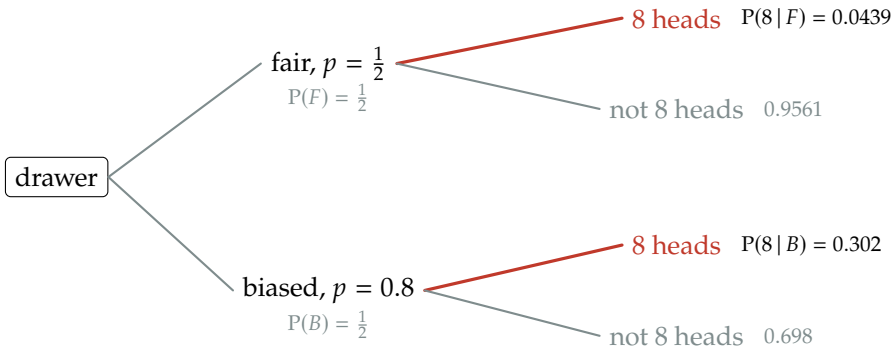


Figure 0.2: The coin as a probability tree. Read left to right, it is a prediction machine: pick a coin, then see what that coin tends to produce. Inference reads it the other way. We have observed "8 heads", so only the two red paths survive; the question is how the total probability of those two paths is split between the fair coin and the biased one.

The backward direction in four steps. Before any equation, put the question into words: (1) what did we see? (2) which hypotheses could have produced it? (3) how probable was it under

each of them? (4) how must the probabilities the hypotheses had beforehand change in view of it?

The event. What we observed is "8 heads in 10 tosses".

What could have produced it? The one thing we do not know is which kind of coin came out of the drawer. So there are two candidate explanations: the coin is fair (F), or the coin is biased (B).

How probable was the event under each? This is a forward calculation, done once per candidate: if the coin is fair, how often do ten tosses give 8 heads? If it is biased, how often? These two numbers are written $\mathrm{P}(8 \mid F)$ and $\mathrm{P}(8 \mid B)$, and Example 0.1 computes them.

Only then the update. Weight each path of the tree by the prior probability of its coin, keep only the paths that end at "8 heads", and rescale so that the surviving weights add up to one.

Step 3 is where conditional probability enters. For each hypothesis it gives the probability that the event we actually saw would have had, if that hypothesis were the one nature chose. To write that in symbols we need its definition.

Working definition 0.1 (conditional probability)

For two events A and B [sets of possible outcomes, e.g. A = "the coin is biased", B = "8 heads"] with $\mathrm{P}(A) > 0$, the **conditional probability** of B given A is

$$\mathrm{P}(B \mid A) := \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(A)}. \tag{0.1}$$

Operational meaning: throw away every outcome outside A, and measure the fraction of what is left that also lies in B. A becomes the new reference space [the "whole world" against which fractions are now measured]. The **prediction** question is a number $\mathrm{P}(E \mid H)$ [fraction of the H-world in which E happens]. The **inference** question is a number $\mathrm{P}(H \mid E)$ [fraction of the E-world in which H was true].

The same overlap $A \cap B$ can be divided by $\mathrm{P}(A)$ or by $\mathrm{P}(B)$. Multiply (0.1) through by its denominator; doing this with A and B in both orders gives the **product rule**

$$\mathrm{P}(A \cap B) = \mathrm{P}(B \mid A)\ \mathrm{P}(A) = \mathrm{P}(A \mid B)\ \mathrm{P}(B). \tag{0.2}$$

This one line already contains both directions. Think of probability as area. $\mathrm{P}(A \cap B)$ is the area of the overlap. The two conditional probabilities divide that *same* overlap by two *different* areas, $\mathrm{P}(A)$ or $\mathrm{P}(B)$. Figure P.2 draws this for the coin.

Now we turn the direction round, from the observed event back to the hypotheses. Let $H_1, \ldots, H_M$ be hypotheses that are **mutually exclusive and exhaustive** [exactly one of them is true]; for the coin, $M = 2$, with $H_1 = F$ and $H_2 = B$. Let E be the observed event, here "8 heads".

$$\begin{aligned}
\mathrm{P}(H_j \mid E) &\overset{(1)}{=} \frac{\mathrm{P}(H_j \cap E)}{\mathrm{P}(E)} \\
&\overset{(2)}{=} \frac{\mathrm{P}(E \mid H_j)\ \mathrm{P}(H_j)}{\mathrm{P}(E)} \\
&\overset{(3)}{=} \frac{\mathrm{P}(E \mid H_j)\ \mathrm{P}(H_j)}{\sum_{i=1}^{M} \mathrm{P}(E \cap H_i)} \\
&\overset{(4)}{=} \frac{\mathrm{P}(E \mid H_j)\ \mathrm{P}(H_j)}{\sum_{i=1}^{M} \mathrm{P}(E \mid H_i)\ \mathrm{P}(H_i)}.
\end{aligned} \tag{0.3}$$

(1) Definition (0.1), with E as the reference space.
(2) Product rule (0.2), used in the other order: $\mathrm{P}(H_j \cap E) = \mathrm{P}(E \mid H_j)\,\mathrm{P}(H_j)$.
(3) The pieces $E \cap H_1, \ldots, E \cap H_M$ do not overlap (the H_i exclude each other) and together make up all of E (one H_i is always true). Probabilities of non-overlapping pieces add.
(4) Product rule again, inside each term of the sum.

This is **Bayes' theorem**. In words, $\mathrm{P}(H_j \mid E) \propto \mathrm{P}(E \mid H_j)\,\mathrm{P}(H_j)$: *take each hypothesis's prior weight, multiply it by how readily that hypothesis produces what we saw, and normalise.* The denominator, $\mathrm{P}(E) = \sum_i \mathrm{P}(E \mid H_i)\,\mathrm{P}(H_i)$, is the **law of total probability**. It is the total weight of all the paths in the tree that end at the observation. The three ingredients have names. $\mathrm{P}(H)$ is the **prior**: the probability of H before the observation. $\mathrm{P}(E \mid H)$, read as a function of the hypothesis H with E held fixed, is the **likelihood**. $\mathrm{P}(H \mid E)$ is the **posterior**: the probability of H after the observation. Chapter 5 develops all of this in full.

Example 0.1 (the forward direction: what does each coin produce?). The question is step 3 of the four: if the coin is fair, how probable are 8 heads in 10 tosses, and if it is biased, how probable? Both answers come from one formula. For a coin with heads probability p, tossed n times independently, the probability of exactly k heads is

$$\mathrm{P}(k \mid p) \overset{(1)}{=} \sum_{\text{sequences with } k \text{ heads}} p^k (1-p)^{n-k} \overset{(2)}{=} \binom{n}{k} p^k (1-p)^{n-k}.$$

(1) Independent tosses multiply, so one particular sequence with k heads and $n-k$ tails has probability $p^k(1-p)^{n-k}$, whatever the order. Different sequences are different outcomes, so their probabilities add.
(2) The number of sequences with k heads among n places is the number of ways of choosing which k places are heads, $\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$.

For our two coins, the fair one with $p = \frac{1}{2}$ and the biased one with $p = 0.8$, put $n = 10$ and $k = 8$. Then $\binom{10}{8} = \frac{10 \cdot 9}{2} = 45$, so

$$\mathrm{P}(8 \mid F) \overset{(1)}{=} 45 \left(\tfrac{1}{2}\right)^{10} = \frac{45}{1024} = 0.04395,$$
$$\mathrm{P}(8 \mid B) \overset{(2)}{=} 45\,(0.8)^8 (0.2)^2 = 45 \times 0.16777 \times 0.04 = 0.3020.$$

(1) With $p = \frac{1}{2}$ every sequence has the same probability 2^{-10}.
(2) $0.8^8 = 0.16777$ and $0.2^2 = 0.04$.

The biased coin produces "8 heads" in about 30% of its ten-toss runs; the fair coin in about 4%.

Example 0.2 (the backward direction: which coin was it?). The question is step 4: having seen 8 heads, how probable is it that the coin is the biased one? We have everything Bayes' theorem (0.3) needs. The two hypotheses are F (fair) and B (biased). The drawer held equal numbers of each, so the priors are $\mathrm{P}(F) = \mathrm{P}(B) = \frac{1}{2}$. The likelihoods are the two forward numbers of Example 0.1. Then

$$\mathrm{P}(B \mid 8) \overset{(1)}{=} \frac{\mathrm{P}(8 \mid B)\,\mathrm{P}(B)}{\mathrm{P}(8 \mid B)\,\mathrm{P}(B) + \mathrm{P}(8 \mid F)\,\mathrm{P}(F)}$$
$$\overset{(2)}{=} \frac{0.3020}{0.3020 + 0.04395} = \frac{0.3020}{0.3459} = 0.873.$$

(1) Bayes' theorem (0.3): numerator is the red path through B in Figure 0.2, denominator is both red paths.

(2) The common factor $\mathrm{P}(B) = \mathrm{P}(F) = \frac{1}{2}$ cancels between numerator and denominator.

So $\mathrm{P}(B \mid 8) = 0.873$ and $\mathrm{P}(F \mid 8) = 0.127$. The coin itself did *not* change. What changed is how we share probability between the two coins it could be.

Key idea: Bayes' theorem as information shrinking the sample space

Before the evidence, the whole sample space is relevant. The evidence E tells us that we are somewhere in the region where E is true, so the sample space *shrinks* to that region. The posterior $\mathrm{P}(H \mid E)$ is the fraction of the shrunken region occupied by H:

$$\mathrm{P}(H \mid E) = \frac{\text{area of } (H \cap E)}{\text{area of } (H \cap E) + \text{area of } (\neg H \cap E)} = \frac{\mathrm{P}(E \mid H)\,\mathrm{P}(H)}{\mathrm{P}(E \mid H)\,\mathrm{P}(H) + \mathrm{P}(E \mid \neg H)\,\mathrm{P}(\neg H)}.$$

Figure P.2 draws this for the coin, with $H = B$ (the biased coin, "bent" in the figure) and $\neg H = F$ (fair; $\neg H$ means "not H"). Take a rectangle Ω, whose area 1 stands for total probability, and cut it into two strips of width $\mathrm{P}(B) = \mathrm{P}(F) = \frac{1}{2}$. The left strip is the "biased" world and the right strip is the "fair" world. In each strip, mark the part of that world in which "8 heads" happens. That part is the forward number: it fills the fraction $\mathrm{P}(8 \mid B) = 0.302$ of the biased strip and the fraction $\mathrm{P}(8 \mid F) = 0.044$ of the fair strip. So a forward probability is a *share of a strip*, and the joint probability $\mathrm{P}(8 \cap B) = \mathrm{P}(8 \mid B)\,\mathrm{P}(B)$ is an *area*, the share times the strip's width. The two marked pieces sit side by side and together make one box: the event "8 heads".

Observing 8 heads rules out everything outside the box, because those regions are worlds in which 8 heads did not happen. The box is the new sample space. The posterior $\mathrm{P}(B \mid 8)$ is the biased piece's share of the box, $0.151/(0.151 + 0.022) = 0.873$. The picture also shows why the posterior depends on the prior: making the biased strip narrower would shrink its piece, and so its share of the box.

The biased piece answers both directions. As a share of the biased strip, it is the forward probability $\mathrm{P}(8 \mid B)$. As a share of the box, it is the posterior $\mathrm{P}(B \mid 8)$. Bayes' theorem is the relation between these two ways of measuring the same piece (Section 5.2).

Example 0.3 (odds, and updating twice). With two hypotheses, Bayes' theorem is simplest as a ratio, because the normalisation drops out. Write Bayes' theorem once for B and once for F, with E the observed event ("8 heads"), and divide the first by the second. Both have the same denominator $\mathrm{P}(E)$, so it cancels:

$$\underbrace{\frac{\mathrm{P}(B \mid E)}{\mathrm{P}(F \mid E)}}_{\text{posterior odds}} \overset{(1)}{=} \underbrace{\frac{\mathrm{P}(E \mid B)}{\mathrm{P}(E \mid F)}}_{\text{likelihood ratio}} \underbrace{\frac{\mathrm{P}(B)}{\mathrm{P}(F)}}_{\text{prior odds}} \overset{(2)}{=} \frac{0.3020}{0.04395} \times 1 = 6.87.$$

(1) Ratio of two copies of (0.3); $\mathrm{P}(E)$ cancels.
(2) Numbers from Example 0.1; equal priors give prior odds 1.

Odds of 6.87 to 1 mean a probability $6.87/(1 + 6.87) = 0.873$, as before.

Now toss the same coin ten more times, and suppose we again see 8 heads. How should the odds change? Two facts answer this. First, what we believed after the first run is our starting point for the second, so the posterior odds of the first run are the prior odds of the second. Second, once the coin is fixed, the second run's tosses do not depend on the first run's, so the second run has the same likelihood ratio as the first. The second run therefore multiplies the odds by 6.87 again:

$$\text{odds after two runs} \overset{(1)}{=} 6.87 \times 6.87 = 47.2, \qquad \mathrm{P}(B \mid \text{both runs}) \overset{(2)}{=} \frac{47.2}{48.2} = 0.979.$$

(1) Posterior odds = likelihood ratio × prior odds, with the prior odds now equal to 6.87.
(2) Odds o correspond to probability $o/(1+o)$.

Independent pieces of evidence multiply the odds, so they add in the logarithm of the odds. This is why many independent data points can overwhelm any reasonable prior.

What this bought us. The observation "8 heads" can be reached by two paths through the tree, one through each coin. The biased path is 6.87 times more likely to produce it, so the posterior odds are 6.87 times the prior odds. Inference is this *reallocation of credibility across possibilities*: the possibilities that fit the data gain credibility, and the others lose it [Kruschke pp. 16–22]. Sherlock Holmes' rule, "eliminate the impossible and whatever remains must be the truth", is the extreme case. There some likelihoods are exactly zero, and those hypotheses lose all their credibility. Kruschke's bouncy-ball example (three noisy measurements of a diameter, four candidate ball sizes) is our coin with a continuous outcome. The two directions are also the two basic problems of statistics [Wasserman p. ix]. Forward: *given a data generating process, what are the properties of the outcomes?* Backward: *given the outcomes, what can we say about the process that generated them?*

The two directions in code

The script below does both directions with numbers. Forward, it simulates 20000 ten-toss experiments with each coin and compares how often each head count k turns up with the exact binomial probabilities of Example 0.1. Backward, it computes the posterior $\mathrm{P}(B \mid k)$ for every possible outcome $k = 0, \ldots, 10$, not only for $k = 8$.

Listing 0.1: The key lines: forward simulation, exact pmf, and Bayes on each column of the tree (`code/ch00/01_coin_two_directions.py`)

```
k = np.arange(N_TOSS + 1)


def binom_pmf(k, n, p):
    """P(k heads in n tosses | p): the forward (prediction) question."""
    return np.array([comb(n, int(j)) * p**j * (1 - p) ** (n - j) for j in k])


# ---- forward direction: each coin produces a distribution of outcomes ----
pmf_fair, pmf_bias = binom_pmf(k, N_TOSS, P_FAIR), binom_pmf(k, N_TOSS, P_BIAS)
sim_fair = rng.binomial(N_TOSS, P_FAIR, size=N_EXP)      # 20000 experiments of 10 tosses
sim_bias = rng.binomial(N_TOSS, P_BIAS, size=N_EXP)
freq_fair = np.bincount(sim_fair, minlength=N_TOSS + 1) / N_EXP
freq_bias = np.bincount(sim_bias, minlength=N_TOSS + 1) / N_EXP

# ---- backward direction: Bayes on the observed column of the tree ----
num_bias = pmf_bias * PRIOR_BIAS                  # path "biased, then k heads"
num_fair = pmf_fair * (1 - PRIOR_BIAS)            # path "fair, then k heads"
post_bias = num_bias / (num_bias + num_fair)      # normalise over the paths that end at k
```

The function `binom_pmf` is Example 0.1 line for line. The generator call `rng.binomial`, with `size=N_EXP`, performs the forward experiment 20000 times and returns the list of head counts. `np.bincount` counts how often each value of k appears in that list; dividing by 20000 turns the counts into empirical frequencies. The last three lines are Bayes' theorem (0.3). `num_bias` and `num_fair` are the weights of the two paths in the tree that end at k, and dividing by their sum is the normalisation by $\mathrm{P}(k)$.

For $k = 8$ the run gives an empirical frequency 0.04105 for the fair coin, against the exact 0.04395, and 0.3024 for the biased coin, against the exact 0.302. A theoretical probability and

an empirical frequency are different objects. A fair coin has probability $\frac{1}{2}$ of heads, yet a run of 10 tosses may give 3 heads and a run of 100 may give 45. The simulation lets us watch the frequency settle onto the probability as the number of runs grows [SeeingTheory p. 5]. The simulated fraction still fluctuates. A fraction of N independent trials, each succeeding with probability q, has standard deviation $\sqrt{q(1-q)/N}$ (derived in Example 0.4). Here that is 0.001449, so both agreements are within the expected scatter. Backward, the posterior at $k = 8$ is 0.873, reproducing Example 0.2; at $k = 5$ it has fallen to 0.09696.

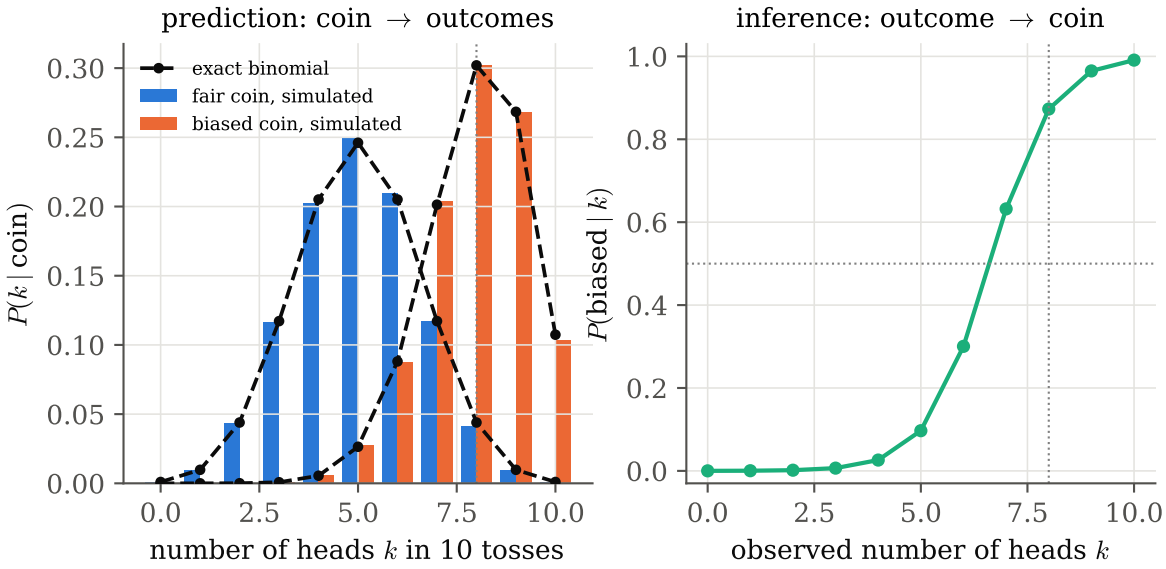


Figure 0.3: Left, prediction: each coin, tossed ten times over and over, spreads its outcomes over $k = 0, \ldots, 10$. Bars are simulated frequencies, dashed curves the exact binomial probabilities; the dotted line marks the observed $k = 8$. Right, inference: for each possible observation k, the probability that the coin was the biased one. The curve crosses the prior value $\frac{1}{2}$ between $k = 6$ and $k = 7$, where the two coins explain the data about equally well.

The forward number $\mathrm{P}(E \mid H)$ and the backward number $\mathrm{P}(H \mid E)$ can be far apart. The classic case is a medical test, and it goes through the same four steps. A disease affects 1% of people. A test detects it with probability 0.95 and gives a false positive with probability 0.10.

The event. A patient tests positive, written +.

What could have produced it? The one thing we do not know is whether the patient is ill. So there are two candidates: the patient has the disease (D), or does not (D^c).

How probable was the event under each? If the patient is ill, the test catches it 95% of the time: $\mathrm{P}(+ \mid D) = 0.95$. If the patient is healthy, a positive result is a false alarm, which happens 10% of the time: $\mathrm{P}(+ \mid D^c) = 0.10$.

Only then the priors, and the normalisation. Before the test, the patient was ill with probability 0.01 and healthy with probability 0.99. Bayes' theorem (0.3) gives

$$\mathrm{P}(D \mid +) = \frac{0.95 \times 0.01}{0.95 \times 0.01 + 0.10 \times 0.99} = \frac{0.0095}{0.1085} = 0.0876.$$

Watch out: $\mathrm{P}(E \mid H)$ **is not** $\mathrm{P}(H \mid E)$

The two numbers answer different questions and can be wildly different. For the positive test, the forward number $\mathrm{P}(+ \mid D) = 0.95$ is not the answer to "is the patient ill?"; the answer is $\mathrm{P}(D \mid +) = 0.0876$. Swapping the two conditionals is the "prosecutor's fallacy".

Its statistical twin appears in Chapter 4: a p-value is *not* the probability that the null hypothesis is true.

The lesson is not in the arithmetic. A positive result can come by two paths: through the disease, or through a false alarm on a healthy person. Each healthy person is unlikely to take the false-alarm path, but there are 99 healthy people for every ill one. The positive result favours the disease only by the factor $\mathrm{P}(+ \mid D)/\mathrm{P}(+ \mid D^c) = 0.95/0.10 = 9.5$. A factor of 9.5 acting on a prior of 1% leaves the posterior below 10%, far from the 95% that the test's sensitivity suggests (Section 5.2 draws it to scale).

Two meanings of "prediction". In this chapter "prediction" means the forward direction, $H \to E$. Statistics and machine-learning books also use the word for guessing a *new* data point (a regression output, a class label) after learning from old data. That second kind of prediction is part of inference, because it starts from data. When Wasserman lists "prediction, classification, clustering, and estimation" as special cases of inference, he means the second sense.

0.2 The questions the book answers

Take a single coin from the magic shop of Section 0.1. Its heads probability p is unknown, but we can toss it as often as we like. Even with this one experiment, the calculation depends on the question we ask about the count. *How likely is 8 heads if $p = \frac{1}{2}$?* is a forward calculation. *What is our best guess for p, and how wrong might it be?* is estimation. *Is 8 heads in 10 surprising for a fair coin?* is a test. *Which values of p are plausible now?* is Bayesian inference. *Could we get any of these numbers by simulation instead of algebra?* is Monte Carlo. *How many tosses would we need to pin p down to ± 0.01?* is a Fisher forecast. Each question has its own object that answers it, and each has its own chapter. If we keep the questions apart, then for any tool we can say which question it answers.

Same system, different measurements. A physicist knows this from the lab. One pendulum supports many experiments. We can measure its period, estimate g from the period, test whether the period depends on the amplitude, or forecast how precisely a longer run would pin down g. The pendulum stays the same. What changes is the question, and with it the quantity we must compute. Probability tools work the same way: the model stays fixed, and each tool computes the quantity that answers one question about it.

The questions in words. Before any formula, each kind of question can be said in one sentence.

1. *Forward:* if the hypothesis is true, how often does this outcome occur?
2. *Spread:* what values can a random quantity take, and how are they distributed?
3. *Estimation:* what value of the parameter do the data point to, and how far off is that guess likely to be?
4. *Testing:* if only chance were at work, would a result like this, or one even more extreme in the same direction, be something we would reasonably expect?
5. *Bayesian inference:* which parameter values could have produced the data, and how probable is each one?

6. *Simulation:* can we get the answer to any of these by drawing random numbers, when the algebra is too hard?
7. *Forecasting:* before the experiment is run, how well could it measure the parameter?
8. *Choosing the summary:* which property of a field (a map, a density) carries the physics, and does it survive the nuisance effects of observation?

The table below is the map of the book. Each row gives one question, the object that answers it, and the chapter where that object is built. The pictogram on the left of each row is a miniature of the central figure of that chapter, the picture to remember for the row.

Most of the rows can be tried on the coin in a few lines. Four of them need a new idea: estimation, testing, Monte Carlo and forecasting. Two definitions carry all four.

Working definition 0.2 (estimator, sampling distribution, p-value)

An **estimator** is a recipe that turns data into a guess for a parameter, $\hat{\theta} = \hat{\theta}(\boldsymbol{d})$ [a function of the data only, e.g. $\hat{p} = k/n$, the fraction of heads]. Because the data are random, so is $\hat{\theta}$. Its distribution over imagined repetitions of the whole experiment is its **sampling distribution** [the histogram of $\hat{\theta}$ if the experiment were rerun many times with the same true θ]. Its **bias** is $\mathbb{E}[\hat{\theta}] - \theta$ and its spread is $\mathrm{Var}(\hat{\theta})$.
A **null hypothesis** H_0 is a hypothesis whose sampling distribution we know ["nothing but chance", e.g. a fair coin]. Given an observed value t_{obs} of a statistic T, the **p-value** is

$$p = \mathrm{P}\left(T \geq t_{\mathrm{obs}} \,|\, H_0\right), \tag{0.4}$$

the probability, computed under H_0, of a result like the observed one *or even more in this direction* [here, larger values of T count as further out]. We can compute it only because "pure chance" is the one case whose distribution we have in hand. For a fair coin we know exactly how often each head count turns up. A small p says that, if the coin really were fair, a result this far out would be a surprise (Chapter 4).

Example 0.4 (the scatter of a fraction). The natural guess for p is the fraction of heads. The question is whether this guess is right on average, and how far it typically misses. Toss the coin n times and let $x_i = 1$ if toss i is heads and $x_i = 0$ if it is tails. The number of heads is $k = \sum_i x_i$, so the estimator is $\hat{p} = k/n = \frac{1}{n}\sum_i x_i$. Its mean and variance are

$$\mathbb{E}[\hat{p}] \overset{(1)}{=} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[x_i] \overset{(2)}{=} \frac{1}{n}\, n\, p = p,$$

$$\mathrm{Var}(\hat{p}) \overset{(3)}{=} \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(x_i) \overset{(4)}{=} \frac{1}{n^2}\, n\, \left(\mathbb{E}[x_i^2] - \mathbb{E}[x_i]^2\right) \overset{(5)}{=} \frac{p - p^2}{n} = \frac{p(1-p)}{n}.$$

(1) The expectation of a sum is the sum of expectations, and constants come out.
(2) $\mathbb{E}[x_i] = 1 \cdot p + 0 \cdot (1-p) = p$.
(3) For *independent* tosses the variance of a sum is the sum of variances; a constant c comes out as c^2.
(4) Definition of the variance, $\mathrm{Var}(x) = \mathbb{E}[x^2] - \mathbb{E}[x]^2$.
(5) $x_i^2 = x_i$ because x_i is 0 or 1, so $\mathbb{E}[x_i^2] = p$.

So $\hat{p}$ is **unbiased**: on average it gives the true p. Its standard deviation, the typical miss, is $\sqrt{p(1-p)/n}$. For $p = 0.8$ this is 0.126 at $n = 10$ and 0.040 at $n = 100$. The same algebra applies to *any* fraction of N independent trials that each succeed with probability q. That is why a Monte Carlo estimate of a probability q from N simulated trials has error $\sqrt{q(1-q)/N}$.

Taxonomy: which object answers which question

	The question, in words	The object that answers it	Ch.
	How likely is this event, if the situation is as assumed? Which paths lead to it?	$\mathrm{P}(E \mid H)$, summed over the paths of a probability tree	1
	What values can a random quantity take, and how are they spread?	a distribution: pmf or pdf, CDF, mean and variance (binomial, Poisson, Gaussian, . . .)	2
	What value of the parameter do the data point to, and how far off is that guess?	an estimator $\hat{\theta}$, its bias and variance, i.e. its sampling distribution; the likelihood $\mathcal{L}(\theta)$	3
	If H_0 (the case whose distribution we know: "it happened by chance") were true, would a result like this, or more extreme in this direction, be expected?	a test statistic and its p-value, the tail probability under H_0	4
	Which parameter values (or models) could have produced the data, and how probable is each?	the posterior $p(\boldsymbol{\theta} \mid \boldsymbol{d}) \propto \mathcal{L}(\boldsymbol{\theta})\, p(\boldsymbol{\theta})$; the evidence Z for comparing models	5
	What is the distribution of a quantity too complicated to compute by hand?	many simulated realisations; histograms and averages, with error $\propto 1/\sqrt{N}$	6
	How do we summarise a random field (a sky map), and how does the summary fluctuate?	correlation function $\xi(r)$, power spectrum $P(k)$, C_ℓ and its estimator $\widehat{C}_\ell$	7
	How do a mask, noise and a beam change the sampling distribution of the summary?	Monte Carlo as the research instrument: simulated skies through the observation pipeline	8
	How do we draw samples from a posterior in many dimensions without ever normalising it?	a Markov chain: a *biased* random walker that lingers where the posterior is high	9
	Before the experiment, how well could it measure $\boldsymbol{\theta}$?	the Fisher matrix $\mathbf{F}$; forecast errors $\sigma_i \geq \sqrt{(\mathbf{F}^{-1})_{ii}}$	10
	What do the positions and motions of galaxies say about how the Universe expands and how structure grows?	galaxy catalogues: $\xi(s)$ and $P(k)$ with weights and windows, the BAO ruler, peculiar velocities	11
	Is there information in how the structures of a field are connected, robust to nuisance effects?	excursion sets, Betti numbers, Euler characteristic, persistence: new random observables S with a $P(S \mid \theta)$	12
	Does the whole chain survive contact with real data?	complete analyses of Planck, SDSS and collider data, each reproducing a published result	13

Chapter 14 puts the rows back together: given a new problem, which row does it belong to?

Example 0.5 (the p-value of 8 heads). *The observation.* Ten tosses gave 8 heads. *The question.* If the coin were fair, would 8 heads, or more than 8 heads, be something we would reasonably expect? *The pieces.* The null hypothesis is H_0: $p = \frac{1}{2}$. The statistic is the number of heads, $T = k$, and "more extreme" means more heads. Then

$$\mathrm{P}(K \geq 8 \mid H_0) \overset{(1)}{=} \mathrm{P}(8) + \mathrm{P}(9) + \mathrm{P}(10) \overset{(2)}{=} \frac{\binom{10}{8} + \binom{10}{9} + \binom{10}{10}}{2^{10}} = \frac{45 + 10 + 1}{1024} = \frac{56}{1024} = 0.0547.$$

(1) "This or more extreme in this direction": the outcomes $8, 9, 10$ are exclusive, so their probabilities add.
(2) Binomial probabilities with $p = \frac{1}{2}$ (Example 0.1): every sequence has probability 2^{-10}.

About 5% of fair ten-toss runs give 8 heads or more (Figure 0.4). That is mildly surprising, not damning. The p-value used only the distribution under H_0; it never mentioned the biased coin. It is *not* $\mathrm{P}(H_0 \mid \text{data})$, the probability that the coin is fair. In the drawer story of Section 0.1 (equal numbers of fair and biased coins) that number was $\mathrm{P}(F \mid 8) = 0.127$, the answer to a different question.

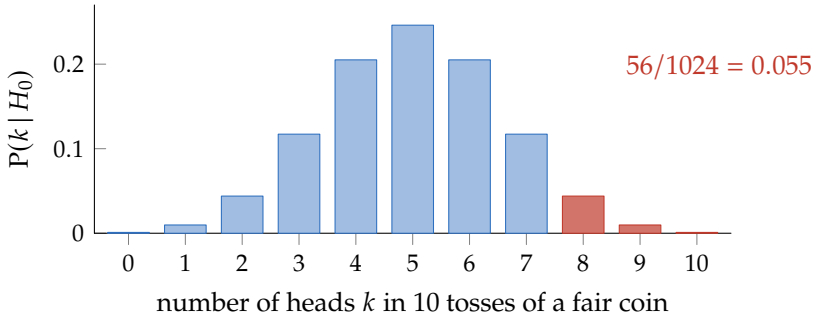


Figure 0.4: The null distribution for ten tosses of a fair coin. The red bars are the observed $k = 8$ and the outcomes more extreme in the same direction; their total height is the p-value. The blue bars never enter.

Example 0.6 (a forecast before any toss: the Fisher information). The question is asked before any toss: how precisely *could* n tosses measure p? The answer depends on how sharply the data single out one value of p. The likelihood of one particular sequence with k heads is $\mathcal{L}(p) = p^k(1-p)^{n-k}$, so its logarithm is $\ln \mathcal{L}(p) = k \ln p + (n-k)\ln(1-p)$. A log-likelihood with a sharp peak pins p down; a flat one does not. The **Fisher information** measures the sharpness. It is minus the expected curvature of $\ln \mathcal{L}$ at the true p:

$$I(p) \overset{(1)}{=} -\mathbb{E}\left[\frac{\mathrm{d}^2 \ln \mathcal{L}}{\mathrm{d}p^2}\right] \overset{(2)}{=} \mathbb{E}\left[\frac{k}{p^2} + \frac{n-k}{(1-p)^2}\right] \overset{(3)}{=} \frac{np}{p^2} + \frac{n(1-p)}{(1-p)^2} = \frac{n}{p} + \frac{n}{1-p} = \frac{n}{p(1-p)}.$$

(1) Definition (Chapter 10): a sharply curved log-likelihood pins the parameter down.
(2) $\frac{\mathrm{d}}{\mathrm{d}p} \ln \mathcal{L} = \frac{k}{p} - \frac{n-k}{1-p}$; differentiate again and change sign.
(3) $\mathbb{E}[k] = np$ and $\mathbb{E}[n-k] = n(1-p)$.

The Cramér–Rao bound (Chapters 3 and 10) says that no unbiased estimator can beat $\mathrm{Var}(\hat{\theta}) \geq 1/I$. Here $1/I = p(1-p)/n$, exactly the variance of $\hat{p}$ found in Example 0.4. So the fraction of heads is as good as any unbiased guess can be. The forecast needs no data, only an assumed p. To reach a standard deviation $\sigma_p = 0.01$ at $p \approx 0.8$ we need $n = p(1-p)/\sigma_p^2 = 0.16/10^{-4} = 1600$ tosses.

What this bought us. The same coin gave three different numbers: a spread $\sqrt{p(1-p)/n}$ (estimation), a tail probability 0.055 (testing) and a required number of tosses, 1600 (forecasting). None of them is a posterior. Each answers its own row of the table. Confusing the rows is the most common error in applied statistics.

Watch out: what a p-value and a confidence interval do *not* say

A p-value of 0.055 does not mean "the coin is fair with probability 0.055". It is $\mathrm{P}(\text{data this extreme} \mid H_0)$, a forward number. Likewise, a 95% confidence interval is a recipe that covers the true value in 95% of repetitions of the experiment; it is not an interval that contains the parameter with probability 95%. Only the posterior of Chapter 5 makes probability statements about parameters.

The questions in code

The script `02_coin_questions.py` checks two of these answers by brute force. The key lines are:

Listing 0.2: Estimation and a p-value by simulation (`code/ch00/02_coin_questions.py`)

```
# (i) sampling distribution of the estimator p_hat = k/n
n_rep = 50000
phat10 = rng.binomial(10, P_TRUE, n_rep) / 10
phat100 = rng.binomial(100, P_TRUE, n_rep) / 100

# (ii) exact p-value: P(K >= 8 | p = 1/2), "this or more extreme in this direction"
p_exact = sum(comb(N_TOSS, j) for j in range(K_OBS, N_TOSS + 1)) / 2**N_TOSS

# (iii) Monte Carlo estimate of the same number: fraction of simulated fair
#       experiments with K >= 8, for N = 10 ... 10^6 experiments
Ns = np.unique(np.logspace(1, 6, 40).astype(int))
fair = rng.binomial(N_TOSS, 0.5, Ns.max())
hits = np.cumsum(fair >= K_OBS)                  # running count of "as or more extreme"
p_mc = hits[Ns - 1] / Ns
sigma_mc = np.sqrt(p_exact * (1 - p_exact) / Ns)   # binomial error of a fraction
```

The first block reruns the whole ten-toss experiment, and the hundred-toss one, 50 000 times with the true $p = 0.8$, and keeps each $\hat{p}$. That list is a sample from the sampling distribution of $\hat{p}$. The second block is Example 0.5 in one line. The third block simulates fair experiments. In it, `np.cumsum(fair >= K_OBS)` counts, for every N at once, how many of the first N experiments were as extreme as the observed 8 heads. The last line is the error bar $\sqrt{q(1-q)/N}$ of Example 0.4.

The run confirms both calculations. The standard deviation of $\hat{p}$ is 0.1267 at $n = 10$ (theory 0.1265) and 0.03997 at $n = 100$ (theory 0.04). The mean at $n = 10$ is 0.7993, so the estimator is unbiased, as derived. The Monte Carlo p-value is 0.0566 after $N = 106$ experiments and $0.0545 \pm 2.274 \times 10^{-4}$ after 10^6, against the exact 0.05469.

How this is used in real research: verification versus research

Monte Carlo has two uses. Here it only *verifies* a number we could compute exactly, 56/1024; the first use is to check theory and code against each other. The second use is as the *research instrument*, where no exact answer exists. Suppose the "coin" were a CMB sky seen through a galactic mask, with noise and a beam. Nobody can write down the distribution of the measured power spectrum. Instead one simulates thousands of skies and makes a histogram. The code is the same: draw, compute the statistic, count. Chapter 6 builds the

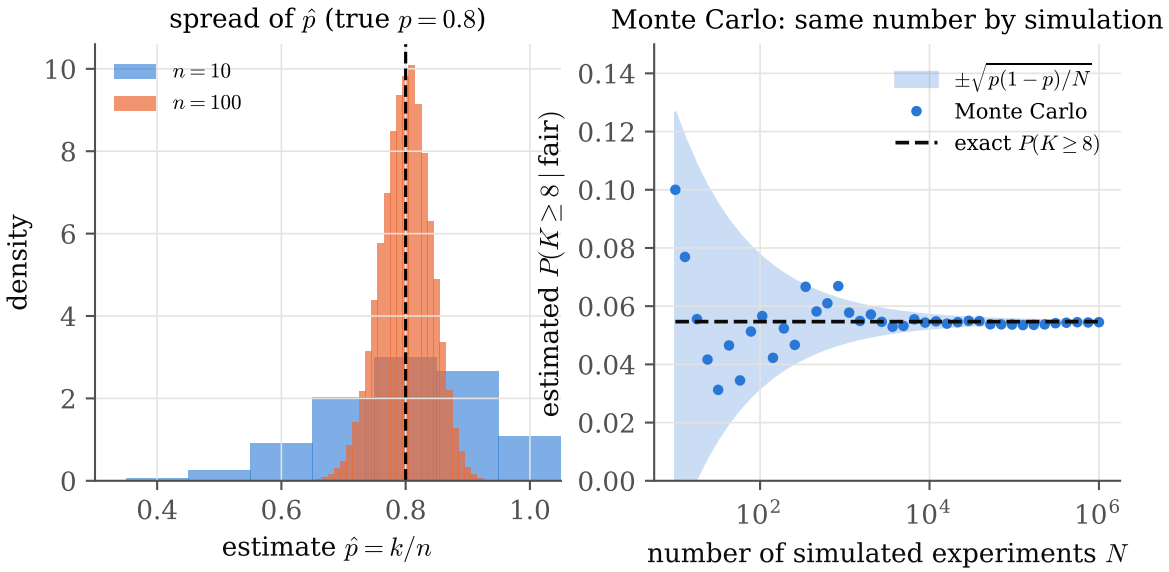


Figure 0.5: Left: the sampling distribution of $\hat{p} = k/n$ for a coin with $p = 0.8$. With 10 tosses the guesses spread over most of the interval; with 100 they crowd around the true value, ten times more tightly in variance. Right: the Monte Carlo estimate of $\mathrm{P}(K \geq 8 \mid \text{fair})$ as the number of simulated experiments grows. The points wander inside the shaded $\pm\sqrt{p(1-p)/N}$ band, which narrows onto the exact value.

first use, Chapter 8 the second.

The same idea under different names. The vocabulary changes between communities, but the ideas do not. What statisticians call *estimation*, the machine-learning literature calls *learning*. The observed values (X_i, Y_i) are there the *training sample*. The inputs X_i are *covariates* in one language and *features* in the other. A few more terms recur. A *hypothesis* is a subset of the parameter space. A *confidence interval* is a recipe that covers the unknown value with a stated frequency. *Bayesian inference* means using data to update beliefs, whereas *frequentist inference* collects the methods whose long-run frequency behaviour is guaranteed [Wasserman p. xi]. Prediction in the second sense of Section 0.1, together with classification, clustering and estimation, counts as a special case of inference [Wasserman p. ix]. Only the Bayesian rows of the table above (Chapters 5 and 9) share out credibility among parameter values [Kruschke p. 15]. The estimation and testing rows instead describe how a *procedure* behaves over repeated experiments. The null hypothesis is the case whose distribution we know, because we assume the result happened by chance [Wasserman p. 149].

0.3 The pipeline: from a physical model to a statement about physics

Real data are never fed into Bayes' theorem whole; they are compressed first. A Planck temperature map at the resolution used for cosmology has about 5×10^7 pixels. The model we want to test, ΛCDM, has six parameters. Nobody writes Bayes' theorem directly for fifty million correlated pixels. Instead the map is compressed into its angular power spectrum $\widehat{C}_\ell$, about 2500 numbers, and those are compared with the theory. Each step of this is a choice. First we decide that the power spectrum is the information that matters. Then we work out how $\widehat{C}_\ell$ fluctuates from one sky to another. Only then do we infer the six parameters. Written

out as one pipeline, this chain of choices has an arrow for every tool in the book. Knowing which arrow we are on is the fastest way to know what a tool is for.

Microstates and macrostates. Statistical mechanics works the same way. A gas has 10^{23} momenta, but we never infer anything from the full microstate. We compress it into a few macroscopic summaries (pressure, energy, magnetisation). Fluctuation theory then tells us how much those summaries scatter; for example, the energy of a system at temperature T with heat capacity C_V fluctuates by $\langle \Delta E^2 \rangle = k_B T^2 C_V$. A summary is chosen for two reasons: it carries the physics we care about, and its fluctuations can be calculated. Data analysis makes the same two demands of every summary statistic.

Six stages, six questions. The whole book runs along one chain, drawn in Figure 0.6:

$$\boxed{\begin{array}{l}\text{a theory with parameters } \boldsymbol{\theta} \;\rightarrow\; \text{the map or data set it produces} \\ \quad \rightarrow \text{ decide which features of the data carry the physics} \\ \quad \rightarrow \text{ compress the data into a statistic } S \;\rightarrow\; \text{work out how } S \text{ scatters, } P(S \mid \boldsymbol{\theta}) \\ \quad \rightarrow \text{ say what the observed } S \text{ tells us about } \boldsymbol{\theta}\end{array}} \tag{0.5}$$

In symbols, the last two stages are $P(S \mid \theta)$ and $P(\theta \mid S) \propto P(S \mid \theta) P(\theta)$, where θ stands for the parameters $\boldsymbol{\theta}$. The stages answer different kinds of question. The choice of summary decides what should be measured. The last two stages decide how the uncertainty in that measurement carries over to the parameters. The same chain applies whether the data vector holds pixelised CMB temperatures or measured bandpowers C_ℓ, and whatever the number M of model parameters. Each stage is a question.

1. *Physical model:* what theory, with which parameters $\boldsymbol{\theta}$, could have produced the data?
2. *Field or dataset:* what exactly did the instrument record, and how does it depend on $\boldsymbol{\theta}$ and on noise?
3. *What information matters:* which property of the data is sensitive to $\boldsymbol{\theta}$, and which is only nuisance?
4. *Summary:* which function $S[\boldsymbol{d}]$ of the data keeps that property and discards the rest?
5. *Sampling distribution:* if the experiment were repeated with the same $\boldsymbol{\theta}$, how would S scatter? That is $P(S \mid \boldsymbol{\theta})$.
6. *Infer physics:* given the one S we observed, which $\boldsymbol{\theta}$ are plausible? That is $P(\boldsymbol{\theta} \mid S) \propto P(S \mid \boldsymbol{\theta}) P(\boldsymbol{\theta})$.

Stages 1 to 5 are the forward direction of Section 0.1. Stage 6 is the backward direction.

One step of the pipeline is not mathematics: choosing what information matters. It is a physics decision, and Section 0.4 takes it up. The next step, building the summary, does have a precise test: it tells us when compressing the data loses nothing.

Working definition 0.3 (summary statistic, sampling distribution, sufficiency)

A **summary statistic** is any function $S = S[\boldsymbol{d}]$ of the data [a number or a short vector computed from the data alone, e.g. the number of heads k]. Its **sampling distribution** $P(S \mid \boldsymbol{\theta})$ is the distribution of S over repetitions of the experiment with the parameters held fixed [the histogram of S over many simulated universes that share the same $\boldsymbol{\theta}$]. The

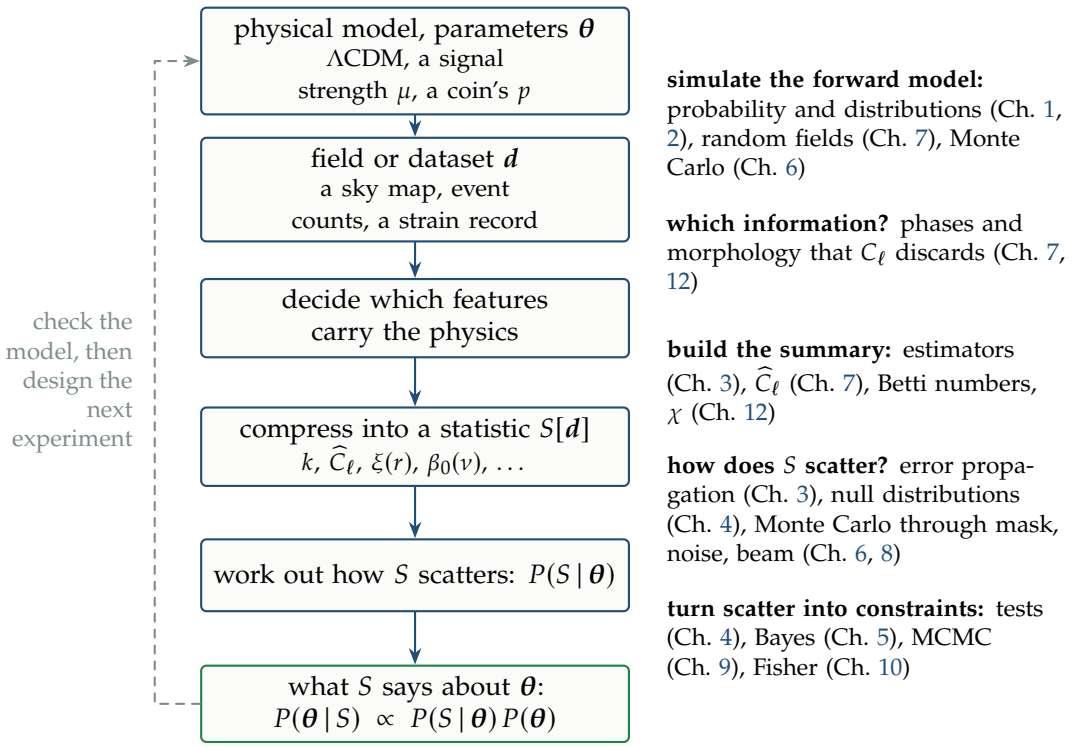


Figure 0.6: The pipeline of the whole book. Boxes are the stages; the notes on the right say which chapters serve the step from one box to the next. The top four boxes are forward modelling and a choice, the fifth is probability theory applied to that choice, and only the last box is inference. The dashed loop is the step statisticians call a posterior predictive check: simulate data from the inferred model and ask whether they look like the real data.

summary is **sufficient** for $\boldsymbol{\theta}$ if the likelihood depends on the data only through S:

$$P(\boldsymbol{d} \mid \boldsymbol{\theta}) = g\big(S[\boldsymbol{d}], \boldsymbol{\theta}\big)\, h(\boldsymbol{d}) \tag{0.6}$$

[h may depend on the data but not on $\boldsymbol{\theta}$, so it cancels from every posterior and every likelihood ratio]. Then $P(\boldsymbol{\theta} \mid \boldsymbol{d}) = P(\boldsymbol{\theta} \mid S)$: nothing about $\boldsymbol{\theta}$ was thrown away.

The coin shows how the test works. The data are the whole sequence $x_1, \ldots, x_n$ of zeros and ones ($x_i = 1$ for heads, as in Example 0.4). Is the number of heads enough, or does the order of the tosses also tell us something about p? Write down the probability of the sequence:

$$\begin{aligned} P(x_1, \ldots, x_n \mid p) &\overset{(1)}{=} \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\ &\overset{(2)}{=} p^{\sum_i x_i}\,(1-p)^{n-\sum_i x_i} \overset{(3)}{=} p^k (1-p)^{n-k}. \end{aligned} \tag{0.7}$$

(1) The tosses are independent; one toss gives p if $x_i = 1$ and $1-p$ if $x_i = 0$, which is what $p^{x_i}(1-p)^{1-x_i}$ says.
(2) Products of powers of the same base add their exponents.
(3) $k = \sum_i x_i$ is the number of heads.

The sequence enters only through k. So (0.7) has the form (0.6) with $S = k$, $g = p^k(1-p)^{n-k}$ and $h = 1$, and the order of the tosses carries no information about p. The sampling distribution of the summary k is the binomial of Example 0.1. The last stage puts it into Bayes' theorem.

Example 0.7 (the coin through the whole pipeline). The last stage asks: after k heads in n tosses, which values of the heads probability p are plausible, and how plausible is each? Now p can be any number in $[0,1]$, not one of two coins. Take a flat prior, $p(p) = 1$ on $[0,1]$, which treats every value as equally plausible beforehand. The posterior for p is then

$$p(p \mid k) \overset{(1)}{=} \frac{P(k \mid p) \cdot 1}{\int_0^1 P(k \mid p')\,\mathrm{d}p'} \overset{(2)}{=} \frac{p^k(1-p)^{n-k}}{\int_0^1 p'^k(1-p')^{n-k}\,\mathrm{d}p'} \overset{(3)}{=} \frac{(n+1)!}{k!\,(n-k)!}\,p^k(1-p)^{n-k}.$$

(1) Bayes' theorem for a continuous parameter: the sum over hypotheses in (0.3) becomes an integral.
(2) The factor $\binom{n}{k}$ does not depend on p and cancels between numerator and denominator.
(3) The Beta integral $\int_0^1 p^a(1-p)^b\,\mathrm{d}p = \frac{a!\,b!}{(a+b+1)!}$ for non-negative integers a, b (proved by b integrations by parts, each moving one power from $(1-p)$ to p).

Where is this posterior centred, and how wide is it? Its mean and variance follow from the same Beta integral:

$$\mathbb{E}[p \mid k] \overset{(1)}{=} \frac{(n+1)!}{k!\,(n-k)!}\int_0^1 p^{k+1}(1-p)^{n-k}\mathrm{d}p \overset{(2)}{=} \frac{(n+1)!}{k!\,(n-k)!}\cdot\frac{(k+1)!\,(n-k)!}{(n+2)!} = \frac{k+1}{n+2},$$

$$\mathbb{E}[p^2 \mid k] \overset{(3)}{=} \frac{(k+1)(k+2)}{(n+2)(n+3)}, \qquad \mathrm{Var}(p \mid k) \overset{(4)}{=} \frac{(k+1)(n-k+1)}{(n+2)^2(n+3)}.$$

(1) Definition of the mean: multiply the posterior by p and integrate.
(2) Beta integral with $a = k+1$, $b = n-k$; then $(k+1)!/k! = k+1$ and $(n+1)!/(n+2)! = 1/(n+2)$.
(3) Same steps with $a = k+2$.
(4) $\mathbb{E}[p^2] - \mathbb{E}[p]^2$, put over the common denominator $(n+2)^2(n+3)$: the numerator is $(k+1)\big[(k+2)(n+2) - (k+1)(n+3)\big] = (k+1)(n-k+1)$.

For large n the variance tends to $\hat{p}(1-\hat{p})/n$ with $\hat{p} = k/n$. This is the same number as the sampling variance $p(1-p)/n$ of Example 0.4, with $\hat{p}$ in place of p. Here are the six stages for the coin. Model: a coin with unknown p. Data: a toss sequence. What matters: the fraction of heads, since the order carries no information. Summary: k. Sampling distribution: binomial. Inference: the Beta-shaped posterior.

Example 0.8 (the CMB through the pipeline). The same six stages apply to a map of the sky. The data are the temperature fluctuation $\Delta T(\hat{n})$ in each direction $\hat{n}$. Expand it in spherical harmonics $Y_{\ell m}$, the sphere's analogue of Fourier modes: $\Delta T(\hat{n}) = \sum_{\ell m} a_{\ell m} Y_{\ell m}(\hat{n})$. A multipole ℓ corresponds to an angular scale of roughly $180^\circ/\ell$, and m runs from $-\ell$ to ℓ. Statistical isotropy [no preferred direction, so the statistics are unchanged by rotations] means $\mathbb{E}[a_{\ell m} a^*_{\ell' m'}] = C_\ell\,\delta_{\ell\ell'}\delta_{mm'}$. So the theory predicts one number per ℓ, the power spectrum $C_\ell(\boldsymbol{\theta})$. The natural summary averages the squared sizes of the $2\ell+1$ measured coefficients at each ℓ. Is it right on average?

$$\mathbb{E}\left[\widehat{C}_\ell\right] \overset{(1)}{=} \mathbb{E}\left[\frac{1}{2\ell+1}\sum_{m=-\ell}^{\ell}|a_{\ell m}|^2\right] \overset{(2)}{=} \frac{1}{2\ell+1}\sum_{m=-\ell}^{\ell}\mathbb{E}\left[a_{\ell m}a^*_{\ell m}\right] \overset{(3)}{=} \frac{1}{2\ell+1}(2\ell+1)\,C_\ell = C_\ell.$$

(1) Definition of the estimator $\widehat{C}_\ell$: the average power in the $2\ell+1$ modes of multipole ℓ.
(2) $|a|^2 = aa^*$, and the expectation of a sum is the sum of expectations.
(3) Isotropy with $\ell' = \ell$, $m' = m$; there are $2\ell+1$ values of m.

So $\widehat{C}_\ell$ is unbiased. It still scatters, because only $2\ell+1$ modes exist at each ℓ and an average of few numbers is noisy. For a Gaussian sky the scatter is $\sqrt{2/(2\ell+1)}\,C_\ell$; this is cosmic variance,

derived in Chapter 7. The pipeline reads: model (ΛCDM) → map → "the physics is in the variance per scale" → $\widehat{C}_\ell$ → $P(\widehat{C}_\ell \mid \boldsymbol{\theta})$ → posterior on six parameters. For a full-sky Gaussian map without noise, $\widehat{C}_\ell$ is sufficient, exactly as k is for the coin (Example 0.13). For a masked or noisy sky, or a non-Gaussian one, it is not, and that is where Chapters 8 and 12 begin.

What this bought us. Compression is safe exactly when the likelihood depends on the data only through the summary. For the coin this holds for k. For a Gaussian isotropic full sky it holds for $\widehat{C}_\ell$. Any other compression is a choice whose cost must be checked.

The pipeline in code

The script `03_pipeline_coin.py` checks the sufficiency claim and runs the pipeline for growing n.

Listing 0.3: Two sequences with the same k have the same likelihood (`code/ch00/03_pipeline_coin.py`)

```
def seq_likelihood(x, p):
    """P(x_1..x_n | p) = prod p^x_i (1-p)^(1-x_i) for a 0/1 sequence x."""
    return np.prod([p**xi * (1 - p) ** (1 - xi) for xi in x], axis=0)


# (a) two different sequences, both with 8 heads in 10
seqA = np.array([1, 1, 1, 1, 1, 1, 1, 1, 0, 0])
seqB = np.array([0, 1, 1, 0, 1, 1, 1, 1, 1, 1])
LA, LB = seq_likelihood(seqA, p), seq_likelihood(seqB, p)
max_diff = np.max(np.abs(LA - LB))
```

Listing 0.4: The posterior after $n = 10, 100, 1000$ tosses (`code/ch00/03_pipeline_coin.py`)

```
for i, n in enumerate([10, 100, 1000]):
    kk = int(tosses[:n].sum())
    logpost = kk * np.log(np.clip(p, 1e-300, None)) + (n - kk) * np.log(np.clip(1 - p,
    1e-300, None))
    post = np.exp(logpost - logpost.max())
    post /= post.sum() * dp                      # flat prior: posterior = normalised
    likelihood
    mean = np.sum(p * post) * dp
    sd = np.sqrt(np.sum((p - mean) ** 2 * post) * dp)
    a2.plot(p, post, color=SERIES[i], label=f"$n={n}$, $k={kk}$")
    out[f"PipeK{names[n]}"] = kk
    out[f"PipeSd{names[n]}"] = sd
    out[f"PipeSdTh{names[n]}"] = np.sqrt((kk / n) * (1 - kk / n) / n)
```

The first listing evaluates (0.7) toss by toss for two sequences with 8 heads each, HHHHHHHHTT and THHTHHHHHH, on a grid of 2001 values of p. The largest difference between the two likelihood curves is 3.469×10^{-18}. That is floating-point round-off, so the curves are the same function, as sufficiency says.

The second listing computes the posterior on the grid. It works in logarithms, because a product of a thousand probabilities underflows. It subtracts the maximum before exponentiating, and then normalises numerically; that sum over the grid is Bayes' denominator. One simulated run with true $p = 0.8$ gives:

- at $n = 10$, $k = 7$, posterior standard deviation 0.1307, against the large-n formula $\sqrt{\hat{p}(1-\hat{p})/n} = 0.1449$;
- at $n = 100$, $k = 76$, 0.04238 against 0.04271;

- at $n = 1000$, $k = 819$, 0.01217 against 0.01218.

At $n = 10$ (here $k = 7$) the grid reproduces the exact value of Example 0.7,

$$\sqrt{\frac{(k+1)(n-k+1)}{(n+2)^2(n+3)}} = 0.131.$$

The large-n formula becomes accurate only as n grows.

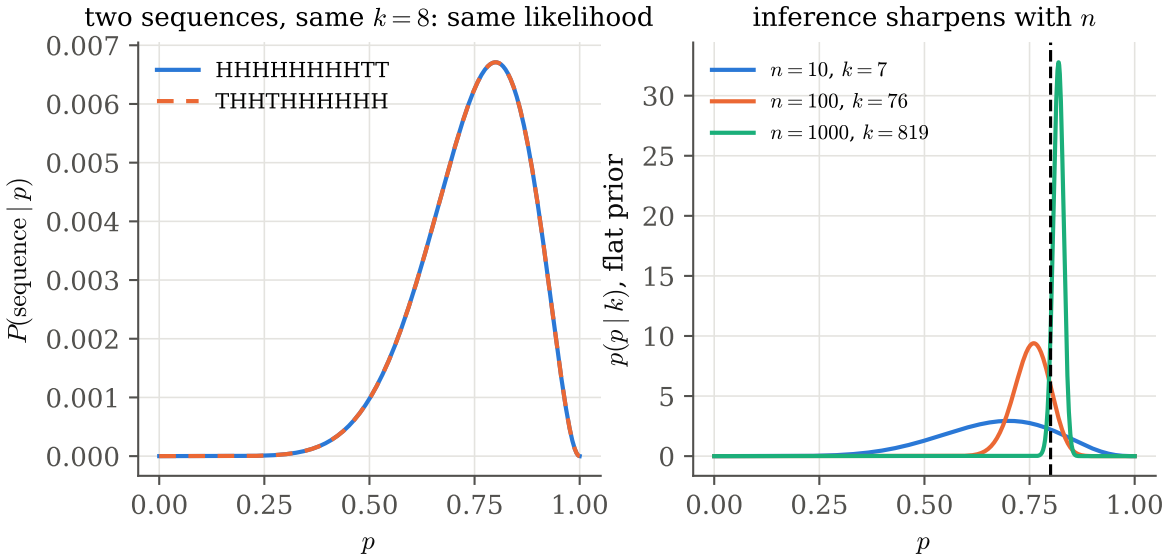


Figure 0.7: Left: the likelihoods of two different toss sequences with the same number of heads lie exactly on top of each other, so the summary k lost nothing. Right: the flat-prior posterior for p after 10, 100 and 1000 tosses of one coin with true $p = 0.8$ (dashed line). The width shrinks like $1/\sqrt{n}$.

How research uses the pipeline. In current research the hard stages are the middle ones. For the coin and the full-sky Gaussian CMB every stage has a closed form, but real problems rarely do. Which summary of a galaxy survey carries the neutrino mass: the power spectrum $P(k)$, the bispectrum, void counts, Betti numbers? What is $P(S \mid \boldsymbol{\theta})$ when S is measured on a masked, noisy, nonlinear field? Usually nobody can write it down. So one simulates the forward model many times at fixed $\boldsymbol{\theta}$ and estimates the mean and covariance of S from the simulations (Chapters 8 and 12). The inference stage then uses the standard machinery of Chapters 5, 9 and 10.

Watch out: the posterior depends on the summary

If S is not sufficient, $P(\boldsymbol{\theta} \mid S)$ is broader than $P(\boldsymbol{\theta} \mid \boldsymbol{d})$, and it can be broader in a direction that matters. That is not a bug of Bayes' theorem; it is the price of the compression chosen at stage 4. Two analyses of the same data with different summaries can legitimately give different constraints.

Kruschke's five steps. The pipeline is the physicist's version of the five steps of Bayesian data analysis [Kruschke p. 25]. They match the six stages one by one. *Identify the data relevant to the research question*: stages 2 and 3, the dataset and the choice of what information matters. *Define a descriptive model for the data*: stages 1 and 5, the model and the sampling distribution. *Specify*

a prior: the factor $P(\boldsymbol{\theta})$ of stage 6. *Reallocate credibility across parameter values*: the posterior of stage 6. *Check the posterior predictions*: the dashed loop of Figure 0.6.

0.4 Which summary? From amplitude to topology

Two simulated universes are handed to us as density maps. In one, the dense matter forms a connected web of filaments. In the other, it sits in isolated clumps. By construction both maps have the same mean, the same variance and the same power spectrum $P(k)$. An analysis that compresses each map to $P(k)$ will call them identical, yet a glance tells them apart.

Whether we catch this difference or miss it is decided at the third stage of the pipeline of Section 0.3: *deciding which features carry the physics*. The power spectrum answers one question: how much variance does each scale carry? The two maps raise a different one. Is there physically meaningful information in the way the high and low regions of a field are connected, information that the summary we already use, here $P(k)$, does not bring out? A field admits several kinds of summary, and two fields can be built to agree on some kinds and differ on the others. Topological summaries are worth the trouble for a second reason too: some of them are protected against the distortions that observation introduces.

A recording with its phases scrambled. Take a recording of speech, $h(t)$, and Fourier transform it to $\tilde{h}(f)$. Keep every amplitude $|\tilde{h}(f)|$ and replace every phase by a random one. The power spectrum is untouched, and so is the loudness, yet the result is a hiss. What made it speech was in the phases: in *where* and *how* the power is arranged in time. For a field, the phases decide whether high values line up into filaments, gather into blobs or scatter as noise. Summaries beyond the power spectrum are ways of reading the phases.

The central question for a field. A theory predicts a field $\phi(\mathbf{x})$ over space. The experiment gives us one copy of it, and an imperfect one. On the way to us its values have been pushed around by perturbations, by noise, by smoothing, by whatever happened to the signal while it travelled. The question is whether something about the *shape and structure* of what we receive survives every change we count as a nuisance. If it does, we can read it off without first undoing those changes. Write such a change as a **nuisance transformation** $\phi \to \phi' = \mathcal{N}[\phi]$; examples are noise, smoothing by a beam, lensing, and a monotonic response of the detector. Three questions follow.

1. What do we want to know: how large the field is, how its values at different points are related, or how its structures are shaped and connected?
2. Which summary T answers that question?
3. Is there a summary that the nuisance cannot change, $T[\mathcal{N}[\phi]] = T[\phi]$, for a stated class of $\mathcal{N}$?

The answers to the first question form a ladder. Each rung up keeps less of the raw values and more of how they are arranged.

A *Gaussian* random field is special: all its information is in the first two rungs. Its higher-order summaries are fixed functions of its power spectrum. Measuring them is therefore a consistency test (a search for non-Gaussianity), not new information. For any non-Gaussian field, from the late-time cosmic web to a map with foreground residuals, the upper rungs carry information that $P(k)$ does not.

Taxonomy: the summaries of a field

kind	the question it answers	examples
amplitude	how large is the field typically?	$\langle\phi\rangle$, Var(ϕ), the one-point pdf $P(\phi)$
correlation	how are values a distance r apart related?	$\xi(r)$, $P(k)$, C_ℓ
higher order	is there information beyond Gaussian two-point statistics?	bispectrum B, trispectrum T, skewness
morphology	what sizes and shapes do the structures have?	areas, perimeters, curvature of excursion sets
topology	how are the structures connected, and how many holes do they have?	Betti numbers β_i, Euler characteristic χ, persistence

These are not rival theories of probability. They are different maps from the same data to different summaries $S[\phi]$. Once S is chosen it is an ordinary random observable with a sampling distribution $P(S \mid \boldsymbol{\theta})$, and Chapters 3 to 5, 9 and 10 apply to it unchanged.

Topology is a property of shapes, not of functions. So to say what topology measures, we must first turn a field into a shape.

Working definition 0.4 (excursion set, Betti numbers, Euler characteristic)

For a threshold ν, the **excursion set** of ϕ on a domain $\mathcal{M}$ is

$$X_\nu = \{\mathbf{x} \in \mathcal{M} : \ \phi(\mathbf{x}) > \nu\} \tag{0.8}$$

[the region of the map where the field exceeds ν, e.g. the sky directions hotter than ν]. In two dimensions the **Betti numbers** are $\beta_0(X_\nu)$ [the number of separate pieces] and $\beta_1(X_\nu)$ [the number of independent holes, i.e. enclosed regions that are not part of the set]. The **Euler characteristic** is $\chi = \beta_0 - \beta_1$. These numbers do not change when the set is stretched or bent without tearing or gluing (Chapter 12). A summary T is a **robust observable** under a class of nuisance transformations if $T[\mathcal{N}[\phi]] = T[\phi]$ for every $\mathcal{N}$ in the class.

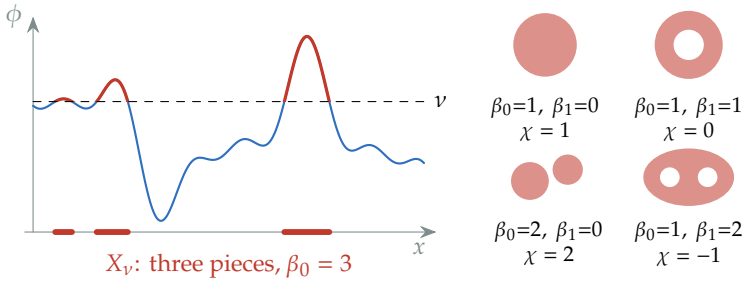


Figure 0.8: Left: a field along a line, a threshold ν, and the excursion set (the red intervals on the axis), which here has three pieces. Right: four two-dimensional sets with their numbers of pieces β_0, holes β_1 and Euler characteristic $\chi = \beta_0 - \beta_1$. Stretching any of them leaves all three numbers unchanged; cutting or gluing does not.

Two facts make these summaries useful. First, a change of phases can leave the lower rungs untouched while moving the upper ones. Second, a distortion that changes each value by the same increasing function leaves the excursion sets themselves untouched. Each takes a few lines to prove.

Example 0.9 (phase randomisation keeps the power spectrum and the variance). The question: if we keep every Fourier amplitude of a field and scramble the phases, what stays the same? Let $\phi(\mathbf{x})$ live on an N-pixel periodic grid with zero mean, and let $\tilde{\phi}_{\mathbf{k}} = \sum_{\mathbf{x}} \phi(\mathbf{x}) e^{-2\pi i \mathbf{k}\cdot\mathbf{x}/L}$ be its discrete Fourier transform (L the side length in pixels). Build a new field ϕ' with $\tilde{\phi}'_{\mathbf{k}} = |\tilde{\phi}_{\mathbf{k}}| e^{i\alpha_{\mathbf{k}}}$, random phases $\alpha_{\mathbf{k}}$ with $\alpha_{-\mathbf{k}} = -\alpha_{\mathbf{k}}$ (so that ϕ' is real). Then

$$P'(k) \overset{(1)}{=} \left\langle |\tilde{\phi}'_{\mathbf{k}}|^2 \right\rangle_{|\mathbf{k}|=k} \overset{(2)}{=} \left\langle |\tilde{\phi}_{\mathbf{k}}|^2 \right\rangle_{|\mathbf{k}|=k} = P(k),$$
$$\mathrm{Var}(\phi') \overset{(3)}{=} \frac{1}{N} \sum_{\mathbf{x}} \phi'(\mathbf{x})^2 \overset{(4)}{=} \frac{1}{N^2} \sum_{\mathbf{k}} |\tilde{\phi}'_{\mathbf{k}}|^2 \overset{(5)}{=} \frac{1}{N^2} \sum_{\mathbf{k}} |\tilde{\phi}_{\mathbf{k}}|^2 = \mathrm{Var}(\phi).$$

(1) The measured power spectrum: average $|\tilde{\phi}_{\mathbf{k}}|^2$ over the wavevectors in a shell of radius k.
(2) $|re^{i\alpha}|^2 = r^2$: the phase drops out of every $|\tilde{\phi}'_{\mathbf{k}}|^2$ individually.
(3) Zero mean (the $\mathbf{k} = 0$ coefficient, the sum of the field, is left at zero).
(4) Parseval's theorem for the discrete Fourier transform, $\sum_{\mathbf{x}} |\phi(\mathbf{x})|^2 = \frac{1}{N} \sum_{\mathbf{k}} |\tilde{\phi}_{\mathbf{k}}|^2$.
(5) Same step as (2).

So every summary built from the amplitudes $|\tilde{\phi}_{\mathbf{k}}|$ alone is identical for ϕ and ϕ'. The skewness, the excursion sets and their Betti numbers depend on the phases, so they are free to change.

Example 0.10 (a monotonic distortion leaves the excursion sets alone). Suppose the detector does not record ϕ itself but $\phi'(\mathbf{x}) = f(\phi(\mathbf{x}))$, with f strictly increasing. Examples are a detector with a nonlinear but monotonic response, or a lognormal model $\phi' = e^{\phi}$. The question: which summaries of ϕ can we still read off ϕ' without knowing f? Compare the excursion sets. For every threshold ν

$$X'_{f(\nu)} \overset{(1)}{=} \{\mathbf{x} : \ f(\phi(\mathbf{x})) > f(\nu)\} \overset{(2)}{=} \{\mathbf{x} : \ \phi(\mathbf{x}) > \nu\} = X_\nu.$$

(1) Definition (0.8) applied to ϕ' at the threshold $f(\nu)$.
(2) A strictly increasing f preserves order both ways: $f(a) > f(b)$ exactly when $a > b$.

The excursion sets are the same regions of the domain $\mathcal{M}$. So the curves $\beta_0(\nu)$, $\beta_1(\nu)$, $\chi(\nu)$ are the same, except that the threshold is relabelled $\nu \to f(\nu)$. Meanwhile the variance, the skewness and $P(k)$ of ϕ' all change; for $f = \exp$ the one-point pdf goes from Gaussian to lognormal. Here topology is a robust observable in the exact sense of the working definition above, $T[\mathcal{N}[\phi]] = T[\phi]$; the power spectrum is not.

Example 0.11 (counting holes by counting cells). How do we count holes in a real map, where nobody can look at every piece? On a pixel grid a set is a union of square pixels, and χ can be computed by counting: $\chi = V - E + F$. Here V is the number of corners, E the number of pixel edges and F the number of pixels of the set, each counted once. A single pixel has $4 - 4 + 1 = 1$. Two separate pixels have $8 - 8 + 2 = 2$. Now take a ring of eight pixels around an empty centre (a 3×3 block with the middle removed). It uses all $4 \times 4 = 16$ corners of the block, 12 horizontal and 12 vertical edges, and 8 pixels:

$$\chi = 16 - 24 + 8 = 0 = \beta_0 - \beta_1 = 1 - 1.$$

So local counting finds a hole without ever looking for it. That is how $\chi(\nu)$ is computed on real maps (Chapter 12).

What this bought us. The ladder of summaries is a ladder of questions. Two fields can agree on the bottom two rungs and differ above them. And some topological summaries are exactly unchanged by a class of distortions that change every rung below.

Two fields with one power spectrum

The script `04_same_spectrum.py` builds the two universes from the opening of Section 0.4: one clumpy, one not, with the same power spectrum. Field A drops 60 points at random on a 256×256 periodic grid and smooths them into blobs. Field B is A with its phases randomised, exactly as in Example 0.9:

Listing 0.5: Keeping every $|\tilde{\phi}_k|$ and replacing the phases (`code/ch00/04_same_spectrum.py`)

```
# ---- field B: keep every |A_k|, replace the phases by those of white noise ----
#      rfft2 of a real white-noise map has exactly the Hermitian symmetry a real field needs
Ak = np.fft.rfft2(A)
phase = np.exp(1j * np.angle(np.fft.rfft2(rng.standard_normal((NPIX, NPIX)))))
Bk = np.abs(Ak) * phase
Bk[0, 0] = Ak[0, 0]                                  # keep the (zero) mean
B = np.fft.irfft2(Bk, s=A.shape)
```

The random phases are taken from the Fourier transform of a real white-noise map. Such phases automatically satisfy the condition $\alpha_{-\mathbf{k}} = -\alpha_{\mathbf{k}}$, so B is real. Pieces are counted with the function `scipy.ndimage.label`, which labels connected groups of pixels above the threshold.

The run shows the lower rungs agreeing and the upper ones not. The variances are $\mathrm{Var}(A) = 1$ and $\mathrm{Var}(B) = 1$, and the two binned power spectra differ by at most a fraction 1.500×10^{-12}, which is round-off. The skewness separates the fields: 2.455 for the clumps of A, 0.02076 for B, which is close to Gaussian. The topology separates them more sharply. At $\nu = 3$ (three standard deviations) A has $\beta_0 = 51$ separate pieces, one per surviving blob, and B has 3. At $\nu = 0$ the counts are 34 and 19.

Topology that survives observation, and topology that records dynamics

The detector with a monotonic response (Example 0.10) is the simplest protected case. A more physical one is CMB polarisation. The linear polarisation is described by the two Stokes parameters Q and U, which combine into $P = Q + iU$. At isolated points where $Q = U = 0$ the polarisation direction is undefined. Walk once round such a **polarisation singularity**, and the phase of P turns by a whole multiple of 2π. This whole number, the winding index or charge, cannot change under a smooth one-to-one remapping of the sky such as weak lensing. Lensing certainly changes the polarisation map, converting E modes into B modes. But it can only move and distort singularities; it cannot change their charges continuously [1–3]. The protection fails where its assumptions fail: pairs of singularities can be created or annihilated, and noise and finite resolution can hide them.

A count at a single threshold is fragile, since nothing makes $\nu = 1\sigma$ special. The remedy is to follow each piece and hole as ν is lowered. Record the threshold where each feature appears (its birth) and the one where it merges or fills in (its death). This gives **persistence**. A feature with a long lifetime $\Delta\nu = \nu_{\mathrm{death}} - \nu_{\mathrm{birth}}$ is a real structure; a short-lived one is likely noise. Chapter 12 builds this with a union–find algorithm.

Watch out: topology is not conserved by dynamics

Invariance holds under a stated class of deformations, not under physics in general. Gravitational collapse is not a monotonic remapping of the initial density: it moves matter, merges clumps and opens voids, and the Betti numbers of the cosmic web change with time [4, 5]. There the changing topology is the *signal*. The scientific question is either "what survives the nuisance?" or "what changes because of the physics?". Both are legitimate; they are different questions. Note also that a mask cuts excursion sets at its edge and so

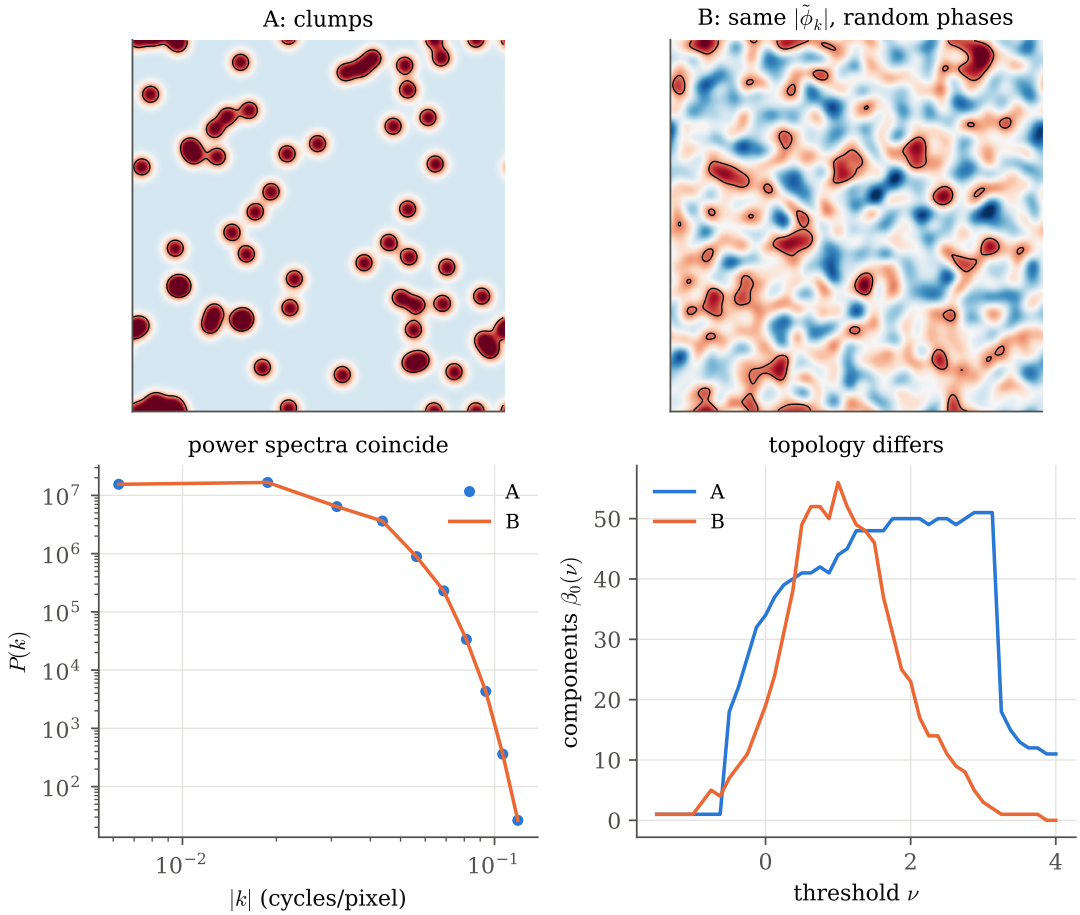


Figure 0.9: Top: the clumpy field A and its phase-randomised twin B, with the contour at $\nu = 1.5$. Bottom left: their power spectra coincide point by point. Bottom right: the number of connected pieces of the excursion set as the threshold is raised. B behaves like a Gaussian field, with most pieces near $\nu \approx 1$ and few at high threshold; A keeps one piece per blob right up to the blob peaks, so its curve stays high until $\nu \approx 3$.

changes their topology, just as it changes the covariance of $\widehat{C}_\ell$ (Chapter 8).

How research uses a topological summary. A CMB map is a function $T : S^2 \to \mathbb{R}$ from directions on the sphere to temperatures. To use its topology, generate many Gaussian skies from the fiducial power spectrum C_ℓ. Pass each sky through the beam, the noise and the mask. Measure $\beta_0(\nu)$, $\beta_1(\nu)$ and $\chi(\nu)$ on every realisation. The mean and covariance of these curves over the realisations define a likelihood for the observed curves, and the posterior follows as for any other summary. This is the Monte Carlo study of Chapter 8 with $\widehat{C}_\ell$ replaced by a Betti curve, and it is where the book ends (Chapter 12).

0.5 One idea, several detectors

A collider experiment counts events in bins of invariant mass and asks whether there is a bump. A CMB satellite measures the temperature of the sky and asks for the cosmological parameters. A gravitational-wave interferometer records a strain time series and asks for the masses of an inspiralling binary. The instruments share nothing, and the statistics looks different in each

community's papers: Poisson log-likelihoods, C_ℓ likelihoods, matched filters. Underneath, all three ask one question: *for a given set of parameters $\boldsymbol{\theta}$, how probable is the noise realisation that would turn the model's prediction into the data we actually recorded?* Because the question is the same, the idea needs to be developed only once. Writing it out for each detector, and for a topological summary, shows how it applies everywhere.

The noise you would need. The likelihood is the probability density of the noise that the data require. Put the model prediction $m(\boldsymbol{\theta})$ on one side and the data $\boldsymbol{d}$ on the other. The difference $\boldsymbol{d} - m(\boldsymbol{\theta})$ is the noise that nature *would have had to supply* if $\boldsymbol{\theta}$ were the truth. A value of $\boldsymbol{\theta}$ that needs a wildly improbable noise realisation is implausible. A value that needs only a typical fluctuation is plausible. What differs between detectors is only the statistics of their noise. Chapter 3 builds the likelihood from this picture one step at a time.

The same four questions for any detector.

1. What does the model predict, as a function of $\boldsymbol{\theta}$?
2. What is the noise: its distribution, and its correlations?
3. For a trial $\boldsymbol{\theta}$, how probable is the noise realisation $\boldsymbol{d} - m(\boldsymbol{\theta})$ (or, for counts, the observed counts given their expectation)?
4. Read that probability as a function of $\boldsymbol{\theta}$, and feed it to Bayes, MCMC or Fisher.

Working definition 0.5 (likelihood from a noise model)

If the data are prediction plus noise, $\boldsymbol{d} = m(\boldsymbol{\theta}) + \boldsymbol{\epsilon}$, and the noise has a known density p_ϵ, then

$$\mathcal{L}(\boldsymbol{\theta}) \equiv p(\boldsymbol{d} \mid \boldsymbol{\theta}) = p_\epsilon\big(\boldsymbol{d} - m(\boldsymbol{\theta})\big) \tag{0.9}$$

[the density of the noise realisation that is required, evaluated at the observed data and read as a function of $\boldsymbol{\theta}$]. The **likelihood** is not a probability distribution over $\boldsymbol{\theta}$: it need not integrate to one over $\boldsymbol{\theta}$, and only its shape and ratios matter.

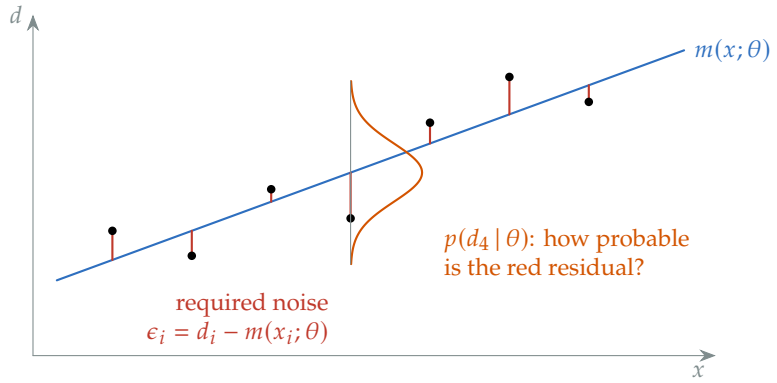


Figure 0.10: Data as prediction plus noise. The blue line is the model for one trial value of θ; the black points are the data; the red segments are the noise each point would need if that θ were true. The orange curve at the fourth point is the Gaussian noise density: a long red segment lands in its tail and costs a lot of probability. The likelihood multiplies these costs over all points; moving θ moves the line and changes every red segment.

The commonest case is independent Gaussian noise: each data point d_i has its own noise $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$, a Gaussian with mean zero and standard deviation σ_i. Then (0.9) gives the workhorse of physics:

$$\ln \mathcal{L}(\boldsymbol{\theta}) \overset{(1)}{=} \ln \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\Big[- \frac{(d_i - m_i(\boldsymbol{\theta}))^2}{2\sigma_i^2} \Big]$$
$$\overset{(2)}{=} -\frac{1}{2} \sum_{i=1}^{N} \frac{(d_i - m_i(\boldsymbol{\theta}))^2}{\sigma_i^2} - \frac{1}{2} \sum_{i=1}^{N} \ln(2\pi\sigma_i^2) \equiv -\frac{1}{2}\chi^2(\boldsymbol{\theta}) + \text{const.} \tag{0.10}$$

(1) Independent noise: the joint density is the product of one Gaussian per data point, evaluated at the required noise $d_i - m_i(\boldsymbol{\theta})$.
(2) The log of a product is the sum of logs, and $\ln e^u = u$.

If the noise in different data points is correlated, with covariance matrix $\mathbf{C}$, the product of separate Gaussians becomes one multivariate Gaussian: $\ln \mathcal{L} = -\frac{1}{2}(\boldsymbol{d} - m)^\mathsf{T} \mathbf{C}^{-1} (\boldsymbol{d} - m) - \frac{1}{2} \ln \det(2\pi\mathbf{C})$. The statistical logic is the same as before. Only the model for the probability of the data has changed.

With a whole data vector and many parameters, the backward question of Section 0.1 is still the same. The data vector is what we saw. Every value of $\boldsymbol{\theta}$ is a candidate explanation. The likelihood says how readily each candidate would have produced those data (Chapter 5). Reading the likelihood as the plausibility of the fluctuation needed to turn the model into the data is also how maximum likelihood is motivated [Wasserman p. 122].

Taxonomy: one question, four implementations

experiment	data	noise model	log-likelihood
collider	counts n_i in bins	Poisson with mean $\lambda_i(\mu) = b_i + \mu s_i$	$\sum_i [n_i \ln \lambda_i - \lambda_i]$
CMB	$a_{\ell m}$, or $\widehat{C}_\ell$	Gaussian random field on the sphere, variance C_ℓ per mode	$-\frac{1}{2} \sum_\ell (2\ell+1) \big[\widehat{C}_\ell / C_\ell + \ln C_\ell \big]$
GW detector	strain $d(t)$	stationary Gaussian noise with PSD $S_n(f)$	$-\frac{1}{2}(d - h \mid d - h)$
topology	Betti curves $S = \beta_0(\nu), \ldots$	unknown in closed form: simulated	$-\frac{1}{2}(S - \bar{S})^\mathsf{T} \widehat{\mathbf{C}}^{-1} (S - \bar{S})$ from simulations

Example 0.12 (the collider: a bump in Poisson counts). The data are event counts n_i in bins i of invariant mass. The question: how probable are these counts for a given signal strength μ? In bin i the expected count is $\lambda_i(\mu) = b_i + \mu s_i$: the background b_i plus a signal template s_i scaled by μ. Counts in different bins are independent Poisson variables. Then

$$\ln \mathcal{L}(\mu) \overset{(1)}{=} \ln \prod_i \frac{\lambda_i^{n_i} e^{-\lambda_i}}{n_i!} \overset{(2)}{=} \sum_i \big[n_i \ln \lambda_i(\mu) - \lambda_i(\mu) \big] - \sum_i \ln n_i!,$$
$$\frac{\mathrm{d} \ln \mathcal{L}}{\mathrm{d}\mu} \overset{(3)}{=} \sum_i s_i \Big(\frac{n_i}{\lambda_i(\mu)} - 1 \Big) = 0 \quad \text{at the maximum-likelihood } \hat{\mu}.$$

(1) Poisson probability of n_i events when λ_i are expected (Chapter 2), multiplied over independent bins.

(2) Log of a product; the last sum does not depend on μ and drops out of every ratio.
(3) $\mathrm{d}\lambda_i/\mathrm{d}\mu = s_i$; the chain rule on $n_i \ln \lambda_i$ gives $n_i s_i/\lambda_i$.

The condition says that at the best fit the observed counts match the expected ones on average, with each bin weighted by its signal s_i. There is no σ_i anywhere. The Poisson distribution carries its own noise: its variance equals its mean.

Example 0.13 (the CMB: the full-sky C_ℓ likelihood). The data are the harmonic coefficients of the sky (Example 0.8); the parameters are the power spectrum values C_ℓ. The question: how probable is the observed sky for given C_ℓ? Write the sky in a real orthonormal basis of spherical harmonics, so that at each multipole ℓ there are $2\ell+1$ real coefficients $r_{\ell j}$. For a Gaussian isotropic sky they are independent, and each is $\mathcal{N}(0, C_\ell)$. Then

$$\begin{aligned}\ln \mathcal{L}(\{C_\ell\}) &\overset{(1)}{=} \sum_\ell \sum_{j=1}^{2\ell+1} \Big[-\frac{r_{\ell j}^2}{2C_\ell} - \frac{1}{2}\ln(2\pi C_\ell)\Big] \\ &\overset{(2)}{=} -\frac{1}{2}\sum_\ell \Big[\frac{(2\ell+1)\widehat{C}_\ell}{C_\ell} + (2\ell+1)\ln C_\ell\Big] + \text{const} \\ &= -\frac{1}{2}\sum_\ell (2\ell+1)\Big[\frac{\widehat{C}_\ell}{C_\ell} + \ln C_\ell\Big] + \text{const}.\end{aligned}$$

(1) (0.10) with $d_i \to r_{\ell j}$, zero mean and variance C_ℓ.
(2) $\sum_j r_{\ell j}^2 = (2\ell+1)\widehat{C}_\ell$, the definition of $\widehat{C}_\ell$ in Example 0.8 (a real orthonormal basis is a rotation of the complex one, so the sum of squares is unchanged); $\ln 2\pi$ goes into the constant.

Three things follow. First, the data enter only through $\widehat{C}_\ell$, so $\widehat{C}_\ell$ is sufficient in the sense of (0.6), as claimed in Example 0.8. Second, setting $\partial \ln \mathcal{L}/\partial C_\ell = \frac{1}{2}(2\ell+1)[\widehat{C}_\ell/C_\ell^2 - 1/C_\ell] = 0$ gives $C_\ell = \widehat{C}_\ell$: the measured spectrum is the best fit. Third, $(2\ell+1)\widehat{C}_\ell/C_\ell = \sum_j (r_{\ell j}/\sqrt{C_\ell})^2$ is a sum of $2\ell+1$ squared unit Gaussians, which is a $\chi^2_{2\ell+1}$ variable (Chapter 2). That is how the script below simulates one sky's spectrum without making a map.

Example 0.14 (the GW detector: Gaussian noise with a PSD). The detector records a strain $d(t) = h(t;\boldsymbol{\theta}) + n(t)$ for a time T: a signal h from the binary with parameters $\boldsymbol{\theta}$, plus noise n. The question: how probable is the required noise $d - h$ for given $\boldsymbol{\theta}$? The noise is simplest in frequency space, so Fourier transform, $\tilde{d}(f) = \int d(t) e^{-2\pi i f t}\mathrm{d}t$, at the frequencies $f_k = k/T$, and write $\tilde{d}_k = \tilde{d}(f_k)$. Stationary Gaussian noise has independent Fourier modes. Its one-sided power spectral density $S_n(f)$ [the noise power per unit frequency, counting positive frequencies only] fixes their size, $\mathbb{E}\,|\tilde{n}_k|^2 = \frac{T}{2} S_n(f_k)$. We also need one fact about complex Gaussians. A complex Gaussian z with $\mathbb{E}\,|z|^2 = \sigma^2$ has density $\frac{1}{\pi\sigma^2} e^{-|z|^2/\sigma^2}$ (its real and imaginary parts are independent, each with variance $\sigma^2/2$). So

$$\begin{aligned}\ln \mathcal{L}(\boldsymbol{\theta}) &\overset{(1)}{=} -\sum_k \frac{|\tilde{d}_k - \tilde{h}_k(\boldsymbol{\theta})|^2}{\frac{T}{2}S_n(f_k)} + \text{const} \overset{(2)}{=} -2\sum_k \Delta f\, \frac{|\tilde{d}_k - \tilde{h}_k|^2}{S_n(f_k)} + \text{const} \\ &\overset{(3)}{=} -2\int_0^\infty \frac{|\tilde{d}(f) - \tilde{h}(f)|^2}{S_n(f)}\,\mathrm{d}f + \text{const} \overset{(4)}{=} -\frac{1}{2}\,(d-h \,|\, d-h) + \text{const},\end{aligned}$$

(1) The required noise in mode k is $\tilde{d}_k - \tilde{h}_k$; the product of independent complex Gaussians becomes a sum in the log. The $-\ln(\pi\sigma^2)$ terms do not depend on $\boldsymbol{\theta}$.

(2) The mode spacing is $\Delta f = 1/T$, so $\frac{1}{T/2} = 2\Delta f$.
(3) For long T the sum over modes becomes an integral over positive frequencies.
(4) Definition of the noise-weighted inner product $(a \mid b) = 4\,\mathrm{Re}\int_0^\infty \tilde{a}(f)\tilde{b}^*(f)/S_n(f)\,\mathrm{d}f$.

This is the χ^2 of (0.10) in the frequency domain, with each frequency weighted by how noisy the detector is there. Maximising it over $\boldsymbol{\theta}$ is matched filtering. Its curvature gives the Fisher matrix of Chapter 10.

What this bought us. Poisson counts, a Gaussian sky and coloured detector noise give three formulas that look unrelated. Each one is (0.9): the probability of the fluctuation that would turn the prediction into the data.

The three detectors in code

The script `05_three_detectors.py` generates data for all three and checks the noise model of each. For the collider and the GW detector the key lines are:

Listing 0.6: Collider: Poisson data and a scan of $\ln \mathcal{L}(\mu)$ (`code/ch00/05_three_detectors.py`)

```
n_obs = rng.poisson(bkg + 1.0 * sig_shape)                    # the data: one Poisson draw per bin
mus = np.linspace(0, 3, 3001)
# ln L(mu) = sum_i [ n_i ln(lambda_i) - lambda_i ]  (+ const, the ln n_i! term)
lam = bkg[None, :] + mus[:, None] * sig_shape[None, :]
lnL = np.sum(n_obs * np.log(lam) - lam, axis=1)
i_best = np.argmax(lnL)
inside = mus[lnL >= lnL.max() - 0.5]                          # Delta lnL = 1/2 interval
```

Listing 0.7: GW: coloured Gaussian noise drawn in the frequency domain (`code/ch00/05_three_detectors.py`)

```
sd = np.sqrt(T * Sn / 4)                                      # Re and Im each get half of T S_n/2
n_k = sd * (rng.standard_normal(f.size) + 1j * rng.standard_normal(f.size))
h_k = np.zeros_like(n_k); f_sig = 40.0
h_k[np.argmin(abs(f - f_sig))] = 12 * np.sqrt(T * Sn[np.argmin(abs(f - f_sig))])   # a weak line
d_k = n_k + h_k
pgram = 2 * np.abs(d_k) ** 2 / T                              # periodogram: E[pgram] = S_n off the line
```

Collider. The first block draws one Poisson count per bin. The expected counts are a falling background plus a peak at 125 GeV with $\mu = 1$. It then evaluates the $\ln \mathcal{L}$ of Example 0.12 on a grid of μ. With 4769 events in total, the maximum is at $\hat{\mu} = 0.959$. The values of μ whose $\ln \mathcal{L}$ lies within $\frac{1}{2}$ of the maximum run from 0.805 to 1.116; for a nearly Gaussian likelihood this is the $\pm 1\sigma$ interval (Chapter 3). The true $\mu = 1$ is inside.

GW detector. The second block draws each Fourier mode with real and imaginary parts of variance $TS_n/4$, so that $\mathbb{E}\,|\tilde{n}_k|^2 = TS_n/2$. The periodogram $2|\tilde{d}_k|^2/T$ should then scatter around S_n. Averaged over 8192 frequency bins away from the injected line, its ratio to S_n is 1.006.

CMB. One full sky is simulated as $\widehat{C}_\ell = C_\ell\, \chi^2_{2\ell+1}/(2\ell+1)$ around the Planck best fit, using the third result of Example 0.13. The mean of $\widehat{C}_\ell/C_\ell$ over $2 \le \ell \le 1500$ is 1.002. Its scatter is 0.2324 for $\ell \le 30$ but only 0.02923 for $\ell > 1000$, as the cosmic-variance formula $\sqrt{2/(2\ell+1)}$ predicts.

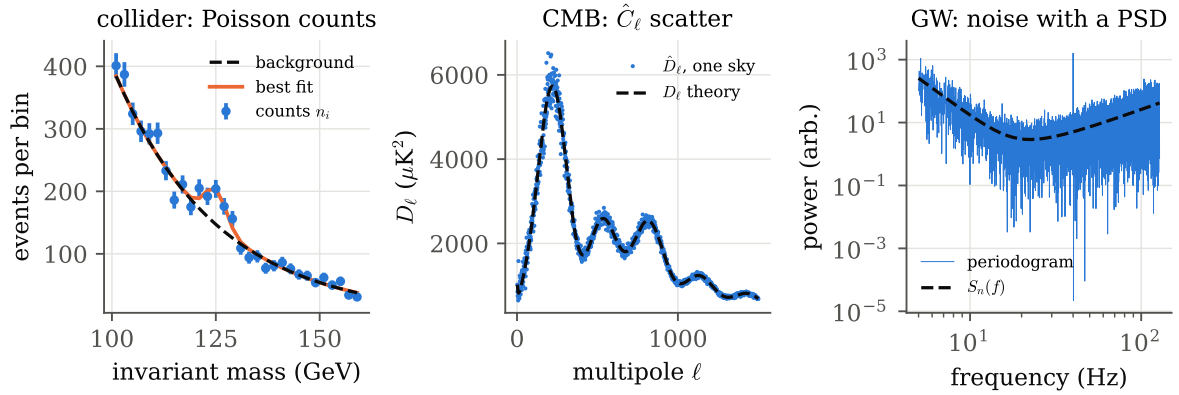


Figure 0.11: The same logic in three instruments. Left: Poisson counts with $\sqrt{n}$ error bars over a falling background, and the best-fit background plus signal. Middle: one simulated sky's $\hat{D}_\ell$ scattering around the theory curve; the scatter is large at low ℓ, where there are few modes, and small at high ℓ. Right: the periodogram of simulated detector noise, wildly fluctuating from bin to bin but centred on the power spectral density.

How research uses these likelihoods. Each simulation here verifies a known noise model: Poisson counts, χ^2 scatter of $\widehat{C}_\ell$, a periodogram centred on S_n. In research the closed forms break. A galactic mask couples multipoles and makes the C_ℓ likelihood non-diagonal. Real interferometer noise has glitches and is not stationary. A Betti curve has no known distribution at all. The fix is always the fourth row of the taxonomy (topology): simulate the forward model, estimate the mean and covariance of the summary from the simulations, and use the Gaussian form. The price is that the estimated covariance $\widehat{\mathbf{C}}$ is itself noisy and must be corrected [6] (Chapter 8).

Watch out: the likelihood is not the posterior

$\mathcal{L}(\boldsymbol{\theta}) = p(\boldsymbol{d} \mid \boldsymbol{\theta})$ is a function of $\boldsymbol{\theta}$, but it is not a density in $\boldsymbol{\theta}$. Here is a test that shows the difference. Change variables from C_ℓ to $\ln C_\ell$. The likelihood's values stay the same, but a density would pick up a Jacobian factor. Only after multiplying by a prior and normalising (Chapter 5) do we get a probability distribution for $\boldsymbol{\theta}$.

0.6 How to run the code

Months from now we will change one line in the Monte Carlo of Chapter 8. A quoted covariance will then move, and we will want to know why: because of the change, or because the random numbers were different this time? That question has an answer only if two things hold. Every simulation in the book must give the same numbers, bit for bit, when it is rerun. And every number in the text must be traceable to the script that printed it. A small amount of machinery guarantees both, so that one command regenerates every figure and every simulated number.

A lab notebook with a calibration source. An experimentalist keeps a notebook that records which run produced which plot, and keeps a calibration source that gives the same signal every time it is measured. The code does both. Each script is one notebook entry: it writes its figure and its numbers to files named after itself. And each script draws its random numbers from a generator seeded by its own name, a calibration source that reproduces the same "noise" on every run.

What we need from the code.

1. Where does each script live, and how is it run?
2. Where do its figures and numbers go, and how do they get into the text?
3. How are random numbers made reproducible, and different scripts independent?
4. How do all figures share one visual style?

```
Notes/
├── main.tex, notes.sty          the book and its style
├── chapters/chNN/part.tex       the text of each chapter
├── code/common.py               shared helpers
├── code/chNN/NN_name.py         one script per result
├── figures/chNN/name.pdf        written by savefig
├── results/chNN/name.tex        written by save_numbers
├── data/                        cached expensive products (CAMB spectra)
└── tools/testpart.sh            build one part and check it
```

Figure 0.12: The layout of the Notes directory. Every script is run from the top directory, reads nothing but its own inputs, and writes its outputs into the `figures` and `results` folders of its own chapter.

Everything is Python with numpy, scipy, matplotlib and healpy, and every sampler in the book is written from scratch. The choice of language does not matter for the statistics; "any computing language can be used" [Wasserman p. ix]. Simulation makes a probability tangible: we watch the empirical frequency settle on the theoretical value, as in the right panel of Figure 0.5 [SeeingTheory pp. 5–7].

Every script is run from the Notes directory, for example

```
python3 code/ch00/02_coin_questions.py
```

and starts with a docstring stating the question it answers, what it computes and what it writes. All scripts share one module, `code/common.py`. Its first part sets the environment, the paths and the colours:

Listing 0.8: Environment, paths and the colour palette (`code/common.py`)

```
import hashlib
import os
import pathlib

# keep numerical libraries from oversubscribing the machine when several scripts run at once
for _v in ("OMP_NUM_THREADS", "OPENBLAS_NUM_THREADS", "MKL_NUM_THREADS"):
    os.environ.setdefault(_v, "4")

import matplotlib

matplotlib.use("Agg")
import matplotlib.pyplot as plt  # noqa: E402
import numpy as np  # noqa: E402

NOTES = pathlib.Path(__file__).resolve().parents[1]
FIG = NOTES / "figures"
RES = NOTES / "results"
# Seed test: SEED_SALT=1 (or 2, ...) shifts every seed and sends figures and numbers to
```

```
# build/seedtest_<salt>/ so the book's own results are untouched. Unset: the book's seeds.
SEED_SALT = os.environ.get("SEED_SALT", "")
if SEED_SALT:
    FIG = NOTES / "build" / f"seedtest_{SEED_SALT}" / "figures"
    RES = NOTES / "build" / f"seedtest_{SEED_SALT}" / "results"
DATA = NOTES / "data"          # cached expensive products (CAMB spectra, simulated maps, chains)

# Categorical order is fixed: slot 1 is always the first series, never cycled.
SERIES = ["#2a78d6", "#eb6834", "#1baf7a", "#eda100", "#e87ba4", "#008300", "#4a3aa7", "#e34948"]
INK = "#0b0b0b"          # primary text
INK2 = "#52514e"         # secondary text, axes
GRID = "#e4e3df"         # recessive grid
THEORY = "#0b0b0b"       # analytic / theoretical curve: black, dashed
SEQ = "Blues"            # sequential colormap (magnitude)
DIV = "RdBu_r"           # diverging colormap (correlations, residuals), centred on zero
```

Lines 24–25 cap the linear-algebra libraries at four threads, so several scripts can run side by side without fighting for cores. Line 29 selects a non-interactive plotting backend, so scripts run without a display. The paths are computed from the location of `common.py` itself. That is why a script finds `figures/` and `results/` whatever directory it is started from. Lines 36–41 are a switch for testing the book itself: setting `SEED_SALT` changes every seed and sends the figures and numbers to a separate folder, so a run with different random numbers can be compared with the book's own without overwriting it. `SERIES` is the fixed order of categorical colours: the first data series in every figure is the same blue, the second the same orange, and colours are never recycled. The function `setup()` (lines 54–80, not shown) applies one matplotlib style: serif fonts, a light grid, no top and right axes. The second part of the module holds the functions every script calls:

Listing 0.9: Figures, numbers, random streams and theory curves (`code/common.py`)

```
def savefig(fig, chapter: str, name: str):
    out = FIG / chapter
    out.mkdir(parents=True, exist_ok=True)
    path = out / f"{name}.pdf"
    fig.savefig(path)
    plt.close(fig)
    print(f"[figure] {path.relative_to(NOTES)}")
    return path


def _fmt(v):
    if isinstance(v, str):
        return v
    if isinstance(v, (int, np.integer)):
        return f"{int(v)}"
    v = float(v)
    if v != 0 and (abs(v) < 1e-3 or abs(v) >= 1e5):
        m, e = f"{v:.3e}".split("e")
        return rf"{m}\times10^{{{int(e)}}}"
    return f"{v:.4g}"


def save_numbers(chapter: str, name: str, values: dict):
    """Write LaTeX macros. Macro names must be letters only (LaTeX rule), e.g. 'VarRatioL'.

    Values are typeset in math mode by the chapter:  $\\VarRatioL$.
    """
    out = RES / chapter
    out.mkdir(parents=True, exist_ok=True)
```

```
    path = out / f"{name}.tex"
    lines = [f"% generated by code/{chapter}/{name}.py -- do not edit"]
    for k, v in values.items():
        assert k.isalpha(), f"LaTeX macro names must be letters only: {k!r}"
        lines.append(f"\\providecommand{{\\{k}}}{{}}\\renewcommand{{\\{k}}}{{{_fmt(v)}}}")
    path.write_text("\n".join(lines) + "\n")
    print(f"[numbers] {path.relative_to(NOTES)}")
    return path


def rng_for(chapter: str, name: str, stream: int = 0) -> np.random.Generator:
    """Reproducible PCG64 generator seeded from the script identity (+ optional stream
    index)."""
    key = f"{chapter}/{name}" + (f"/salt{SEED_SALT}" if SEED_SALT else "")
    seed = int.from_bytes(hashlib.sha256(key.encode()).digest()[:8], "little")
    return np.random.Generator(np.random.PCG64(np.random.SeedSequence([seed, stream])))


_last_progress = [0.0]
```

There are four functions. `savefig(fig, "ch00", "name")` writes `figures/ch00/name.pdf`, which the text includes with `\includegraphics`. `save_numbers` writes a file of LaTeX macros, one per number, formatted by `_fmt` (four significant figures, and $m \times 10^e$ for very large or small values). The text loads that file with `\input` and writes `$\PiHat$` where the number belongs. `rng_for` builds a random generator whose seed is a hash of the script's name. `theory_line` draws every analytic curve the same way, black and dashed, on top of the simulation it is compared with.

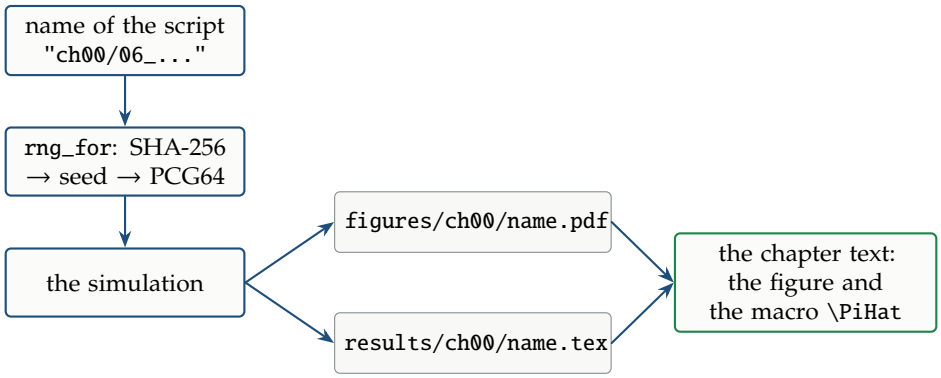


Figure 0.13: How a random number becomes a number in the text. The script's own name fixes the seed, the seed fixes every random draw, and the outputs are written to files that the chapter reads. Rerunning the script regenerates the figure and the quoted numbers together, so they cannot drift apart.

Working definition 0.6 (pseudo-random stream, seed, results macro)

A **pseudo-random generator** is a deterministic algorithm whose output passes statistical tests for independent uniform draws [PCG64 here, built in Chapter 6]. Its whole output is fixed by an integer, the **seed**. `rng_for(chapter, name, stream)` takes the first 8 bytes of the SHA-256 hash of `"chapter/name"` as a 64-bit seed and feeds it, together with the `stream` index, to numpy's `SeedSequence`, which scrambles the pair into a full generator state. Same name and stream [same seed] give the identical stream; different names give streams with no detectable correlation. A **results macro** is a LaTeX command such as `\PiHat` defined in `results/chNN/name.tex` by the script, so that the text quotes the number the code actually produced.

Example 0.15 (can two scripts share a seed by accident?). If two scripts got the same seed, their "independent" random numbers would be identical. How likely is that? The hash scatters names over 2^{64} possible seeds, effectively at random. With M scripts, the probability that some pair collides is bounded as in the birthday problem:

$$\mathrm{P}(\text{some collision}) \overset{(1)}{\le} \sum_{\text{pairs}} \mathrm{P}(\text{this pair collides}) \overset{(2)}{=} \binom{M}{2} 2^{-64}$$
$$\overset{(3)}{\le} \frac{M^2}{2}\, 2^{-64} \overset{(4)}{=} \frac{10^6}{2} \times 5.4 \times 10^{-20} = 2.7 \times 10^{-14}$$

for $M = 1000$.

(1) The probability of a union is at most the sum of the probabilities (the union bound).
(2) Two independent uniform seeds agree with probability 2^{-64}; there are $\binom{M}{2}$ pairs.
(3) $\binom{M}{2} = M(M-1)/2 \le M^2/2$.
(4) $M = 1000$ and $2^{-64} = 5.4 \times 10^{-20}$.

So with a thousand scripts a clash is, for practical purposes, impossible. And because `SeedSequence` mixes in the `stream` index as well, one script can take several independent streams; the three detectors of Section 0.5 use streams $0, 1, 2$.

Example 0.16 (a Monte Carlo number with its error bar: π from darts). Every Monte Carlo number in the book comes with an error bar; here is the simplest case. Throw N points uniformly in the unit square. The fraction $\hat{q}$ that lands inside the quarter disc $x^2 + y^2 < 1$ estimates the disc's area, $q = \pi/4$, so $\hat{\pi} = 4\hat{q}$ estimates π. How far off is it likely to be? A fraction of N independent trials scatters by $\sqrt{q(1-q)/N}$ (Example 0.4), so

$$\sigma_{\hat{\pi}} \overset{(1)}{=} 4\,\sigma_{\hat{q}} \overset{(2)}{=} 4\sqrt{\frac{q(1-q)}{N}} \overset{(3)}{=} 4\sqrt{\frac{0.7854 \times 0.2146}{10^6}} = 4 \times 4.106 \times 10^{-4} = 1.642 \times 10^{-3}.$$

(1) Multiplying a random variable by 4 multiplies its standard deviation by 4.
(2) The scatter of a fraction of N independent trials, Example 0.4.
(3) $q = \pi/4 = 0.7854$ and $N = 10^6$.

The script `06_reproducible_seeds.py` does exactly this with $N = 1000000$. It finds $\hat{\pi} = 3.143$ with estimated error 0.001641, within about one standard error of $\pi = 3.14159$. One more decimal digit means an error ten times smaller, and since the error falls like $1/\sqrt{N}$ that costs 100 times more darts. This $1/\sqrt{N}$ law dominates every Monte Carlo budget in the book.

Example 0.17 (the results file). The same script also checks the seeding. It builds two generators from the same name and one from a different name. The first draw is 0.217 for both generators with the same name (the streams are identical: yes), and 0.4034 for the other name. The file the script wrote, which the text reads, is:

Listing 0.10: The generated macro file (`results/ch00/06_reproducible_seeds.tex`)

```
% generated by code/ch00/06_reproducible_seeds.py -- do not edit
\providecommand{\SeedIdentical}{}\renewcommand{\SeedIdentical}{yes}
\providecommand{\SeedFirstDraw}{}\renewcommand{\SeedFirstDraw}{0.217}
\providecommand{\SeedOtherFirstDraw}{}\renewcommand{\SeedOtherFirstDraw}{0.4034}
\providecommand{\PiHat}{}\renewcommand{\PiHat}{3.143}
\providecommand{\PiErr}{}\renewcommand{\PiErr}{0.001641}
\providecommand{\PiN}{}\renewcommand{\PiN}{1000000}
```

Each line is `\providecommand` followed by `\renewcommand`. The first defines the macro if it does not exist yet; the second sets its value. So the file can be loaded twice without error, and the latest values always win. Macro names contain letters only, because LaTeX requires it, and `save_numbers` enforces this with an assertion.

What this bought us. Rerunning a script regenerates its figure and its numbers together, with the same random draws. The text never contains a simulated number typed by hand.

Watch out: reproducible is not the same as correct

A fixed seed makes a wrong answer reproducible too. Reproducibility lets us see *whether* a change in the code changed a result. It does not replace two other checks. One is the test of Chapter 6: does the generator pass statistical tests? The other is the comparison of every simulation with a known analytic limit. That is why every script in the book plots a black dashed theory curve wherever one exists.

Running everything. To regenerate every figure and number, run each script from the Notes directory, e.g. `for f in code/ch*/[0-9]*.py; do python3 "$f"; done`. Then build the book with `latexmk -pdf main.tex`. Expensive products (the fiducial CAMB spectrum, large simulated map sets, long chains) are cached in `data/` and reused. To force a recomputation, delete the cache file. One part can be built on its own with `./tools/testpart.sh`.

Chapter summary

concept	operational meaning	used later in
prediction, $\mathrm{P}(E \mid H)$	forward: fraction of the H-world in which E happens; compute along a tree path	Chapters 1 and 2
inference, $\mathrm{P}(H \mid E)$	backward: keep the paths that end at E, renormalise	Chapter 5
Bayes' theorem	posterior $\propto$ likelihood $\times$ prior; the evidence E shrinks the sample space	Chapters 5 and 9
odds form	posterior odds = likelihood ratio $\times$ prior odds; independent data multiply	Chapters 4 and 5
estimator, sampling distribution	a function of data; its histogram over repeated experiments, with bias and variance	Chapter 3
p-value	tail probability under H_0 of "this or more extreme in this direction"	Chapter 4
Monte Carlo error	a simulated fraction scatters by $\sqrt{q(1-q)/N}$	Chapters 6 and 8
Fisher information	minus the expected curvature of $\ln \mathcal{L}$; best possible variance $1/I$	Chapter 10
pipeline	model $\to$ data $\to$ what matters $\to S \to P(S \mid \boldsymbol{\theta}) \to P(\boldsymbol{\theta} \mid S)$	every chapter
sufficient summary	likelihood depends on the data only through S; compression is free	Chapters 3 and 7
summary ladder	amplitude, correlation, higher order, morphology, topology	Chapters 7 and 12
excursion set, β_0, β_1, χ	region above ν; pieces, holes, $\chi = \beta_0 - \beta_1 = V - E + F$	Chapter 12
robust observable	$T[\mathcal{N}[\phi]] = T[\phi]$ for a stated class of nuisances	Chapter 12
likelihood from noise	density of the required noise $\boldsymbol{d} - m(\boldsymbol{\theta})$, read as a function of $\boldsymbol{\theta}$	Chapters 3 and 5
Poisson, C_ℓ, Whittle likelihoods	$\sum[n \ln \lambda - \lambda]$; $-\frac{1}{2}\sum(2\ell+1)[\widehat{C}_\ell / C_\ell + \ln C_\ell]$; $-\frac{1}{2}(d-h \mid d-h)$	Chapters 4, 7 and 10
reproducible streams	seed from the script name; numbers reach the text as macros	every chapter

Reading the source (Reading the literature)

- L. Wasserman, *All of Statistics*, Preface (the probability / inference figure and the statistics–data-mining dictionary), Ch. 1, and the discussion of the null hypothesis (p. 149) and maximum likelihood (p. 122).
- J. K. Kruschke, *Doing Bayesian Data Analysis*, Chs. 1–2: reallocation of credibility, Holmes and the bouncy balls, the five steps. The clearest first read.
- *Seeing Theory*, "Basic probability": theoretical versus empirical frequency, interactively.
- On topology of CMB polarisation and the cosmic web: [1–5].

Chapter T0

Python for This Book

Where this chapter is going

Short Python scripts produce the numerical results and figures throughout these notes. A first encounter with names such as `np.linspace`, `ax.loglog`, `rng.standard_normal` and `optimize.brentq` can make even a familiar calculation hard to follow. Fortunately, the scripts reuse a small vocabulary. Two dozen names account for about half the work, and knowing roughly a hundred is enough to follow almost every line.

We can find that vocabulary by reading the scripts as text, recording which language features and function calls appear and counting how often each one occurs. Those counts help us decide what to learn first and which specialised calls to look up when we need them.

A script is read from the top down (Figure T0.1). It opens with a docstring that states its question, then a fixed block of imports, then the computation, then a figure and a file of numbers. The computation is written in plain Python (numbers, text, lists and dictionaries, loops and conditions, functions) and almost all of its arithmetic is done by NumPy, the array library: arrays, indexing, masks, broadcasting, random numbers, linear algebra and Fourier transforms. On top of NumPy sit a few SciPy tools (integrals, roots, minimisers, probability distributions) and the few Matplotlib calls that draw every figure. A handful of features appear in only a few scripts, and two physics libraries, healpy and CAMB, can be read as black boxes with known inputs and outputs. And a few traps catch every newcomer once.

Knowing the syntax is not the end. A script can run without an error message and still print a number that is partly noise: rounding error, a step size chosen badly, a Monte Carlo average from too few draws. So every printed number deserves one more question: *is it real?* Telling numerical noise from signal, knowing when sixteen digits are not enough, and predicting how long a computation will take before starting it are part of the same small toolkit.

The book's pipeline (Chapter 0) runs from a physical model to a dataset, to a summary of the data, to the sampling distribution of that summary, and back to the physics. Every arrow of it is a computation, and the same small toolkit serves all of them. With it, every script in the book reads like a derivation written in a slightly different notation.

T0.1 How much Python does the book use? A census

Suppose we want to learn just enough French to read one particular novel. A sensible first step is to count the words in the novel itself. Most of the text will turn out to be made of a few hundred very common words, and the rare ones can be looked up when we meet them. We can treat the book's code the same way. There are 345 Python files under `code/`, with 45734 lines between them. The question is how much of the language they actually use.

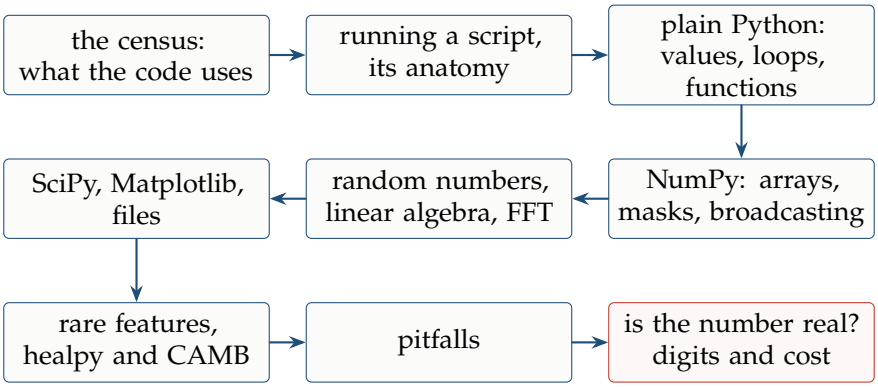


Figure T0.1: From the census to trusting a number. The census (top left) decides what is worth learning. The blue boxes are the language and its libraries, in the order a reader meets them in a script. The red box asks whether a printed number can be trusted, how many of its digits are real, and how long it took to compute.

The question, in words. Three counts answer it.

1. Which pieces of Python *syntax* (loops, conditions, function definitions, . . .) appear, and in what share of the scripts?
2. Which *libraries* do the scripts import?
3. Which *functions* do they call, how often each, and how many distinct names are needed to cover most of the calls?

Reading code without running it. A Python program is text. Before Python runs it, a part of Python called the **parser** turns the text into a tree of statements, much as a grammar teacher turns a sentence into subject, verb and object. The tree for the line `x = np.sqrt(2.0)` says: an assignment, whose target is the name `x`, and whose value is a call; the thing called is the attribute `sqrt` of the name `np`; the call has one argument, the number 2.0. Python's standard module `ast` ("abstract syntax tree") gives a program access to this tree. So a census script can read every file, build its tree, and walk through the tree counting what it finds, without executing a single line of the code it is counting. Nothing is run, so nothing can go wrong, and the census of all 345 files takes a few seconds.

Figure T0.2 shows the tree for that one line. The census script's job is to visit every node of every such tree. For a node that is a call it writes down the name of the thing being called, the way a reader would see it on the page. The function that does this is short:

Listing T0.1: Turning the called object into the name a reader sees (`code/chT0/01_census.py`)

```
36 def dotted(node):
37     """Turn the function part of a call into text: np.linalg.solve, ax.plot, len."""
38     if isinstance(node, ast.Name):
39         return node.id
40     if isinstance(node, ast.Attribute):
41         inner = dotted(node.value)
42         return f"{inner}.{node.attr}" if inner else f"obj.{node.attr}"
43     return None                                  # e.g. axes[0].plot -> an object, handled
       above
```

If the called thing is a plain name, like `len` or `print`, the name is the answer (line 38–39). If it is an attribute, like `np.sqrt`, the function calls itself on the part before the dot and glues the pieces back together with a dot (lines 40–42). So `np.linalg.solve` is rebuilt piece by piece, from `np` outwards. Calls on an object that has no simple name, such as `axes[0].plot`, are

filed under the method name, as `x.plot`. Elsewhere the script maps the local short names back to their libraries: `from scipy import stats` followed by `stats.norm.pdf` is counted as `scipy.stats.norm.pdf`. Calls of functions that the book defines itself are counted separately, because they are part of the book's own vocabulary, not of Python's.

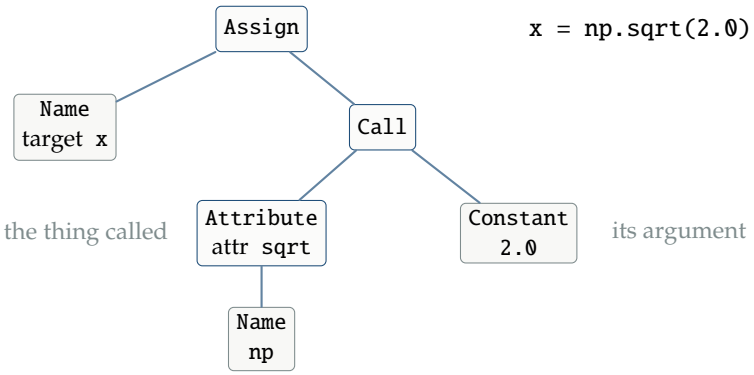


Figure T0.2: The tree that Python's parser builds from the line `x = np.sqrt(2.0)`, before anything is run. The census counts one call here, named `np.sqrt` by reading the `Attribute` node from the bottom up. Grey boxes are the leaves: names and numbers.

What the count found. The scripts make 27297 calls to functions that are not the book's own, and those calls use 759 distinct names. That sounds like a lot of names to learn. It is not, because the calls are very unevenly shared among the names. Sort the names from the most used to the least used. The first five are `np.sqrt` (1255 calls), `x.mean` (836), `x.plot` (826), `np.array` (721) and `x.set_xlabel` (707). Then ask: how many names, taken from the top of that list, are needed before they account for a given share of all the calls?

Working definition T0.1 (coverage of a vocabulary)

Rank the distinct names by the number of times they are called, $n_1 \geq n_2 \geq \cdots \geq n_D$, where D is the number of distinct names [here $D = 759$]. The **coverage** of the top K names is the share of all calls they make,

$$c(K) = \frac{\sum_{k=1}^{K} n_k}{\sum_{k=1}^{D} n_k}.$$

The number of names needed for a share q is the smallest K with $c(K) \geq q$, written K_q.

Example: the top five names. How much of the code do the five most used names cover on their own? We add their counts and divide by the total:

$$c(5) \overset{(1)}{=} \frac{n_1 + n_2 + n_3 + n_4 + n_5}{\sum_k n_k} \overset{(2)}{=} \frac{4345}{27297} = 15.9\,\%.$$

(1) The definition of $c(K)$ with $K = 5$.
(2) The five counts quoted above add up to 4345; the total number of calls is 27297.

So one call in about every six or seven in the whole book is one of five names: a square root, a mean, a plot command, the creation of an array, and the label of a horizontal axis.

Figure T0.3 is the whole answer. Half of all calls use $K_{0.5} = 25$ names. Eighty per cent use $K_{0.8} = 83$. Ninety per cent use $K_{0.9} = 159$. The curve then flattens: to reach 99% takes $K_{0.99} = 526$ names, because the last per cent is a long tail of names that each appear once or twice. There are 362 names called three times or fewer in the whole book, and together they

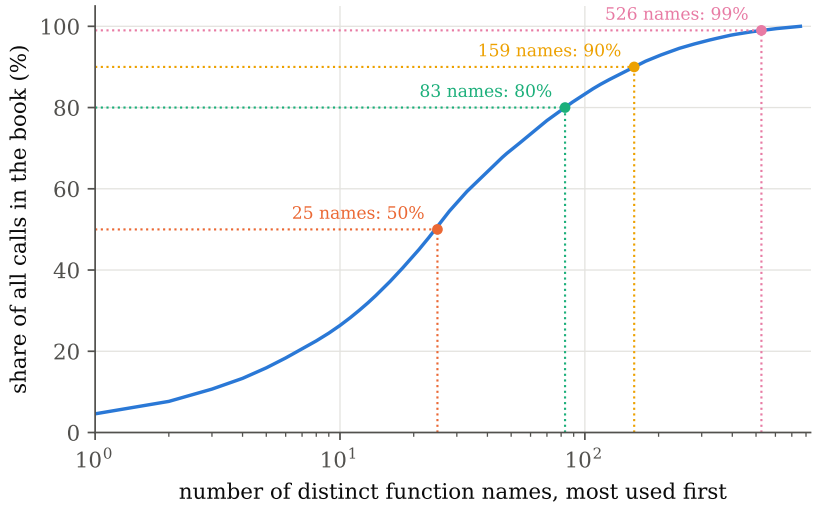


Figure T0.3: The share of all the book's library calls made by the most used names. The horizontal axis counts distinct names, most used first, on a logarithmic scale. Half of all calls use only 25 names; 83 names cover 80%, 159 cover 90%, and the last 1% of calls is spread over hundreds of names that each appear a handful of times.

make only 2.1% of the calls. Those are exactly the names it is pointless to learn in advance. Each is used once, for one special job, and the script's comment or the text next to it says what it does.

Example: what the eighty per cent means for one script. A typical script calls 36 distinct library names (the median over all files). Suppose a reader knows the 83 most used names. How many unfamiliar calls will a script of a hundred calls contain? If the script's calls are drawn like the book's as a whole, each call is one of the known names with probability $c(83) \approx 0.8$. The expected number of unknown calls is then

$$\mathbb{E}[\text{unknown calls}] \overset{(1)}{=} \sum_{i=1}^{100} \mathrm{P}(\text{call } i \text{ is unknown}) \overset{(2)}{=} 100 \times \big(1 - c(83)\big) \approx 100 \times 0.2 = 20.$$

(1) The expectation of a count is the sum of the probabilities of the events it counts (each call contributes 1 if unknown, 0 if not).
(2) Each call is unknown with probability $1 - c(K)$, the share of calls outside the known names.

Twenty calls sounds like many, but most of them repeat: a script that calls `hp.anafast` once usually calls it inside a loop or several times with different arguments. In practice the unfamiliar part of a script is one or two library functions, and the comment on that line usually says what they return.

Syntax: the same story, more strongly. The census also asks, for each piece of syntax, what share of the scripts use it at least once (Figure T0.4). Some pieces are everywhere: assignment, function calls, imports, indexing, keyword arguments and tuple unpacking appear in practically every file, `for` loops in 88%, f-strings (a way of putting numbers into text) in 84%, slices in 79%, the definition of a function in 72%. Then comes a sharp drop. The `while` loop appears in 6% of the scripts, `try/except` in 3%, classes in 2%, decorators in 2%, generators (`yield`) in none. A course on Python spends weeks on classes, inheritance, generators, exceptions and decorators. The book barely uses them.

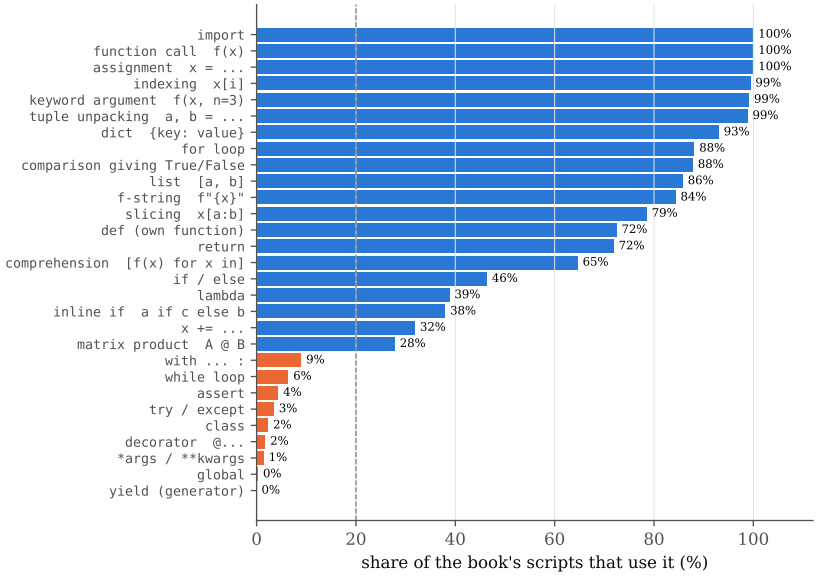


Figure T0.4: For each piece of Python syntax, the share of the book's scripts that use it at least once. Blue bars (at least 20% of the scripts) are the core every reader needs; orange bars are rare and are collected in Section T0.8. The dashed line marks 20%.

Libraries. The import statements tell the same story. NumPy is imported by 99% of the scripts and Matplotlib's plotting module by 88%. SciPy is used by 53%, almost always for one or two functions. healpy, the library for maps on the sphere, appears in 10%, and CAMB, the cosmology code, in 5%; numba and the multiprocessing module in about 1% each. By number of calls, NumPy names make 34% of all library calls, plotting calls about 22% and SciPy only 3%: SciPy functions are powerful, so one call does a lot. The rest are Python's own built-in functions, such as `len`, `range`, `print`, and methods of arrays and lists, such as `x.mean()` or `results.append(...)`.

Figure T0.5 lists the forty most used names, coloured by library. It is worth reading slowly once: together these forty names cover more than half of everything the book's code does, and each of them reappears below with an example.

Key idea: A small vocabulary does almost everything

About 25 names cover half of all the function calls in the book, about 83 cover 80%, and the syntax needed is the core of the language: assignment, calls with keyword arguments, indexing and slicing, `for` and `if`, short functions, lists, tuples and dictionaries. Everything else is rare and can be looked up when met.

That vocabulary is best learned in the order in which a reader meets it inside a script, each idea with its minimal syntax, a real excerpt from one of the book's scripts, and the one mistake most often made with it.

T0.2 Running a script and reading what it prints

Much like a written lab procedure, each script states the question, identifies the tools it needs, carries out the calculation and records the results. We can follow all four steps in `code/ch02/02_binomial.py`, which checks the binomial distribution of Section 2.2 by simulation. Each simulated experiment counts the successes in n independent trials with

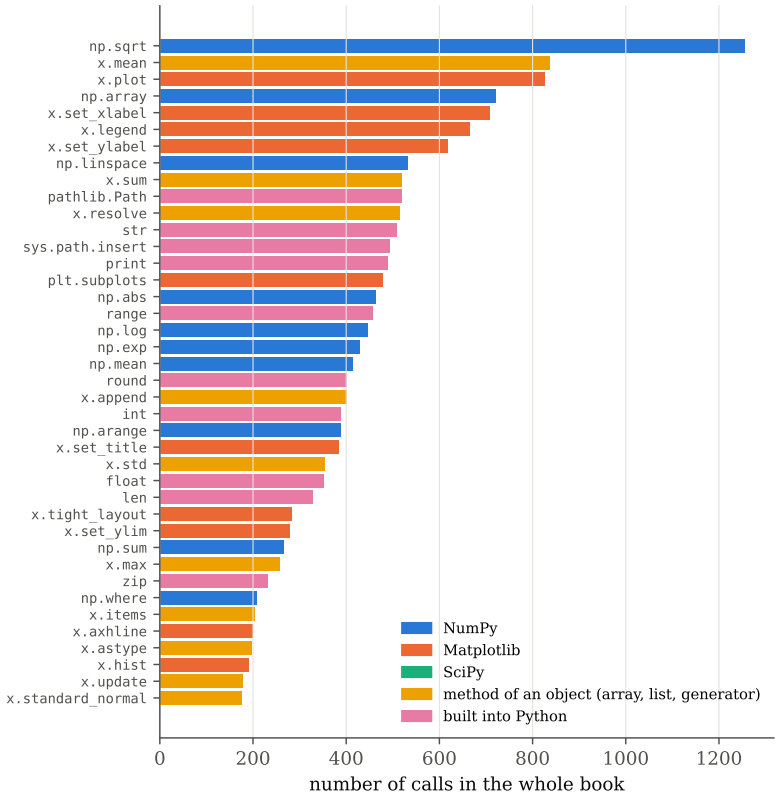


Figure T0.5: The forty most used names in the book's code, by number of calls. `x.` stands for a method called on some object: `x.mean` is the mean of an array, `x.plot` draws on a set of axes, `x.append` adds an item to a list. Blue are NumPy functions, orange Matplotlib, green SciPy, yellow methods of arrays, lists and random generators, pink Python's built-in functions.

success probability p. Across repeated experiments, does the histogram of counts follow the binomial formula $\binom{n}{k}p^k(1-p)^{n-k}$?

Running it. From the top directory of the book (the folder `Notes`), one types in a terminal

```
python3 code/ch02/02_binomial.py
```

Python reads the file from the top and carries out its lines one after another. When it finishes, two new files exist: the figure `figures/ch02/02_binomial.pdf` and the numbers file `results/ch02/02_binomial.tex`. The terminal shows what the script chose to print, here two short lines naming those files. The book's own runs were made the same way, on a computing cluster, and the two files copied back. The random numbers are seeded by the script's name (Section 0.6), so the outputs are identical wherever the script runs.

When it fails. If something goes wrong, Python stops and prints a **traceback**: a list of the lines that were being executed, innermost last, followed by one line naming the error. The last line is the one to read first. For example, adding an array of three numbers to an array of four numbers stops a script with

```
ValueError: operands could not be broadcast together with shapes (3,) (4,)
```

which says exactly what is wrong: two arrays whose shapes do not fit (Section T0.4.4). The line above it in the traceback gives the file and line number where it happened. Reading the traceback from the bottom up is the single most useful debugging habit.

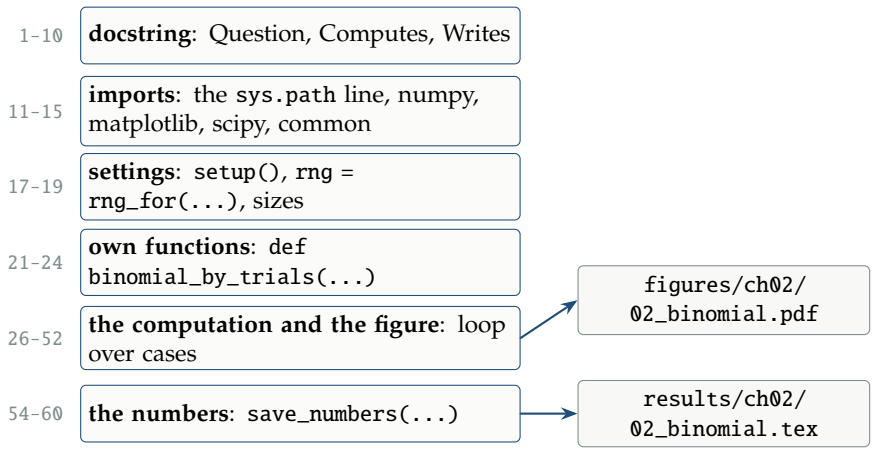


Figure T0.6: The six blocks of every script in the book, with the line numbers they occupy in `02_binomial.py`. The two grey boxes are the files the script writes: the chapter text includes the figure and reads its numbers from the second file.

The whole script. Here it is, all sixty lines.

Listing T0.2: A complete book script: the binomial distribution by simulation (`code/ch02/02_binomial.py`)

```
"""Many successive trials: the binomial distribution from counting paths.

Question:  if we repeat a Bernoulli(p) trial n times and count the successes,
           does the histogram of that count match C(n,k) p^k (1-p)^(n-k)?
Computes:  (a) 20000 repetitions of "n trials, count successes" for six (n,p)
           pairs, compared with the pmf; (b) a 3-out-of-4 coincidence trigger
           with per-layer efficiency 0.95, exact and simulated; (c) the sample
           mean and variance for n=20, p=0.2 against np and np(1-p).
Writes:    figures/ch02/02_binomial.pdf, results/ch02/02_binomial.tex
"""
import sys, pathlib; sys.path.insert(0, str(pathlib.Path(__file__).resolve().parents[1]))
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import binom
from common import setup, savefig, save_numbers, rng_for, SERIES

setup(6.4, 4.6)
rng = rng_for("ch02", "02_binomial")
reps = 20_000

def binomial_by_trials(n, p, reps):
    """Run 'reps' experiments; each is n independent Bernoulli(p) trials."""
    trials = rng.uniform(size=(reps, n)) <= p    # True = success
    return trials.sum(axis=1)                     # number of successes per experiment

# left column: p = 0.5 and growing n;  right column: n = 20 and growing p
cases = [(5, 0.5), (10, 0.5), (20, 0.5), (20, 0.1), (20, 0.2), (20, 0.6)]
fig, axes = plt.subplots(3, 2, sharex=True, sharey=True)
for ax, (n, p) in zip(axes.T.ravel(), cases):
    k = binomial_by_trials(n, p, reps)
    ks = np.arange(0, 21)
    freq = np.bincount(k, minlength=21)[:21] / reps
    ax.bar(ks, freq, width=0.8, color=SERIES[0], alpha=0.45, label="simulated")
    ax.plot(ks, binom.pmf(ks, n, p), "k_", ms=9, mew=1.6, label="pmf")
    ax.text(0.97, 0.9, f"$n={n},\\ p={p}$", transform=ax.transAxes, ha="right", va="top",
    fontsize=9)
    ax.set_ylim(0, 0.42)
for ax in axes[-1]:
```

```
38      ax.set_xlabel("number of successes $k$")
39  for ax in axes[:, 0]:
40      ax.set_ylabel("$P(k)$")
41  axes[0, 0].legend(loc="center right", fontsize=8, bbox_to_anchor=(1.0, 0.45))
42  fig.tight_layout()
43  savefig(fig, "ch02", "02_binomial")
44
45  # --- (b) coincidence trigger: at least 3 of 4 layers fire ------------------
46  eps, layers = 0.95, 4
47  exact = binom.pmf(3, layers, eps) + binom.pmf(4, layers, eps)
48  hits = binomial_by_trials(layers, eps, 1_000_000)
49  sim = np.mean(hits >= 3)
50
51  # --- (c) moments -----------------------------------------------------------
52  k = binomial_by_trials(20, 0.2, 200_000)
53
54  save_numbers("ch02", "02_binomial", {
55      "twoaTrigExact": exact,
56      "twoaTrigSim": sim,
57      "twoaTrigFour": eps**4,
58      "twoaBinMean": float(k.mean()),
59      "twoaBinVar": float(k.var(ddof=1)),
60  })
```

We walk through it block by block (Figure T0.6).

The docstring (lines 1–10). Text between triple quotes at the top of a file is the file's **docstring**. Python stores it but does not execute it. Every script in the book states three things there: the *Question* it answers, what it *Computes*, and what it *Writes*. Reading the docstring first tells us what to look for in the rest.

The imports (lines 11–15). Line 11 is the same in every script and is explained once below. Lines 12–14 load NumPy under the short name `np`, Matplotlib's plotting module as `plt`, and the binomial distribution from SciPy. Line 15 loads four helpers from the book's own module `common.py`: `setup` (the figure style), `savefig` (write a figure into the right folder), `save_numbers` (write numbers for the text) and `rng_for` (a random generator seeded by the script's name). What each helper does inside is described in Section 0.6; here it is enough to know what goes in and what comes out.

The settings (lines 17–19). `setup(6.4, 4.6)` sets the figure size in inches and the house style. `rng = rng_for("ch02", "02_binomial")` creates the random generator. Every random number in the script comes from `rng`. `reps = 20_000` fixes the number of repetitions; the underscore only makes the number easier to read.

The script's own function (lines 21–24). `binomial_by_trials(n, p, reps)` draws a table of uniform random numbers with `reps` rows and `n` columns, marks each entry `True` when it is below p (a success), and counts the successes in each row. Line 24 does the counting with `trials.sum(axis=1)`: adding `True` as 1 and `False` as 0 along each row. Every idea in these four lines (arrays, comparisons that make masks, sums along an axis) appears again in Section T0.4.

The computation and the figure (lines 26–52). Line 27 lists six pairs (n, p). Line 28 creates a figure with a 3×2 grid of panels. Line 29 walks through the panels and the pairs side by side: for each panel `ax` and each pair `(n, p)`, line 30 simulates, line 32 turns the counts into frequencies, and lines 33–34 draw the simulated bars and the exact probabilities as short black marks. Line 43 saves the figure. Lines 46–52 then compute two more checks: a trigger that fires when at least three of four detector layers fire, exactly and by simulation, and the mean and variance of the count.

The numbers (lines 54–60). `save_numbers` writes the five numbers to `results/ch02/02_binomial.tex`, each as a LaTeX command, for example `\twoaTrigExact`. The chapter

text writes that command where the number belongs, so the number in the text is always the one the script printed.

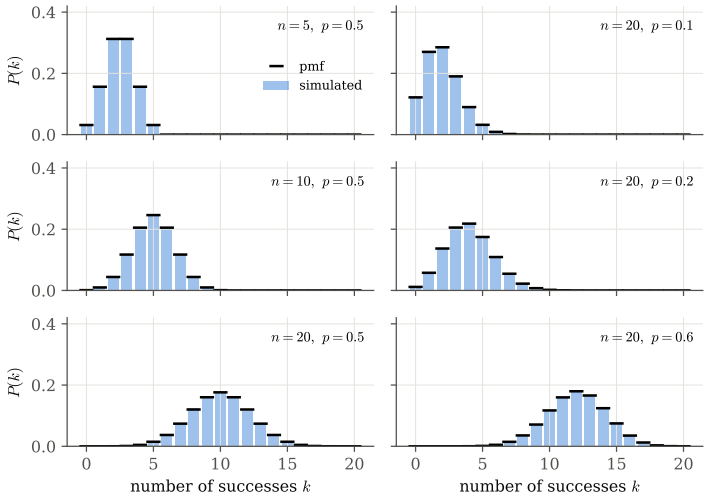


Figure T0.7: What `02_binomial.py` draws: for six choices of the number of trials n and the success probability p, the frequencies of k successes in 20 000 simulated experiments (blue bars) and the binomial probabilities (black marks). The left column grows n at $p = 0.5$; the right column grows p at $n = 20$.

The one line that is the same everywhere. Line 11 reads

```
import sys, pathlib; sys.path.insert(0,
str(pathlib.Path(__file__).resolve().parents[1]))
```

It looks cryptic, but it is a chain of small steps, each turning one thing into the next. Its job is to tell Python where `common.py` lives, so that line 15 can import it. Python looks for modules in a list of folders called `sys.path`, and the folder `code/` is not on that list by default. So the line computes the path of `code/` and puts it at the front of the list. For this script, read inside out:

$$\texttt{__file__} \overset{(1)}{=} \texttt{"code/ch02/02_binomial.py"}$$

$$\texttt{pathlib.Path(__file__).resolve()} \overset{(2)}{=} \texttt{/.../Notes/code/ch02/02_binomial.py}$$

$$\texttt{....parents[0]} \overset{(3)}{=} \texttt{/.../Notes/code/ch02}$$

$$\texttt{....parents[1]} \overset{(4)}{=} \texttt{/.../Notes/code}$$

$$\texttt{sys.path.insert(0, str(...))} \overset{(5)}{=} \text{put that folder first in the search list.}$$

(1) Python sets `__file__` to the path by which the script was started.
(2) `pathlib.Path` turns the text into a path object, and `resolve()` makes it a full path from the root of the disk, whichever folder we started from.
(3) `parents[0]` is the folder that contains the file.
(4) `parents[1]` is one folder further up: `code/`, where `common.py` is.
(5) `str` turns the path object back into text, and `insert(0, ...)` puts it at position 0, the front, so it is searched first.

The tour script `code/chT0/06_python_tour.py` prints the last names of these folders for itself: `here.parents[0].name` is `'chT0'` and `here.parents[1].name` is `'code'`. Because the path

is computed from the script's own location, the script runs correctly whether it is started from `Notes`, from `Notes/code`, or from anywhere else. Library files of one chapter used by another (such as `code/ch07/lib_fields.py`) are reached the same way, with one more `sys.path.insert` line naming their folder.

Example: reading a printed summary. Some scripts print a short report as they run. The census script of Section T0.1 prints three lines and keeps them in a text file:

```
345 files, 45734 lines, 27297 library calls, 759 distinct names
names covering 50/80/90/99 %: 25 83 159 526
names used at most 3 times: 362, carrying 2.1% of calls
```

The first line gives the size of the job (files, lines, calls, distinct names). The second gives $K_{0.5}, K_{0.8}, K_{0.9}, K_{0.99}$ of Figure T0.3, in that order. The third gives the long tail. Printing a few such lines, with words next to the numbers, is what makes a script's output readable a month later.

Watch out: Never edit a results file by hand

The files in `results/` are written by the scripts. If a number in the text looks wrong, the fix is in the script: change it, rerun it, and the text updates itself. A number changed by hand in `results/` is silently overwritten at the next run, and until then the text disagrees with the code.

T0.3 Plain Python: values, containers, loops and functions

Strip away the libraries and a script is a sequence of small instructions: give a number a name, collect several numbers together, repeat something for each item of a collection, do one thing or another depending on a condition, and package a recipe so that it can be used again. That is the whole of the plain Python the book needs. Every block below was run by `code/chT0/06_python_tour.py`, and each block shows a few lines of that script followed by exactly what they printed. The helper `say` prints a line and keeps a copy for the book; many lines use the f-string form `f"{x = }"`, which prints the expression itself followed by its value (Section T0.3.2).

T0.3.1 Numbers and names

The line `n = 20` does not state an equation. It is an instruction: compute the right-hand side and attach the name on the left to the result. The **assignment** `n = n + 1` is therefore perfectly sensible: take the current value of `n`, add one, and attach the name `n` to the new value.

Python has two kinds of number that matter here. An **integer** (type `int`) is a whole number, exact and of any size: a count of events, an index, a number of simulations. A **float** is a real number stored with about sixteen significant digits (Section T0.10 says exactly how many and why). `20` is an integer, `20.0` and `0.2` are floats, and `1.5e-3` means 1.5×10^{-3}.

Listing T0.3: The tour: numbers and names (`code/chT0/06_python_tour.py`)

```
35 n = 20                                   # an integer: a count
36 p = 0.2                                  # a float: a real number, about 16 significant
       digits
37 say(f"{n * p = }")
38 say(f"{n / 8 = }")                       # / always gives a float
39 say(f"{n // 8 = }   {n % 8 = }")         # whole-number division, and the remainder
40 say(f"{2 ** 10 = }   {1.5e-3 * 2 = }")   # powers, and scientific notation
```

```
say(f"{1_000_000 = }")                    # underscores only make long numbers readable
say(f"{type(n) = }   {type(p) = }")
```

It prints:

```
n * p = 4.0
n / 8 = 2.5
n // 8 = 2   n % 8 = 4
2 ** 10 = 1024   1.5e-3 * 2 = 0.003
1_000_000 = 1000000
type(n) = <class 'int'>   type(p) = <class 'float'>
```

The operators are `+ - * /`, `**` for powers (2^{10} is `2 ** 10`, never `2^10`, which means something else), `//` for whole-number division and `%` for the remainder. Division with `/` always gives a float, even for two integers: 20/8 = 2.5. Underscores inside a number are ignored, so `1_000_000` is a million written readably. The book uses this everywhere for sizes of simulations: `reps = 20_000` in Section T0.2, `M = 200_000` below.

Watch out: The caret is not a power

`2 ** 10` is 1024. `2 ^ 10` is 8: the caret combines the bits of two integers, an operation the book never needs. A formula copied from paper with `^` for powers runs without complaint and gives nonsense.

T0.3.2 Text and f-strings

Text is a **string**: characters between quotes, single or double. Strings are used for file names, labels on axes and lines of printed output. Two operations cover most uses. Strings can be joined with +, and single characters or pieces can be taken out by position, exactly as for lists (Section T0.3.3).

Listing T0.4: The tour: text (`code/chT0/06_python_tour.py`)

```
name = "binomial"                         # text goes between quotes, single or double
say(f"{len(name) = }   {name.upper() = }")
say(f"{'02_' + name = }")                 # + joins two strings
say(f"{name[0] = }   {name[-1] = }   {name[:3] = }")
```

It prints:

```
len(name) = 8   name.upper() = 'BINOMIAL'
'02_' + name = '02_binomial'
name[0] = 'b'   name[-1] = 'l'   name[:3] = 'bin'
```

The operation that the book uses constantly is the **f-string**: a string with an `f` before the opening quote, in which anything between curly braces is computed and written into the text. A colon inside the braces sets the format: `:.3f` means three digits after the decimal point, `:.2e` scientific notation with two digits after the point, `:,` thousands separators. Writing `=` at the end, as in `f"{n = }"`, prints the expression and its value; the tour script uses this for every line it prints.

Listing T0.5: The tour: formatting numbers in f-strings (`code/chT0/06_python_tour.py`)

```
tau_hat, sig = 1.0234567, 0.1447
say(f"tau = {tau_hat:.3f} +/- {sig:.2f}")       # :.3f -> three digits after the point
say(f"p-value = {2.87e-7:.2e}")                  # :.2e -> scientific notation
say(f"N = {1000000:,}")                          # :,   -> thousands separators
say(f"$n={n},\\ p={p}$")                         # a plot label; a backslash is written
    twice
```

It prints:

```
tau = 1.023 +/- 0.14
p-value = 2.87e-07
N = 1,000,000
$n=20,\ p=0.2$
```

Here is how a real script uses f-strings. `code/ch01/01_events_as_masks.py` (Section 1.1) sends its numbers to the text, rounding some of them to four digits on the way:

Listing T0.6: Numbers sent to the text, some formatted by f-strings (`code/ch01/01_events_as_masks.py`)

```
49 save_numbers("ch01", "1a_events_masks", {
50     "OneAMaskN": N,
51     "OneAMaskFreqA": A.mean(), "OneAMaskFreqB": B.mean(),
52     "OneAMaskFreqAB": A_and_B.mean(), "OneAMaskFreqAorB": A_or_B.mean(),
53     "OneAMaskExA": f"{pA:.4f}", "OneAMaskExB": f"{pB:.4f}",
54     "OneAMaskExAB": f"{pAB:.4f}", "OneAMaskExAorB": f"{pAorB:.4f}",
55     "OneAMaskDeMorgan": int(demorgan_violations), "OneAMaskIEGap": int(ie_gap),
56 })
```

The keys on the left (`"OneAMaskExA"` and so on) become the names of LaTeX commands. The values are either numbers, which `save_numbers` formats itself, or strings already formatted with f-strings, such as `f"{pA:.4f}"`, which writes the probability `pA` with four decimals. Plot labels are f-strings too: line 35 of the binomial script writes `f"$n={n},\\ p={p}$"`, which becomes the label $n = 20,\ p = 0.2$ in the figure.

Watch out: Backslashes in labels

A backslash inside an ordinary Python string starts a special character: `"\n"` is a new line and `"\t"` a tab. So the LaTeX label `"$\theta$"` arrives as a tab followed by `heta`. Either double every backslash (`"$\\theta$"`) or put an `r` before the quote (`r"$\theta$"`), which tells Python to leave backslashes alone. The book's scripts use both forms.

T0.3.3 Lists, tuples and dictionaries

Three containers hold several values under one name.

A **list** is an ordered sequence in square brackets, `[1, 2, 5, 30]`, which can grow. Its items are numbered from 0, not from 1, so `ns[0]` is the first item. Negative positions count from the end: `ns[-1]` is the last item. A **slice** `ns[a:b]` takes the items from position `a` up to, but not including, position `b`. `.append(x)` adds an item at the end. The method `.append` is called 400 times in the book's code, almost always in one pattern: start with an empty list, and append one result each time round a loop.

Listing T0.7: The tour: a list, and the start-empty-then-append pattern (`code/chT0/06_python_tour.py`)

```
61 ns = [1, 2, 5, 30]                          # a list: an ordered sequence that can grow
62 say(f"{ns[0] = }   {ns[-1] = }   {len(ns) = }")
63 say(f"{ns[1:3] = }")                        # a slice: positions 1 and 2, stopping before 3
64 ns.append(100)                              # add one item at the end
65 say(f"{ns = }")
66 results = []                                # the usual pattern: start empty, append in a loop
67 for k in ns:
68     results.append(k * k)
69 say(f"{results = }")
```

It prints:

```
ns[0] = 1   ns[-1] = 30   len(ns) = 4
ns[1:3] = [2, 5]
ns = [1, 2, 5, 30, 100]
results = [1, 4, 25, 900, 10000]
```

The rule "from a, stopping before b" becomes obvious once positions are pictured as the *gaps between items*, not the items themselves (Figure T0.8). Then `ns[1:3]` is everything between gap 1 and gap 3: two items. The length of a slice is `b - a`, and `ns[:k]` followed by `ns[k:]` gives back the whole list with nothing repeated and nothing lost.

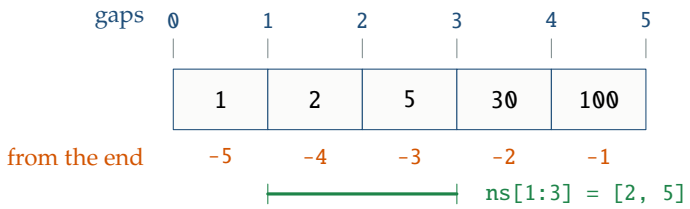


Figure T0.8: The list `ns = [1, 2, 5, 30, 100]`. Blue numbers mark the gaps between items: a slice `ns[a:b]` runs from gap `a` to gap `b`, so `ns[1:3]` holds two items. An index `ns[i]` is the item just after gap `i`; orange numbers are the negative indices, counting back from the end.

A **tuple** is a fixed group of values in round brackets, `(20, 0.2)`, which cannot be changed after it is made. The book uses tuples to keep things together that belong together, such as the pair (n, p) of a binomial distribution. Its most useful property is **unpacking**: `n, p = case` gives one name to each entry. Unpacking works in `for` loops too, and is used in practically every script.

Listing T0.8: The tour: tuples and unpacking (`code/chT0/06_python_tour.py`)

```
case = (20, 0.2)                          # a tuple: a fixed group of values
n, p = case                               # unpacking: one name for each entry
say(f"{n = }   {p = }")
cases = [(5, 0.5), (20, 0.1)]             # a list of tuples, as in the binomial script
for n, p in cases:                        # unpack each tuple as the loop meets it
    say(f"n = {n}, p = {p}, mean n p = {n * p}")
```

It prints:

```
n = 20   p = 0.2
n = 5, p = 0.5, mean n p = 2.5
n = 20, p = 0.1, mean n p = 2.0
```

A **dictionary** (`dict`) is a table that maps *keys* to *values*, written `{"H0": 67.36, "ns": 0.9649}`. A value is fetched by its key, `fid["H0"]`, not by a position. New entries are added by assignment, `fid["tau"] = 0.0544`, and `.items()` walks through the (key, value) pairs.

Listing T0.9: The tour: a dictionary (`code/chT0/06_python_tour.py`)

```
fid = {"H0": 67.36, "ombh2": 0.02237, "ns": 0.9649}    # a dictionary: names -> values
say(f"{fid['H0'] = }")
fid["tau"] = 0.0544                       # add a new entry
say(f"{len(fid) = }   {'tau' in fid = }")
for key, value in fid.items():            # walk through (name, value) pairs
    say(f"{key:>6} = {value}")
```

It prints:

```
fid['H0'] = 67.36
len(fid) = 4   'tau' in fid = True
    H0 = 67.36
```

```
ombh2 = 0.02237
   ns = 0.9649
  tau = 0.0544
```

All three containers appear together in the binomial script of Section T0.2, lines 27–30: a list of tuples `cases = [(5, 0.5), (10, 0.5), ...]`, walked through with unpacking, `for ax, (n, p) in zip(...)`. The dictionary `{...}` handed to `save_numbers` on line 54 is the commonest dictionary in the book. A slightly richer use is in `code/ch03/01_sampling_distribution.py` (Section 3.1.4), which compares four ways of estimating g from ten readings:

Listing T0.10: A dictionary of estimators, and a dictionary built from it (`code/ch03/01_sampling_distribution.py`)

```
27
28 # ---------- (A) ten Gaussian readings of g ----------
29 g_true, sigma, N = 9.81, 0.05, 10
30 x = g_true + sigma * rng.standard_normal((M, N))      # each row = one experiment
31 est = {
32     "sample mean": x.mean(axis=1),
33     "sample median": np.median(x, axis=1),
34     "mid-range": 0.5 * (x.min(axis=1) + x.max(axis=1)),
35     "first reading": x[:, 0],
36 }
37 sd = {k: v.std(ddof=1) for k, v in est.items()}
```

The dictionary `est` maps a name (a string) to an array of 200 000 estimates, one per simulated experiment. Line 37 builds a second dictionary from the first: for every key `k` and value `v`, the standard deviation of `v`. That one-line construction is a *comprehension* (Section T0.3.4).

Example: why a dictionary and not a list? The four estimators could be kept in a list, `[mean, median, midrange, first]`. Then the standard deviation of the median would be `sd[1]`, and anyone reading the code must remember that position 1 is the median. With a dictionary it is `sd["sample median"]`: the code says what it means, and adding a fifth estimator does not shift the positions of the others. The rule of thumb is: a list when the items are of the same kind and their order matters (a sequence of sizes N), a dictionary when each item has a name (estimators, parameters, numbers for the text).

Watch out: Counting from zero

The first item is `x[0]` and the last is `x[len(x) - 1]`, also written `x[-1]`. `x[len(x)]` does not exist and stops the script with `IndexError`. When a formula on paper runs over $i = 1, \ldots, N$, the code runs over `i = 0, ..., N-1`.

T0.3.4 Doing things repeatedly, and conditionally

A `for` loop repeats a block of lines once for each item of a collection. The block is marked by **indentation**: the lines that belong to the loop are shifted to the right by four spaces, and the first line that is not shifted is the first line after the loop. Python has no `end` statement; the indentation *is* the structure.

`range(n)` produces the integers $0, 1, \ldots, n-1$ (`n` numbers, stopping before `n`), and `range(a, b)` those from `a` up to `b - 1`. `enumerate(xs)` gives the position together with each item, and `zip(xs, ys)` walks two collections side by side.

Listing T0.11: The tour: three kinds of for loop (`code/chT0/06_python_tour.py`)

```
91 total = 0
```

```
92 for k in range(1, 6):                    # k = 1, 2, 3, 4, 5  (stops before 6)
93     total += k                           # the same as total = total + k
94 say(f"{total = }")
95 for i, n in enumerate([5, 10, 20]):      # enumerate: the position and the item
96     say(f"i = {i}, n = {n}")
97 for n, p in zip([5, 10], [0.5, 0.1]):    # zip: walk two lists side by side
98     say(f"n = {n}, p = {p}")
```

It prints:

```
total = 15
i = 0, n = 5
i = 1, n = 10
i = 2, n = 20
n = 5, p = 0.5
n = 10, p = 0.1
```

An `if` statement runs a block only when a condition is true. `elif` ("else if") adds further conditions, tried in order, and `else` catches everything left. Conditions are comparisons (<, <=, == for "is equal to", != for "is not equal to") joined by `and`, `or`, `not`.

Listing T0.12: The tour: if, elif, else (`code/chT0/06_python_tour.py`)

```
102 for z in [0.5, 3.4, 5.3]:
103     if z >= 5:
104         verdict = "discovery"
105     elif z >= 3:
106         verdict = "evidence"
107     else:
108         verdict = "nothing yet"
109     say(f"z = {z}: {verdict}")
110 say(f"{(2 < 3) and (3 < 1) = }   {not (3 < 1) = }")
```

It prints:

```
z = 0.5: nothing yet
z = 3.4: evidence
z = 5.3: discovery
(2 < 3) and (3 < 1) = False   not (3 < 1) = True
```

Both appear together in the dice script of Section 1.1, which colours the 36 outcomes of two dice according to the two events "first die even" and "sum at least 9":

Listing T0.13: Two nested loops and a three-way condition (`code/ch01/01_events_as_masks.py`)

```
62 for i in range(1, 7):
63     for j in range(1, 7):
64         inA, inB = (i % 2 == 0), (i + j >= 9)
65         colour = "white"
66         if inA and inB:
67             colour = SERIES[2]
68         elif inA:
69             colour = SERIES[0]
70         elif inB:
71             colour = SERIES[1]
72         ax0.add_patch(Rectangle((i - 0.45, j - 0.45), 0.9, 0.9, facecolor=colour,
73                                 edgecolor="#52514e", lw=0.6, alpha=0.75))
```

Lines 62–63 are two loops, one inside the other: for each value `i` of the first die, for each value `j` of the second, so the inner block runs $6 \times 6 = 36$ times. Line 64 computes two true/false values. Lines 66–71 choose a colour: both events (green), only A (blue), only B (orange), and otherwise the white set on line 65 stays. Line 72 draws one square.

A `while` loop repeats as long as a condition holds, for when the number of repetitions is not known in advance. It is rare in the book (6% of the scripts). A typical use is accept–reject sampling (Chapter 6): keep drawing until enough draws have been accepted. `code/ch03/10_angular_ml.py` (Section 3.9.3) does exactly that:

Listing T0.14: Draw until there are enough: a while loop (`code/ch03/10_angular_ml.py`)

```
27 def sample(size):
28     """accept-reject from the flat envelope on [xmin, xmax]"""
29     out = np.empty(0)
30     gmax = 1 + abs(a_true) + abs(b_true)
31     while out.size < size:
32         x = rng.uniform(xmin, xmax, 2 * size)
33         keep = rng.uniform(0, gmax, 2 * size) < 1 + a_true * x + b_true * x**2
34         out = np.concatenate([out, x[keep]])
35     return out[:size]
```

Line 31 says: while fewer than `size` values have been collected, draw another batch of 2×`size` candidates, keep those that pass the test (line 33), and add them to the collection (line 34). Line 35 returns exactly `size` of them.

A **comprehension** is a loop that builds a list (or a dictionary) in one line: `[k ** 2 for k in range(5)]` is the list of the squares of $0, \ldots, 4$. An `if` at the end keeps only some items. The same construction inside `sum(...)` or `max(...)` needs no brackets.

Listing T0.15: The tour: comprehensions (`code/chT0/06_python_tour.py`)

```
142 squares = [k ** 2 for k in range(5)]              # a list built by a loop in one line
143 evens = [k for k in range(10) if k % 2 == 0]      # ... keeping only some items
144 lengths = {s: len(s) for s in ["pi", "K", "p"]}   # the same for a dictionary
145 say(f"{squares = }")
146 say(f"{evens = }")
147 say(f"{lengths = }")
148 say(f"{sum(k ** 2 for k in range(5)) = }")       # inside a function call, no brackets needed
```

It prints:

```
squares = [0, 1, 4, 9, 16]
evens = [0, 2, 4, 6, 8]
lengths = {'pi': 2, 'K': 1, 'p': 1}
sum(k ** 2 for k in range(5)) = 30
```

Comprehensions are in 65% of the scripts. Line 37 of the sampling-distribution excerpt above, `{k: v.std(ddof=1) for k, v in est.items()}`, is one. A one-line `if` also exists for values: `a if condition else b`. It is how the book's code writes a default, as in `rng = np.random.default_rng() if rng is None else rng`.

Example: the same sum three ways. The sum $\sum_{k=1}^{5} k$ is computed by an explicit loop in the tour (`total = 15`). With a comprehension it is `sum(k for k in range(1, 6))`. With NumPy (Section T0.4) it is `np.arange(1, 6).sum()`. All three give 15. The loop is the clearest; the NumPy form is the fastest for long sums, by a factor of about 137 for a million terms (Section T0.4.5).

Watch out: Indentation is part of the program

Moving a line four spaces to the left takes it out of the loop: it then runs once, after the loop, instead of once per item. Python cannot warn about this, because both versions are valid programs. When a result looks like "only the last case was used", check the indentation of the line that stores the result.

T0.3.5 Functions

A **function** packages a recipe with inputs and outputs, so that it can be used many times and read once. It starts with `def`, its name and its inputs (the **parameters**) in brackets; the indented block is the recipe; `return` hands back the result. A string on the first line of the block is the function's docstring. Parameters can have **default values**, used when the caller does not give them; and a caller can give any parameter by name, as a **keyword argument**, in any order. To return several results, a function returns them separated by commas, which makes a tuple, and the caller unpacks it.

Listing T0.16: The tour: two functions and a lambda (`code/chT0/06_python_tour.py`)

```
def gaussian(x, mu=0.0, sigma=1.0):
    """Normal density at x; mu and sigma have default values."""
    return np.exp(-0.5 * ((x - mu) / sigma) ** 2) / (sigma * np.sqrt(2 * np.pi))


def mean_and_error(x):
    """Two answers at once: the sample mean and its standard error."""
    return np.mean(x), np.std(x, ddof=1) / np.sqrt(len(x))


say(f"{gaussian(0.0) = :.4f}")                    # defaults used: mu = 0, sigma = 1
say(f"{gaussian(1.0, sigma=2.0) = :.4f}")        # sigma given by name, mu left at 0
m, e = mean_and_error(np.array([9.80, 9.71, 9.86, 9.83]))
say(f"m = {m:.3f} +/- {e:.3f}")
square = lambda t: t ** 2                        # a one-line function without a name
say(f"{square(3) = }")
```

It prints:

```
gaussian(0.0) = 0.3989
gaussian(1.0, sigma=2.0) = 0.1760
m = 9.800 +/- 0.032
square(3) = 9
```

The first two lines call `gaussian(x, mu=0.0, sigma=1.0)`: once with both defaults, once giving `sigma` by name and leaving `mu` at its default. The third calls `mean_and_error`, which returns two values, unpacked into `m, e`. The last line uses a **lambda**: a one-line function without a name, `lambda t: t ** 2`, convenient when a function is needed only once, as an argument to another function.

Every one of these features is in daily use in the book. The log-likelihood of the decay-time example (Section 3.12.4) is a two-line function, and the function whose root gives the likelihood interval is a lambda:

Listing T0.17: A function, and lambdas handed to SciPy (`code/ch03/09_mle_decay.py`)

```
def lnL(tau, t):
    """log-likelihood of mean life tau for decay times t (constants dropped)."""
    return -len(t) * np.log(tau) - t.sum() / tau


# ---------- one experiment ----------
t = rng.exponential(tau_true, n)
tau_hat = t.mean()                                  # analytic MLE
tau_num = optimize.minimize_scalar(lambda s: -lnL(s, t), bounds=(0.1, 5),
    method="bounded").x
sig_curv = tau_hat / np.sqrt(n)                     # 1/sqrt(-d2 lnL/dtau2) at tau_hat
lmax = lnL(tau_hat, t)
f = lambda s: lnL(s, t) - lmax + 0.5                # zero where ln L has dropped by 1/2
lo = optimize.brentq(f, 0.3 * tau_hat, tau_hat)
```

```
40 hi = optimize.brentq(f, tau_hat, 3 * tau_hat)
```

Line 29 is the whole recipe: $\ln \mathcal{L}(\tau) = -n \ln \tau - \sum_i t_i/\tau$ for decay times t_i, with `len(t)` giving n and `t.sum()` the sum. Line 35 hands a lambda to SciPy's minimiser: "the function of `s` that returns minus the log-likelihood at `s`". Line 38 builds the function whose zeros are the points where $\ln \mathcal{L}$ has dropped by $\frac{1}{2}$ from its maximum, and lines 39–40 find those zeros (Section T0.6).

Library functions in the book use defaults and keyword arguments to stay flexible. The Gaussian-field simulator of Section 7.4.2 begins

Listing T0.18: Defaults and keyword arguments in a library function (`code/ch07/lib_fields.py`)

```
45 def gaussian_field(P, N: int, L: float, d: int = 2, rng=None, nsim: int | None = None):
46     """Gaussian random field(s) with power spectrum P(k) on a periodic box.
47
48     P    : callable P(kmag) (vectorised), in units of length^d
49     nsim : if given, return an array of nsim independent fields (axis 0)
50     The k = 0 mode is set to zero, so each box has exactly zero mean.
51     """
52     rng = np.random.default_rng() if rng is None else rng
53     D = L / N
```

The parameters `d`, `rng` and `nsim` have defaults (2 dimensions, no generator given, one field), so the simplest call is `gaussian_field(P, N, L)`, and a three-dimensional batch of fields is `gaussian_field(P, N, L, d=3, nsim=100, rng=rng)`. The words after the colons (`N: int`, `L: float`) are **type hints**: notes for the reader saying what kind of value is expected. Python ignores them.

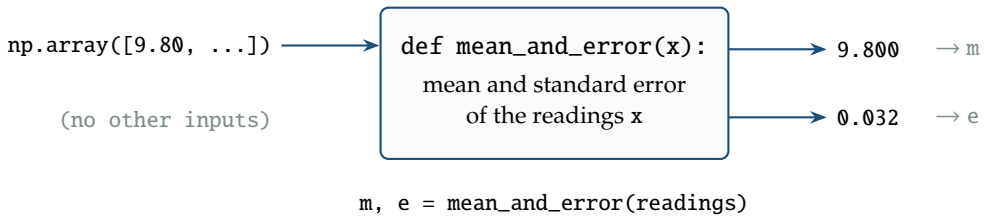


Figure T0.9: A function as a box with a name: an array of readings goes in, two numbers come out as a tuple, and the caller gives each a name by unpacking. Nothing inside the box is visible to the caller except what `return` hands back.

Example: why a function at all? The binomial script of Section T0.2 simulates the binomial count three times: for the six panels (line 30), for the trigger (line 48) and for the moments (line 52). Without the function `binomial_by_trials`, its two lines would be written three times, and a correction (say, < instead of <=) would have to be made in three places. With the function, it is made once. This is the reason almost three quarters of the scripts (72%) define at least one function of their own.

Watch out: A function that forgets to return

A function without a `return` line returns the special value `None`. The mistake looks like this: the last line computes the answer, `chi2 = np.sum(r ** 2)`, but never returns it. The caller then gets `None`, and the error appears later, somewhere else, as `TypeError: unsupported operand type(s) ... 'NoneType'`. When that word appears in an error message, look for a missing `return`.

T0.3.6 Modules and imports

A **module** is a file of Python code whose functions can be used elsewhere. `import numpy as np` loads the module `numpy` and calls it `np` inside the script, so its square root is `np.sqrt`. `from scipy import stats` loads one part of SciPy under its own name, and `from math import comb` loads a single function. The book's short names are fixed: `np` for NumPy, `plt` for Matplotlib's plotting module, `hp` for healpy.

Listing T0.19: The tour: imports (`code/chT0/06_python_tour.py`)

```
import math                                          # a whole module: use as math.something
from math import comb                                # one name from a module
say(f"{math.pi = }   {comb(10, 3) = }")
here = pathlib.Path(__file__).resolve()              # the full path of this script
say(f"{here.name = }")
say(f"{here.parents[0].name = }   {here.parents[1].name = }")
```

It prints:

```
math.pi = 3.141592653589793   comb(10, 3) = 120
here.name = '06_python_tour.py'
here.parents[0].name = 'chT0'   here.parents[1].name = 'code'
```

The book's own modules (`common.py`, `camb_fiducial.py` and the `lib_*.py` files of each chapter) are imported the same way once their folder is on `sys.path` (Section T0.2). One more construction appears at the bottom of a few of them. `code/camb_fiducial.py` ends with

Listing T0.20: Run main() only when the file is started as a script (`code/camb_fiducial.py`)

```
if __name__ == "__main__":
    main()
```

Python sets the name `__name__` to `"__main__"` in the file that was started, and to the module's own name in a file that was imported. So `python3 code/camb_fiducial.py` computes and saves the CAMB spectra, while `from camb_fiducial import load_fiducial` in another script only makes the function available, without running anything.

T0.4 NumPy: computing with whole arrays

A physicist writes $\chi^2 = \sum_i (d_i - m_i)^2/\sigma_i^2$ without thinking about the order in which the terms are added. The formula is about whole vectors: the data vector $\boldsymbol{d}$, the model vector $\boldsymbol{m}$, the vector of errors. NumPy lets code be written the same way. Its central object is the **array**: a block of numbers, all of the same kind, laid out in a grid of one or more dimensions, on which arithmetic acts element by element. `(d - m) / s` is the vector of normalised residuals, and `np.sum(((d - m) / s) ** 2)` is χ^2. NumPy names make 34% of all library calls in the book, and NumPy is imported by 99% of the scripts: this section is the heart of the toolkit. The printed lines below come from `code/chT0/07_numpy_tour.py`.

T0.4.1 Arrays, shape and type

An array is made from a list with `np.array([...])`, or directly by a function that fills it: `np.zeros`, `np.ones`, `np.full`, `np.arange`, `np.linspace`, or a random generator. Three attributes describe it. Its **shape** is a tuple giving the length along each dimension: `(4,)` for four readings, `(2, 3)` for two rows of three. `ndim` is the number of dimensions, and `dtype` the kind of number stored, almost always `float64` (a real number in double precision) or `int64`

(a whole number). Arithmetic and functions such as `np.sqrt`, `np.exp`, `np.log` act on every element.

```
g.shape = (4,)   g.dtype = dtype('float64')   g.ndim = 1
counts.dtype = dtype('int64')
X.shape = (2, 3)   X.size = 6
np.arange(5) = array([0, 1, 2, 3, 4])
np.linspace(0, 1, 5) = array([0.  , 0.25, 0.5 , 0.75, 1.  ])
np.logspace(0, 3, 4) = array([   1.,   10.,  100., 1000.])
g * 2 = array([19.6 , 19.42, 19.72, 19.66])
np.sqrt(g) = array([3.1305, 3.1161, 3.1401, 3.1353])
```

Three ways of making evenly spaced numbers need to be told apart, because they are used constantly. `np.arange(a, b, step)` starts at `a` and adds `step` until it reaches `b`, which it does *not* include; it is meant for integers. `np.linspace(a, b, n)` gives `n` points from `a` to `b` *including* both ends; it is the right choice for a grid of real numbers, such as the x values at which a density is plotted. `np.logspace(p, q, n)` gives `n` points from 10^p to 10^q, evenly spaced in the logarithm, and `np.geomspace(a, b, n)` the same between `a` and `b` themselves; they are the right choice for sample sizes, scales and frequencies that span several decades.

Example: how many points, and what spacing? A grid of θ from 0 to 1 with steps of 0.1 has eleven points, $0, 0.1, \ldots, 1$. With `np.linspace(0, 1, 11)` the spacing is

$$\Delta\theta \overset{(1)}{=} \frac{b-a}{n-1} = \frac{1-0}{11-1} = 0.1,$$

(1) n points with both ends included leave $n-1$ gaps between them.

and the last point is exactly 1. With `np.arange(0, 1, 0.1)` one gets only ten points, stopping at 0.9; and because 0.1 is not exactly representable in binary (Section T0.10), `np.arange` with a fractional step can occasionally include or exclude the end by rounding. Hence the rule: `arange` for integers, `linspace` for real grids.

T0.4.2 Indexing and slicing

Indexing an array works like indexing a list (Figure T0.8), with one addition: an array with several dimensions takes one index per dimension, separated by commas. `A[1, 2]` is the element in row 1, column 2 (both counted from 0). A colon on its own means "all of this dimension", so `A[:, 0]` is the first column and `A[1]` (or `A[1, :]`) the second row. A slice `a:b:step` can take every second or third element, and `::-1` reverses.

```
v[0] = 0   v[-1] = 90   v[2:5] = array([20, 30, 40])
v[::3] = array([ 0, 30, 60, 90])   v[::-1][:3] = array([90, 80, 70])
A =
[[ 0  1  2  3]
 [ 4  5  6  7]
 [ 8  9 10 11]]
A[1, 2] = 6   A[1] = array([4, 5, 6, 7])   A[:, 0] = array([0, 4, 8])
A[0:2, 1:3] =
[[1 2]
 [5 6]]
```

The book stores many repetitions of an experiment as the *rows* of a two-dimensional array: one row per simulated experiment, one column per reading. Then `x[:, 0]` is "the first reading of every experiment", which is how the sampling-distribution script of Section 3.1.4 builds its fourth estimator (line 34 of the excerpt in Section T0.3.3).

T0.4.3 Masks: selecting by a condition

A comparison between an array and a number, such as `x > 0`, does not give one `True` or `False`. It gives an array of them, one per element: a **boolean mask**. A mask can do three things. Used as an index, `x[m]` keeps only the elements where the mask is `True`. Summed, `m.sum()` counts the `True` entries, because `True` counts as 1 and `False` as 0. Averaged, `m.mean()` is the fraction of `True` entries, which is how a simulated probability is computed. Masks combine with `&` (and), `|` (or) and `~` (not). `np.where(m, a, b)` takes `a` where the mask is true and `b` elsewhere.

```
m = array([ True, False,  True,  True, False])
x[m] = array([0.3, 2.5, 0.8])   m.sum() = 3   m.mean() = 0.6
np.where(x > 0, x, 0.0) = array([0.3, 0. , 2.5, 0.8, 0. ])
x[(x > 0) & (x < 1)] = array([0.3, 0.8])
```

Masks are the bridge between probability and code, and the very first script of Chapter 1 is built on them (Section 1.1). An event is a set of outcomes; in a simulation, an event *is* a mask over the simulated outcomes:

Listing T0.21: Events as boolean masks over 100 000 simulated rolls of two dice (`code/ch01/01_events_as_masks.py`)

```
20 rng = rng_for("ch01", "1a_events_masks")
21 N = 100_000
22
23 # --- one row per repetition of the experiment: the outcome omega = (d1, d2)
24 d1 = rng.integers(1, 7, size=N)
25 d2 = rng.integers(1, 7, size=N)
26
27 # --- events are boolean arrays: A[i] is True when outcome i lies in A
28 A = (d1 % 2 == 0)            # first die even
29 B = (d1 + d2 >= 9)           # sum at least 9
30
31 # --- set operations are logical operations on the masks
32 A_and_B = A & B              # intersection
33 A_or_B = A | B               # union
34 not_A = ~A                   # complement
35
36 # De Morgan: (A u B)^c == A^c n B^c must hold outcome by outcome
37 demorgan_violations = np.count_nonzero(~A_or_B != (~A & ~B))
38
39 # inclusion-exclusion on counts: |A u B| = |A| + |B| - |A n B|  (exact identity)
40 ie_gap = A_or_B.sum() - (A.sum() + B.sum() - A_and_B.sum())
```

Lines 24–25 roll two dice $N = 100\,000$ times: `rng.integers(1, 7, size=N)` draws integers from 1 to 6 (the upper end 7 is excluded, as for `range`). Line 28 is the event "first die even" as a mask: `d1 % 2 == 0` is `True` where the remainder on division by 2 is zero. Lines 32–34 are intersection, union and complement, written `&`, `|`, `~`. Line 40 checks the inclusion–exclusion identity $|A \cup B| = |A| + |B| - |A \cap B|$ on the counts, each count being a `.sum()` of a mask. The frequency of A is then `A.mean()`, which estimates $\mathrm{P}(A)$.

Example: a probability as the mean of a mask. Why is `A.mean()` an estimate of $\mathrm{P}(A)$? Write $I_i = 1$ if roll i lies in A and $I_i = 0$ otherwise; the mask `A` is the array $(I_1, \ldots, I_N)$. Then

$$\texttt{A.mean()} \overset{(1)}{=} \frac{1}{N}\sum_{i=1}^{N} I_i \overset{(2)}{=} \frac{\#\{i:\ \text{roll } i \in A\}}{N}, \qquad \mathbb{E}[\texttt{A.mean()}] \overset{(3)}{=} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[I_i] \overset{(4)}{=} \mathrm{P}(A).$$

(1) `.mean()` adds the entries and divides by their number, with `True` counted as 1.
(2) The sum of the indicators counts the rolls that land in A.

(3) The expectation of a sum is the sum of the expectations.
(4) Each indicator has expectation $1 \cdot \mathrm{P}(A) + 0 \cdot (1 - \mathrm{P}(A)) = \mathrm{P}(A)$.

So the mean of a mask is the relative frequency, an unbiased estimate of the probability, with the binomial standard error $\sqrt{\mathrm{P}(A)(1-\mathrm{P}(A))/N}$ (Section T0.10).

Watch out: and is not &

For masks, write `(x > 0) & (x < 1)`, with the brackets. The word `and` works only for single `True/False` values and stops with `ValueError: The truth value of an array with more than one element is ambiguous`. Without the brackets, `x > 0 & x < 1` is read as `x > (0 & x) < 1`, because `&` binds more tightly than `>`.

T0.4.4 Broadcasting: arrays of different shapes

`g * 2` multiplies every element of `g` by 2: the single number is "stretched" to the shape of the array. NumPy generalises this to arrays of different shapes, and the generalisation, called **broadcasting**, is what lets the book write tables of pairs without loops.

Working definition T0.2 (broadcasting)

Two arrays can be combined element by element when their shapes are compatible. Compare the shapes from the right (the last dimension first). Two lengths are compatible if they are equal or if one of them is 1 [a length-1 dimension is copied as often as needed]. A missing dimension on the left counts as length 1. The result has, in each dimension, the larger of the two lengths.

The tool for making a length-1 dimension on purpose is `None` as an index. If `r` has shape `(3,)`, then `r[:, None]` has shape `(3, 1)` (a column) and `r[None, :]` has shape `(1, 3)` (a row). Their difference has shape `(3, 3)`, and its entry (i, j) is $r_i - r_j$ (Figure T0.10).

```
r[:, None].shape = (3, 1)   r[None, :].shape = (1, 3)
r[:, None] - r[None, :] =
[[ 0. -1. -3.]
 [ 1.  0. -2.]
 [ 3.  2.  0.]]
(A - A.mean(axis=0)).mean(axis=0) = array([0., 0., 0., 0.])
```

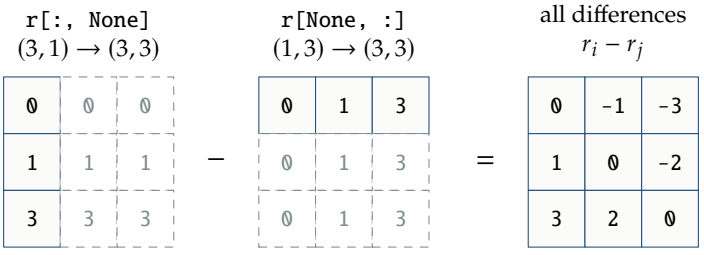


Figure T0.10: Broadcasting a column against a row. The column of shape $(3, 1)$ is copied sideways and the row of shape $(1, 3)$ is copied downwards (dashed copies are never actually stored), so the difference is the full 3×3 table of $r_i - r_j$.

The book uses exactly this to build covariance matrices that depend on the distance between two points. The correlated one-dimensional field of Section 6.9.2 has covariance $C_{ij} = e^{-|x_i - x_j|/\ell}$ between grid points x_i and x_j, and the whole 200×200 matrix is one line:

Listing T0.22: A covariance matrix from all pairwise distances, by broadcasting (`code/ch06/23_cholesky_field.py`)

```
38 # ------------------------------------------------------------------ 2. a correlated 1-D field
39 n, Lbox, ell, M = 200, 20.0, 2.0, 4000
40 xg = np.linspace(0, Lbox, n)
41 C = np.exp(-np.abs(xg[:, None] - xg[None, :]) / ell)
42 L = np.linalg.cholesky(C)
```

Line 39 sets the sizes and line 40 makes the 200 grid positions. In line 41, `xg[:, None] - xg[None, :]` is the 200×200 table of differences, `np.abs` takes their absolute values, and the exponential is applied to every entry. Line 42 factorises the matrix (Section T0.5.2).

Example: when shapes do not fit. Adding `r` of shape `(3,)` to `np.ones(4)` of shape `(4,)` compares the last lengths, 3 and 4. They differ and neither is 1, so the operation is refused with the error quoted in Section T0.2. The error is a friend: NumPy refuses to guess what was meant. The dangerous case is the opposite one, when the shapes *do* fit but not in the way intended. Subtracting a vector of 3 column means from a 3×3 matrix, `A - A.mean(axis=0)`, subtracts each column's mean from its column, as intended; `A - A.mean(axis=1)` subtracts the *row* means from the *columns*, also without complaint. For square arrays, only the numbers reveal the mistake. Writing `A.mean(axis=1, keepdims=True)`, which keeps the reduced dimension with length 1 (shape $(3, 1)$), makes the intention explicit.

T0.4.5 Vectorisation: why loops are avoided

Python can of course compute χ^2 with a loop over the data points. The question is how much slower it is, and why. `code/chT0/02_vectorise.py` computes the same χ^2 in both ways for N from 10 to 10^6 or 10^7:

Listing T0.23: The same sum, one element at a time and as one array expression (`code/chT0/02_vectorise.py`)

```
19 def chi2_loop(d, m, s):
20     """One element at a time, the way one would write it on paper."""
21     total = 0.0
22     for i in range(len(d)):
23         total += ((d[i] - m[i]) / s[i]) ** 2
24     return total
25
26
27 def chi2_vec(d, m, s):
28     """The same sum as one expression on whole arrays."""
29     return np.sum(((d - m) / s) ** 2)
```

For $N = 10^6$ the loop takes 0.3575 s and the array expression 2.614 ms, a factor of 137; the two answers agree to a relative difference of 3.731×10^{-15}, which is rounding error from adding the terms in a different order (Figure T0.11).

Where the factor comes from. Both versions do the same arithmetic, so the difference must be in the overhead around it. In the loop, Python handles each number as a separate object: it looks up `d[i]`, checks its type, creates a new object for each intermediate result, and only then does the floating-point operation, which itself takes a nanosecond or less. In the array expression, Python is consulted once per *array*; the loop over elements runs inside NumPy's compiled code, on numbers stored side by side in memory. A simple cost model makes this quantitative. Let each version cost a fixed overhead plus a cost per element, $T_{\text{loop}} = o_{\text{loop}} + c_{\text{loop}} N$

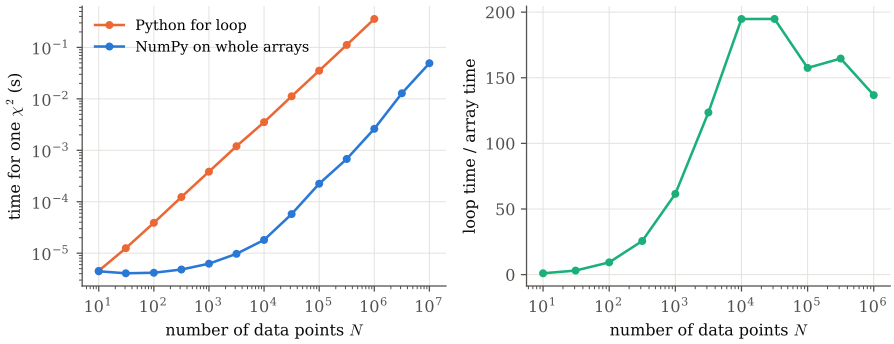


Figure T0.11: Left: the time to compute one χ^2 of N data points with a Python loop (orange) and with one NumPy expression (blue). Both grow in proportion to N, but the loop pays far more per point. Right: the ratio of the two times, close to 1 for a handful of points and about a hundred for large N.

and $T_{\text{vec}} = o_{\text{vec}} + c_{\text{vec}}N$. Then

$$\frac{T_{\text{loop}}}{T_{\text{vec}}} \overset{(1)}{=} \frac{o_{\text{loop}} + c_{\text{loop}}N}{o_{\text{vec}} + c_{\text{vec}}N} \xrightarrow{N\to\infty} \frac{c_{\text{loop}}}{c_{\text{vec}}} \overset{(2)}{=} \frac{358\text{ ns}}{2.6\text{ ns}} \approx 137.$$

(1) For large N the overheads are negligible next to the terms proportional to N.
(2) The measured costs per element at $N = 10^6$: the total times divided by 10^6.

For $N = 10$ the ratio is about 1: both versions are dominated by the fixed overhead of a function call, and vectorising buys nothing. The lesson is practical: a loop over a few dozen cases (six binomial panels, ten sample sizes) is fine and often clearest; a loop over a million data points or simulated values is written as an array expression.

T0.4.6 Reductions along an axis

A **reduction** turns many numbers into fewer: `sum`, `mean`, `std`, `var`, `min`, `max`, `median`, `percentile`. On a two-dimensional array, the keyword `axis` says which dimension is reduced, that is, which dimension *disappears*. With experiments in rows and readings in columns, `E.mean(axis=1)` averages along each row and leaves one mean per experiment; `E.mean(axis=0)` averages down each column and leaves one mean per reading slot; with no `axis`, all numbers are averaged (Figure T0.12).

```
E.mean() = 9.8067
E.mean(axis=1) = array([9.8  , 9.805, 9.815])
E.mean(axis=0) = array([9.83 , 9.7667, 9.8167, 9.8133])
E.std(axis=1, ddof=1) = array([0.0648, 0.0342, 0.0661])
E.max() = 9.9   E.argmax() = 8   np.cumsum([1, 2, 3]) = array([1, 3, 6])
```

This is how the book studies sampling distributions: simulate M experiments of N readings each as one $M \times N$ array, reduce along `axis=1` to get M estimates, and histogram them. The $1/\sqrt{N}$ study of Section 3.3.5 does it for several parent distributions and several N:

Listing T0.24: Many experiments at once: an $M \times N$ array reduced along its rows (`code/ch03/03_one_over_sqrtN.py`)

```
for N in Ns:
    for k, (draw, _) in parents.items():
        m = draw((M, N)).mean(axis=1)
        var_of_mean[k].append(m.var(ddof=1))
        if k == "Exponential(1)":
            q = np.percentile(m, [25, 75]); iqr[k].append(0.5 * (q[1] - q[0]))
```

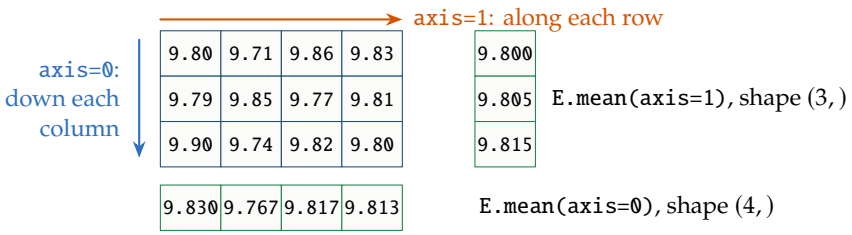


Figure T0.12: Reducing a 3×4 array of readings (three experiments in rows, four readings each). The axis named in the call is the one that disappears: `axis=1` leaves one mean per experiment, `axis=0` one mean per column.

```
40      mc = rng.standard_cauchy((M, N)).mean(axis=1)
41      q = np.percentile(mc, [25, 75]); iqr["Cauchy"].append(0.5 * (q[1] - q[0]))
```

Line 36 is the whole simulation for one parent distribution and one N: `draw((M, N))` returns $M = 20\,000$ rows of N draws, and `.mean(axis=1)` turns each row into its sample mean. Line 37 stores the variance of those M means. Line 39 computes two percentiles at once from a list, `[25, 75]`. The loop over N (line 34) and over the three parents (line 35) stays a Python loop, because it runs only 11×3 times; the $M \times N$ work inside it is vectorised.

Example: the standard error from one array. Four readings of g are in the first row of the array `E` above. Their mean and its standard error are

$$\bar{g} = \texttt{E[0].mean()}, \qquad \sigma_{\bar{g}} \overset{(1)}{=} \frac{s}{\sqrt{n}} \overset{(2)}{=} \texttt{E[0].std(ddof=1) / np.sqrt(E.shape[1])}.$$

(1) The standard error of a mean of n independent readings with sample standard deviation s (Chapter 3).

(2) `ddof=1` divides by $n - 1$ instead of n (the unbiased sample variance of Section 2.13.4); `E.shape[1]` is the number of columns, $n = 4$.

Forgetting `ddof=1` is a common slip. NumPy's default is `ddof=0`, the divide-by-n version, which is the maximum-likelihood estimate of the variance and is biased low by the factor $(n-1)/n$.

T0.4.7 Grids in two and more dimensions

A function of two variables, such as a likelihood $\mathcal{L}(a, b)$ or a Fourier-space amplitude $P(|\boldsymbol{k}|)$, is evaluated on a grid. `np.meshgrid(xs, ys, indexing="ij")` returns two arrays of the same two-dimensional shape: in the first, entry (i, j) is `xs[i]`; in the second, it is `ys[j]`. Any formula in `XX` and `YY` is then evaluated at every grid point at once. The keyword `indexing="ij"` makes the first index run over `xs`, as in matrix notation; without it, NumPy uses the plotting convention, with the axes swapped, which is a classic source of transposed images.

```
XX =
[[-1. -1. -1.]
 [ 0.  0.  0.]
 [ 1.  1.  1.]]
YY =
[[0. 1. 2.]
 [0. 1. 2.]
 [0. 1. 2.]]
np.hypot(XX, YY) =
[[1.     1.4142 2.2361]
 [0.     1.     2.    ]
 [1.     1.4142 2.2361]]
```

The Gaussian random fields of Chapter 7 need $|\boldsymbol{k}|$ on the grid of the fast Fourier transform. The library function that builds it is short:

Listing T0.25: The grid of wavenumbers for a d-dimensional box (`code/ch07/lib_fields.py`)

```
24 def kgrid(N: int, L: float, d: int, real: bool = True):
25     """|k| on the FFT grid (rfft layout if real=True), and the list of component arrays."""
26     k1 = 2 * np.pi * np.fft.fftfreq(N, d=L / N)
27     kl = 2 * np.pi * np.fft.rfftfreq(N, d=L / N) if real else k1
28     axes = [k1] * (d - 1) + [kl]
29     comps = np.meshgrid(*axes, indexing="ij")
30     kmag = np.sqrt(sum(c ** 2 for c in comps))
31     return kmag, comps
```

Line 26 gives the wavenumbers $k = 2\pi n/L$ in the order the FFT uses (Section T0.5.3). Line 28 builds a list with one copy for each dimension: `[k1] * (d - 1)` repeats a list, and `+` joins two lists. Line 29 hands that list to `np.meshgrid`; the star in `*axes` unpacks the list into separate arguments. Line 30 computes $|\boldsymbol{k}| = \sqrt{k_x^2 + k_y^2 + \dots}$ at every grid point with a generator expression inside `sum`. Line 31 returns two things.

T0.5 Random numbers, linear algebra and Fourier transforms

Three parts of NumPy carry the physics of most chapters. Random numbers make the simulated experiments of every Monte Carlo study. Linear algebra handles covariance matrices, least-squares fits and Fisher matrices. The Fourier transform turns a field or a time series into its modes, where Gaussian random fields and detector noise become simple. Each takes only a few names.

T0.5.1 Random numbers with a seed

A computer cannot toss a coin. It runs a deterministic recipe, a **pseudo-random generator**, whose output passes every statistical test for independent uniform draws (Chapter 6 builds such generators from scratch). The whole sequence is fixed by one integer, the **seed**: the same seed gives the same sequence, every time, on every machine. NumPy packages a generator as an object of type `Generator`, and every random draw is a method of that object:

call	draws
`rng.random(n)`, `rng.uniform(a, b, n)`	uniform on $[0,1)$ or $[a,b)$
`rng.standard_normal(n)`, `rng.normal(mu, s, n)`	Gaussian, $\mathcal{N}(0,1)$ or $\mathcal{N}(\mu, s^2)$
`rng.integers(a, b, n)`	integers from a to `b-1`
`rng.poisson(lam, n)`, `rng.exponential(tau, n)`	Poisson counts, exponential waiting times
`rng.choice(x, n)`, `rng.permutation(x)`	resampling, shuffling

The last argument may be a shape instead of a number: `rng.standard_normal((M, N))` is an $M \times N$ array, the "M experiments of N readings" layout of Section T0.4.6.

The book never creates a generator by hand. It calls `rng_for(chapter, name)` from `common.py`, which makes the seed from the script's own name (Section 0.6). The seeding script of Chapter 0 shows the consequence in four lines:

Listing T0.26: Same name, same random numbers; different name, unrelated ones (`code/ch00/06_reproducible_seeds.py`)

```
14 r1 = rng_for("ch00", "06_reproducible_seeds")      # seed = hash of
       "ch00/06_reproducible_seeds"
```

```
15 r2 = rng_for("ch00", "06_reproducible_seeds")     # same name -> same seed -> same stream
16 r3 = rng_for("ch00", "some_other_script")          # different name -> unrelated stream
17 a, b, c = r1.random(5), r2.random(5), r3.random(5)
```

Lines 14 and 15 build two generators from the same name, so `a` and `b` are identical arrays. Line 16 uses another name, so `c` is unrelated. An optional third argument, the **stream**, gives one script several independent generators: `rng_for("ch00", "05_three_detectors", 1)` is a different sequence from stream 0 (Section 6.5 shows why streams from one seed are independent).

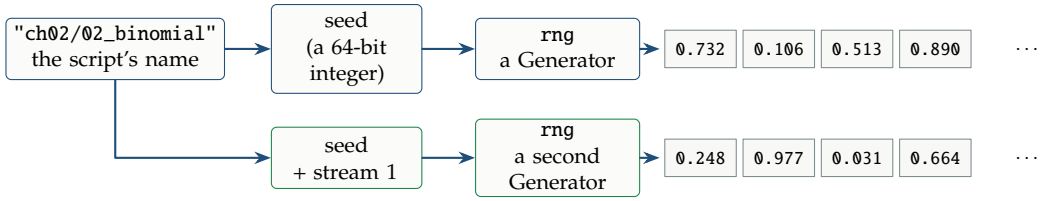


Figure T0.13: Where the random numbers of a script come from. The script's name fixes a seed, the seed fixes a generator, and the generator's sequence of draws (the numbers shown are only illustrative) is the same on every run. A stream number mixed into the seed gives a second, independent sequence.

The census checks that the book keeps this rule. Of the 216 files that draw random numbers, 197 create their generator with `rng_for`; nearly all of the rest are library files (`lib_*.py`, 15 of them) that receive a generator from the script that calls them, as `gaussian_field(..., rng=rng)` does in Section T0.3.5. A generator without a seed, `np.random.default_rng()`, appears 3 times, almost always as the fallback `if rng is None`. The old style of NumPy, a global generator set by `np.random.seed(...)` and used through `np.random.normal(...)`, appears 3 times.

Example: why a fixed seed matters for a comparison. Suppose we change the mask of a CMB simulation and the estimated variance of $\hat{C}_{100}$ moves by 2%. Is that the mask, or just different random skies? If both runs use the same seed, every sky is the same in the two runs, so the random part of the two estimates is largely shared and cancels in the difference; what remains is the mask effect plus whatever random part the two estimators do not share. If they use unrelated seeds, the 2% is the mask plus the Monte Carlo noise of two independent runs, of the same order. Fixed seeds do not make the comparison exact. They make the two runs share their noise, which shrinks the noise of the difference, and Section T0.10.5 shows the size of the gain.

Watch out: A generator made inside a loop

Writing `rng = rng_for(...)` *inside* a loop restarts the sequence at every pass, so every "independent" repetition gets the same random numbers. The generator is made once, at the top of the script, and then drawn from repeatedly. (When fresh but reproducible streams are wanted per repetition, pass the repetition number as the stream: `rng_for(chapter, name, i)`.)

T0.5.2 Linear algebra

Matrices are two-dimensional arrays, and the matrix product is the operator `@`: `A @ B`, `A @ x`, `x @ C @ x` for $x^{\mathsf{T}}Cx$ (`x.T` or `A.T` is the transpose, which does nothing to a one-dimensional array). It appears in 28% of the scripts. The star `A * B` is *not* the matrix product: it multiplies element by element. Four functions from `np.linalg` do the rest:

- `np.linalg.solve(A, b)` returns the x with $Ax = b$;
- `np.linalg.cholesky(C)` returns the lower-triangular L with $LL^\mathsf{T} = \mathbf{C}$, for a symmetric positive-definite $\mathbf{C}$ such as a covariance matrix;
- `np.linalg.eigh(C)` returns the eigenvalues (in increasing order) and the eigenvectors (as columns) of a symmetric matrix;
- `np.linalg.inv(A)` returns the inverse matrix.

```
L =
[[2.     0.    ]
 [0.9    0.4359]]
np.allclose(L @ L.T, C) = True
lam = array([0.1569, 4.8431])
np.linalg.solve(C, b) = array([-3.4211,  8.1579])
np.linalg.inv(C) @ b = array([-3.4211,  8.1579])
np.linalg.cond(C) = 30.9
Hilbert n = 10: cond = 1.6e+13, residual solve = 8.5e-17, inv = 3.6e-05, error of x =
    2.3e-05
n = 2000: solve 44 ms, inv @ b 180 ms, cholesky 57 ms
```

Cholesky: making correlated noise. The commonest use of `cholesky` in the book is to draw correlated Gaussian vectors with a given covariance $\mathbf{C}$. Draw a vector z of independent standard Gaussians and set $x = Lz$. Why does x have covariance $\mathbf{C}$? Because

$$\mathrm{Cov}(x) \overset{(1)}{=} \mathbb{E}[xx^\mathsf{T}] \overset{(2)}{=} \mathbb{E}[Lzz^\mathsf{T}L^\mathsf{T}] \overset{(3)}{=} L\,\mathbb{E}[zz^\mathsf{T}]\,L^\mathsf{T} \overset{(4)}{=} L\,\mathbb{1}\,L^\mathsf{T} \overset{(5)}{=} \mathbf{C}.$$

(1) x has zero mean, because z does, so its covariance is $\mathbb{E}[xx^\mathsf{T}]$.
(2) $x^\mathsf{T} = (Lz)^\mathsf{T} = z^\mathsf{T}L^\mathsf{T}$.
(3) L is a fixed matrix, so it comes out of the expectation.
(4) Independent standard Gaussians have unit variances and zero covariances.
(5) The defining property of the Cholesky factor.

In code, with many draws stored as the rows of an array `z`, the product Lz for every row at once is `z @ L.T`. That is exactly line 32 of the correlated-field script of Section 6.9.2:

Listing T0.27: Correlated Gaussian pairs from independent ones (`code/ch06/23_cholesky_field.py`)

```
29 C2 = np.array([[1.0, 1.6], [1.6, 4.0]])
30 L2 = np.linalg.cholesky(C2)
31 z = rng.standard_normal((50_000, 2))
32 xy = z @ L2.T                                  # each row is L z
33 S2 = np.cov(xy, rowvar=False)
```

Line 31 draws 50 000 rows of two independent Gaussians, line 32 turns each row into a correlated pair, and line 33 checks the result with `np.cov(..., rowvar=False)`, the sample covariance when the variables are the columns.

Solve, do not invert. The quantity $d^\mathsf{T}\mathbf{C}^{-1}d$ appears in every Gaussian likelihood and every Fisher matrix. It is tempting to write `d @ np.linalg.inv(C) @ d`. The better form is `d @ np.linalg.solve(C, d)`: it asks for $y = \mathbf{C}^{-1}d$ by solving $\mathbf{C}y = d$, without ever forming the inverse. The Fisher matrix of the CMB forecast in Section 10.6.1 is written this way:

Listing T0.28: A Fisher matrix summed over multipoles, with solve instead of inv (`code/ch10/lib_fisher10.py`)

```
183 def fisher_spectra(dC, C, NT, NP, fsky, lmin, lmax):
184     """The same T+E Fisher matrix written as sum_l dC^T Cov^-1 dC, with Cov the 3x3
        covariance of the
185     estimated (TT, EE, TE) spectra (Verde 2010, eq. 27 with noise added)."""
186     F = np.zeros((6, 6))
187     for l in range(lmin, lmax + 1):
188         a = C["TT"][l] + NT[l]; b = C["TE"][l]; e = C["EE"][l] + NP[l]
189         nu = (2 * l + 1) * fsky
190         cov = np.array([[2 * a * a, 2 * b * b, 2 * a * b],
191                         [2 * b * b, 2 * e * e, 2 * e * b],
192                         [2 * a * b, 2 * e * b, b * b + a * e]]) / nu
193         d = np.array([dC["TT"][:, l], dC["EE"][:, l], dC["TE"][:, l]])
194         F += d.T @ np.linalg.solve(cov, d)
195     return F
```

For each multipole ℓ (line 187), lines 190–192 build the 3×3 covariance of the estimated TT, EE and TE spectra, line 193 stacks the derivatives of the three spectra with respect to the six parameters into a 3×6 array `d`, and line 194 adds $\boldsymbol{d}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{d}$, a 6×6 matrix, to the running sum.

Why is `solve` better? There are two reasons. The first is speed: solving needs one factorisation of $\mathbf{C}$, about $\frac{2}{3}n^3$ operations, while inverting needs about $2n^3$ and then a matrix product. The tour measured, for $n = 2000$, 44 ms for `solve` against 180 ms for `inv` followed by a product. The second is accuracy, and it matters for badly conditioned matrices, where small rounding errors are amplified. The tour tried both on the Hilbert matrices $H_{ij} = 1/(i+j+1)$, which become nearly singular as their size grows (Figure T0.14, left). For $n = 10$, with condition number $\kappa \approx 1.602\times10^{13}$ (Section T0.10.4 explains this number), `solve` returns an $\boldsymbol{x}$ that satisfies the equations to 8.502×10^{-17} in relative terms, the best that double precision allows, while `inv` leaves a residual of 3.584×10^{-5}. Neither recovers $\boldsymbol{x}$ itself better than 2.272×10^{-5}, because no method can beat the conditioning of the problem.

So when is `inv` right? When the inverse matrix itself is the answer wanted. Turning a Fisher matrix $\mathbf{F}$ into the covariance of the parameters, $\mathbf{F}^{-1}$, is the commonest case, and it accounts for most of the 105 calls of `np.linalg.inv` in the book (against 29 of `solve`, 22 of `cholesky` and 13 of `eigh`). Those matrices are small (6×6 or so) and well conditioned, and there `inv` is harmless. The rule is: `inv` when the matrix A^{-1} is the result, `solve` (or `cholesky`) when A^{-1} only stands inside a formula such as $A^{-1}\boldsymbol{b}$ or $\boldsymbol{b}^{\mathsf{T}}A^{-1}\boldsymbol{b}$.

Example: eigenvectors as the axes of an ellipse. For the covariance $\mathbf{C} = \left(\begin{smallmatrix} 4 & 1.8 \\ 1.8 & 1 \end{smallmatrix}\right)$ of the tour, `eigh` returns the eigenvalues $\lambda_1 \approx 0.157$ and $\lambda_2 \approx 4.843$. Their sum is $5 = \operatorname{tr}\mathbf{C}$ and their product $0.76 = \det\mathbf{C} = 4 - 1.8^2$, as they must be. The 1σ contour of a two-dimensional Gaussian with this covariance is an ellipse whose axes point along the eigenvectors, with half-lengths $\sqrt{\lambda_1} \approx 0.40$ and $\sqrt{\lambda_2} \approx 2.20$. Section 1.13 draws it with exactly this recipe (the function `ellipse` in `code/ch01/18_covariance_transform.py`).

T0.5.3 The fast Fourier transform

The discrete Fourier transform of N numbers $x_0, \dots, x_{N-1}$ is

$$X_k = \sum_{n=0}^{N-1} x_n\, e^{-2\pi i k n/N}, \qquad k = 0, \dots, N-1,$$

the same sign convention as $\tilde{h}(f) = \int h(t)e^{-2\pi i f t}\,\mathrm{d}t$. Computed term by term it costs N^2 operations. The **fast Fourier transform** (FFT) computes the same numbers in about $N\log_2 N$ operations, by splitting the sum into its even and odd terms again and again; for a million

points that is fifty thousand times less work, which is why every power-spectrum calculation of the book is done in Fourier space. NumPy's functions live in `np.fft`:

- `np.fft.fft(x)` gives all N complex X_k; `np.fft.ifft` inverts it.
- For real data, $X_{N-k} = X_k^*$, so half of the numbers are redundant. `np.fft.rfft(x)` returns only $k = 0, \dots, N/2$, that is $N/2 + 1$ numbers, and `np.fft.irfft` inverts it. `rfftn` and `irfftn` do the same in several dimensions (the random fields of Chapter 7).
- `np.fft.rfftfreq(N, d=dt)` gives the frequency of each entry, $f_k = k/(N\,\Delta t)$, and `fftfreq` the same for the full transform, with the negative frequencies in the second half.

```
t.size = 2000   F.size = 1001   f[1] = 0.5   f[-1] = 500.0
peak at f = 50.0 Hz
Parseval: sum x^2 = 2105.229,  sum |X|^2 / N = 2105.229
```

The tour puts a weak tone, amplitude 0.5 at 50 Hz, into Gaussian noise of standard deviation 1, samples it at 1000 Hz for 2 s, and transforms (Figure T0.14, right). Two numbers follow from the sampling alone. The frequency step is $\Delta f = 1/T = 0.5$ Hz and the highest frequency is the Nyquist frequency $f_s/2 = 500$ Hz, which is why `f[1] = 0.5` and `f[-1] = 500.0`. The tone, invisible by eye in the time series, stands out at 50.0 Hz, about 180 times above the median noise power. The last printed line checks Parseval's theorem in NumPy's convention, $\sum_n x_n^2 = \frac{1}{N}\sum_k |X_k|^2$: the factor $1/N$ is where conventions differ between libraries and books, and checking Parseval is the quickest way to find out which one a function uses.

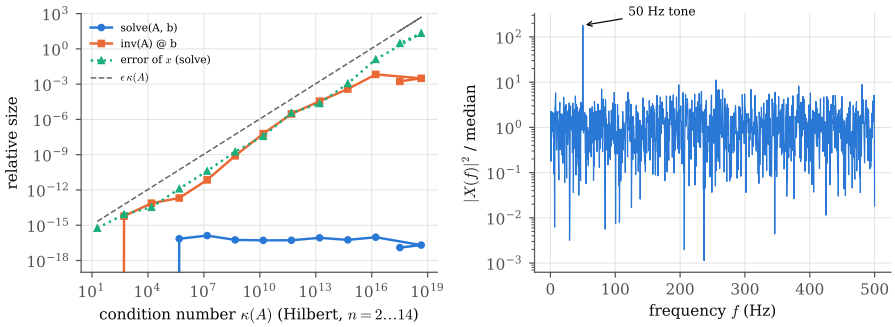


Figure T0.14: Left: solving $A\boldsymbol{x} = \boldsymbol{b}$ for Hilbert matrices of growing size. With `solve` (blue) the residual stays at the rounding level for every condition number; with `inv` followed by a product (orange) it grows with the condition number. The error of $\boldsymbol{x}$ itself (green) follows $\epsilon\kappa$ (dashed) whatever the method. Right: the power spectrum of 2 s of noise containing a weak 50 Hz tone, normalised to its median: one Fourier coefficient carries the tone.

In the book, a typical FFT is two lines inside a larger recipe. The autocorrelation of a Markov chain (Section 9.12.2) is computed with the Wiener–Khinchin theorem: the autocorrelation is the inverse transform of the power spectrum.

Listing T0.29: The autocorrelation of a chain through two FFTs (`code/ch09/lib_mcmc9c.py`)

```
74 def acf(x, maxlag=None):
75     """Normalised autocorrelation rho(k) of a 1-D series, via FFT (zero padding)."""
76     x = np.asarray(x, dtype=float) - np.mean(x)
77     n = x.size
78     f = np.fft.rfft(x, n=2 * n)
79     r = np.fft.irfft(f * np.conj(f))[:n]
80     r /= r[0]
81     return r if maxlag is None else r[: maxlag + 1]
```

Line 76 subtracts the mean. Line 78 transforms with `n=2 * n`, which pads the series with zeros to twice its length, so that the circular wrap-around of the discrete transform does not mix the end of the chain with its beginning. Line 79 multiplies the transform by its complex conjugate (the power) and transforms back; the first n entries are the autocovariance at lags $0, \ldots, n-1$. Line 80 divides by the value at lag 0, so that $\rho(0) = 1$. Done directly, this would cost n^2 operations, about 10^{10} for a chain of 10^5 steps; through the FFT it takes milliseconds.

T0.6 SciPy: distributions, integrals, roots and minima

NumPy does arithmetic on arrays. SciPy adds the classic numerical recipes that a physicist would otherwise copy from a handbook: the densities and tail probabilities of the standard distributions, numerical integration, root finding, minimisation, special functions and interpolation. It is used in 53% of the scripts but makes only 3% of the calls, because one SciPy call does a lot of work. Six tools cover almost all of it. Each is shown answering one small physics question in `code/chT0/03_scipy_tour.py`, with an answer we can check by hand, and then in a real script.

T0.6.1 Probability distributions

`scipy.stats` contains one object per distribution: `stats.norm`, `stats.poisson`, `stats.binom`, `stats.chi2`, `stats.t`, `stats.expon`, and about a hundred more. Every object answers the same five questions, which are the five things the earlier chapters ask of a distribution (Figure T0.16, left):

method	question	formula
`pdf(x)` (`pmf` if discrete)	how dense is the probability at x?	$p(x)$
`cdf(x)`	how likely is a value at most x?	$F(x) = \mathrm{P}(X \le x)$
`sf(x)`	how likely is a value above x?	$1 - F(x)$, computed directly
`ppf(q)`	below which value lies a fraction q?	$F^{-1}(q)$
`rvs(size=n)`	give me n random draws	

Shape parameters come first (`stats.poisson.pmf(k, mu)`, `stats.chi2.sf(x, df)`), and every continuous distribution also takes `loc` and `scale`, which shift and stretch it: `stats.norm.pdf(x, loc=mu, scale=sigma)` is the Gaussian of mean μ and standard deviation σ. The logarithmic versions `logpdf`, `logpmf` and `logsf` are what likelihoods use, because a product of many small densities underflows while a sum of their logarithms does not (Section T0.11).

Listing T0.30: One distribution object, several questions (`code/chT0/03_scipy_tour.py`)

```
24 # (a) one distribution object, five questions
25 z95 = stats.norm.ppf(0.975)                    # where does the cdf reach 0.975?
26 p5 = stats.norm.sf(5.0)                        # P(Z > 5) = 1 - cdf(5), computed without
       cancellation
27 x = stats.norm.rvs(loc=2.0, scale=0.5, size=10_000, random_state=rng)
28 out.update(TzZninetyfive=f"{z95:.4f}", TzPfive=p5, TzBackCdf=f"{stats.norm.cdf(z95):.4f}",
29            TzRvsMean=f"{x.mean():.4f}", TzRvsStd=f"{x.std(ddof=1):.4f}")
```

Line 25 asks where the Gaussian cdf reaches 0.975: the answer, 1.9600, is the familiar "1.96σ" of a two-sided 95% interval, and `cdf` of it gives back 0.9750. Line 26 asks for the probability of a fluctuation above 5σ: 2.867×10^{-7}, the discovery threshold of particle physics. Line 27 draws 10^4 values from $\mathcal{N}(2, 0.5^2)$ using the book's generator (`random_state=rng`); their sample mean and standard deviation are 2.0049 and 0.5039.

The same two functions do the work of the p-value-to-sigma conversion of Section 4.9.1:

Listing T0.31: Sigmas to one- and two-sided p-values (`code/ch04/09_p_to_z.py`)

```
23 Z = np.arange(1, 7)
24 p1 = stats.norm.sf(Z)
25 p2 = 2 * stats.norm.sf(Z)
```

`stats.norm.sf(Z)` acts on the whole array `Z` = 1, . . . , 6 at once, like any NumPy function. The reverse direction, from a p-value to a number of sigmas, is `stats.norm.isf(p)`, the inverse of `sf`.

Watch out: `1 - cdf` in the tail

For x far in the upper tail, $F(x)$ is a number like 0.9999999997, and $1 - F(x)$ keeps only the few digits that differ from 1. `sf(x)` computes the tail directly and keeps all sixteen. At 5σ the difference is still small; beyond about 8σ, `1 - cdf` returns exactly 0 while `sf` is still accurate (Section T0.11).

T0.6.2 Integrals

`integrate.quad(f, a, b)` integrates a function of one variable from `a` to `b` (either limit may be `np.inf`) and returns *two* numbers: the value and an estimate of its absolute error. It evaluates `f` at cleverly chosen points, compares two rules of different order on each piece of the interval, and keeps subdividing the pieces where they disagree. The tour checks it against the 5σ tail of line 26:

Listing T0.32: The 5σ tail again, by integrating the density (`code/chT0/03_scipy_tour.py`)

```
31 # (b) the same tail by integrating the density from 5 to infinity
32 tail, tail_err = integrate.quad(stats.norm.pdf, 5.0, np.inf)
33 out.update(TzQuadTail=tail, TzQuadErr=tail_err, TzQuadRel=abs(tail - p5) / p5)
```

quad returns 2.867×10^{-7} with an error estimate of 8.853×10^{-10}, and the value differs from `sf(5)` by a relative 5.328×10^{-9}. The comma on the left, `tail, tail_err = ...`, unpacks the two returned numbers. Many scripts keep only the value, writing `quad(...)[0]`, as the detector-smearing check of Section 1.10 does:

Listing T0.33: Expected counts per histogram bin from integrals of a density (`code/ch01/13_detector_smearing.py`)

```
60 frac_neg_exact = quad(p_meas, -np.inf, 0)[0]
61 counts, edges = np.histogram(E_meas, bins=140, range=(-20, 120))
62 expected = np.array([N * quad(p_meas, a, b)[0] for a, b in zip(edges[:-1], edges[1:])])
```

Line 62 is a list comprehension that integrates the density over every bin, `zip(edges[:-1], edges[1:])` pairing each left edge with the next one. Section T0.10.3 shows how the error estimate can be wrong, and what to do about it.

T0.6.3 Roots

Many statistical answers are the solution of one equation in one unknown: the parameter value where $\ln \mathcal{L}$ has dropped by $\frac{1}{2}$, the mean μ at which observing so few events would have probability 5%, the Lagrange multiplier of a maximum-entropy problem. `optimize.brentq(f, a, b)` finds a zero of `f` between `a` and `b`. It requires a **bracket**: `f(a)` and `f(b)` must have opposite signs, so that a continuous `f` must cross zero in between. It then shrinks the bracket,

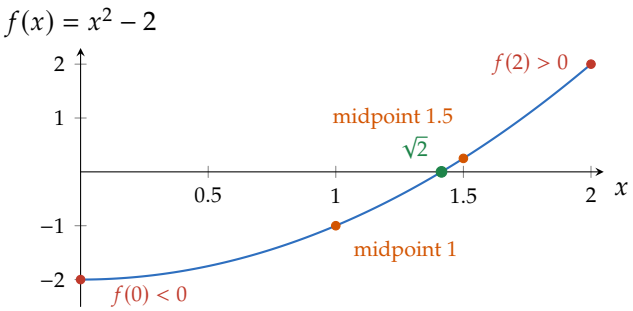


Figure T0.15: A bracket and its shrinking. $f(0) = -2 < 0$ and $f(2) = 2 > 0$ (red points), so a zero lies in $[0, 2]$. The midpoint 1 has $f < 0$, so the zero is in $[1, 2]$; the next midpoint 1.5 has $f > 0$, so it is in $[1, 1.5]$ (orange points). `brentq` mixes such halvings with faster interpolation steps and stops at $\sqrt{2}$ (green) to twelve digits. On $[0, 1]$ there is no sign change and `brentq` refuses to start.

combining halving (always safe) with interpolation steps (fast when f is smooth), until the zero is pinned to about 10^{-12} (Figure T0.15).

The tour computes the 95% upper limit on a Poisson mean after observing n events: the μ at which $\mathrm{P}(N \le n \mid \mu) = 0.05$.

Listing T0.34: An upper limit as the root of an equation (`code/chT0/03_scipy_tour.py`)

```
35 # (c) Poisson upper limit: the mean mu at which seeing n or fewer events has probability
       0.05
36 def upper_limit(n, cl=0.95):
37     return optimize.brentq(lambda mu: stats.poisson.cdf(n, mu) - (1 - cl), 1e-6, 50.0)
38 mu0, mu3 = upper_limit(0), upper_limit(3)
39 out.update(TzUpperZero=f"{mu0:.4f}", TzUpperZeroExact=f"{np.log(20):.4f}",
       TzUpperThree=f"{mu3:.3f}")
```

For $n = 0$ the equation is $e^{-\mu} = 0.05$, whose exact root is $\ln 20 = 2.9957$; `brentq` gives 2.9957. For $n = 3$ it gives 7.754 (Figure T0.16, right). The likelihood interval of the decay-time example (Section T0.3.5, lines 39–40 there) is the same pattern, with two brackets, one on each side of the maximum.

T0.6.4 Minima

Maximum likelihood means finding the parameters that make $-\ln \mathcal{L}$ smallest. `optimize.minimize(f, x0)` starts at the guess `x0` and walks downhill. It returns a result object whose most useful entries are `res.x` (the minimising parameters), `res.fun` (the value there), `res.success` (whether the method believes it converged) and `res.nit` (how many iterations it took). The `method` keyword chooses the walk: `"Nelder-Mead"` needs only function values and is robust; `"BFGS"` and `"L-BFGS-B"` use gradients (estimated numerically if not given) and are faster on smooth problems; `"L-BFGS-B"` also accepts `bounds`. Extra fixed inputs, such as the data, are handed over with `args=(...,)`.

Listing T0.35: Maximum likelihood for Gaussian data by minimisation (`code/chT0/03_scipy_tour.py`)

```
41 # (d) maximum likelihood by minimising minus the log-likelihood over (mu, ln sigma)
42 def neg_lnL(theta, data):
43     mu, ln_sig = theta
44     return -np.sum(stats.norm.logpdf(data, mu, np.exp(ln_sig)))
45 res = optimize.minimize(neg_lnL, x0=[0.0, 0.0], args=(x,), method="Nelder-Mead")
46 mu_hat, sig_hat = res.x[0], np.exp(res.x[1])
47 out.update(TzMinMu=f"{mu_hat:.4f}", TzMinSig=f"{sig_hat:.4f}",
       TzMeanCheck=f"{x.mean():.4f}",
```

```
48              TzStdCheck=f"{x.std(ddof=0):.4f}", TzMinIters=res.nit)
```

Two habits are visible here. The width is fitted as $\ln\sigma$ (line 43), so that any real number is an allowed value and the minimiser cannot wander to a negative σ. And the result is checked against an answer known in closed form: for Gaussian data the maximum-likelihood estimates are the sample mean and the divide-by-n standard deviation. The minimiser finds $\hat{\mu} = 2.0050$ and $\hat{\sigma} = 0.5039$ in 69 iterations; the closed forms give 2.0049 and 0.5039.

The Poisson-binned fit of Section 3.11.4 wraps `minimize` in a small function, so that three different goodness-of-fit functions can be minimised with the same call:

Listing T0.36: A minimiser wrapped in a function, with tightened tolerances (`code/ch03/12_poisson_bins.py`)

```
57 def fit(fun, n):
58     start = np.array([n.sum(), 1.0])
59     r = optimize.minimize(fun, start, args=(n,), method="Nelder-Mead",
60                           options=dict(xatol=1e-7, fatol=1e-9, maxiter=4000))
61     return r.x, r.fun
```

The function to minimise, `fun`, is itself an argument (functions can be passed around like numbers). `options=dict(xatol=..., fatol=..., maxiter=...)` tightens the stopping rules, because the default tolerances of Nelder–Mead are loose for a study that compares estimators to many digits. Section T0.10.3 shows what happens when a minimiser stops too early, and how `res.success` reports it.

T0.6.5 Special functions and interpolation

`scipy.special` contains the functions of mathematical physics: Γ and its logarithm `gammaln`, the beta function, binomial coefficients `comb`, error functions, Bessel functions, the Voigt profile. The one the book uses most is `gammaln`, because likelihoods of counts contain factorials, and factorials overflow: $200! \approx 10^{375}$ is far beyond the largest double, about 1.8×10^{308}. Its logarithm is a modest number:

Listing T0.37: $\ln 200!$ is fine; 200! as a float is not (`code/chT0/03_scipy_tour.py`)

```
50 # (e) logs of huge numbers: ln(200!) is fine, 200! as a float is not
51 ln_fact = special.gammaln(201)
52 try:
53     float(math.factorial(200))
54     overflow = "no"
55 except OverflowError:
56     overflow = "yes"
57 out.update(TzLnFact=f"{ln_fact:.3f}", TzLnFactCheck=f"{math.lgamma(201):.3f}",
       TzFactOverflow=overflow)
```

$\ln 200! = \ln\Gamma(201) = 863.232$, and Python's own `math.lgamma` agrees (863.232). Converting 200! itself to a float overflows (`OverflowError`: yes); the `try`/`except` block of lines 52–56 catches that error and records it instead of stopping (Section T0.8). The de Moivre–Laplace check of Section 2.9.4 uses `special.gammaln(n_st + 1)` for exactly this reason.

Interpolation estimates a function between tabulated points. `np.interp(x, xt, yt)` joins the table (x_t, y_t) by straight lines; `interpolate.CubicSpline(xt, yt)` by smooth cubic pieces whose first and second derivatives are continuous. For a coarse table of six values of $\sin x$ on $[0, \pi]$ the tour finds a largest error of 0.0489 for straight lines and 0.0027 for the spline (lines 59–66). Straight lines are the right choice when the table is fine and the curve may have kinks; a spline when the table is coarse and the curve is smooth. `np.interp` is the one the book

calls most often, typically to read a tabulated CAMB spectrum or a transfer function at new multipoles.

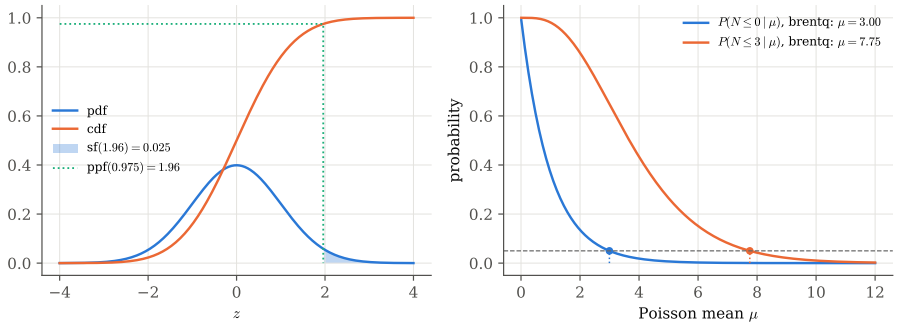


Figure T0.16: Left: the standard Gaussian's density (pdf) and cumulative distribution (cdf). The shaded tail above 1.96 is `sf(1.96)` = 0.025, and the dotted path from 0.975 on the cdf down to the axis is `ppf(0.975)` = 1.96. Right: the probability of observing at most n events as a function of the Poisson mean, for $n = 0$ and 3; `brentq` finds where each curve crosses 0.05 (dashed), the 95% upper limits.

T0.7 Matplotlib, files and folders

Every figure drawn from data in the book is made by a dozen Matplotlib calls, and every expensive result is kept on disk by two NumPy calls. Both are learned once.

T0.7.1 A figure is a canvas holding axes

Matplotlib thinks of a picture in two layers (Figure T0.17). The **figure** is the whole canvas, the thing saved to a PDF file. On it sit one or more **axes**: rectangular panels, each with its own coordinate system, ticks, labels and legend. Everything that is drawn is drawn *on an axes*. The book always creates both at once,

```
fig, ax = plt.subplots()      or      fig, axes = plt.subplots(2, 3)
```

which returns the figure and one axes, or the figure and a 2×3 array of axes, `axes[row, column]`. Everything else is a method of `ax`:

call	draws
`ax.plot(x, y)`, `ax.plot(x, y, "o")`	a line through the points, or markers only
`ax.loglog`, `ax.semilogx`, `ax.semilogy`	the same with logarithmic axes
`ax.hist(data, bins=40, density=True)`	a histogram, normalised to a density
`ax.errorbar(x, y, yerr=s, fmt="o")`	points with error bars
`ax.fill_between(x, lo, hi, alpha=0.3)`	a shaded band
`ax.imshow(image, origin="lower")`	a two-dimensional array as an image
`ax.axhline(y)`, `ax.axvline(x)`	a horizontal or vertical reference line
`ax.set_xlabel`, `set_ylabel`, `set_title`, `set_xlim`	labels and ranges
`ax.legend()`	a key built from the `label=...` of each curve

Finally `fig.tight_layout()` packs the panels so that labels do not overlap, and the book's `savefig(fig, "chNN", "name")` writes the PDF into `figures/chNN/` (Section 0.6). Colours come from the house palette set by `setup()`, and every analytic curve is drawn by `theory_line(ax, x, y)` as a black dashed line on top of the simulation it is compared with.

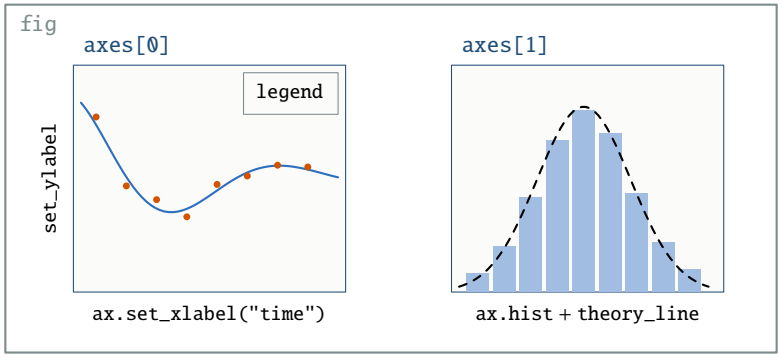


Figure T0.17: The two layers of a Matplotlib picture. The figure (`fig`) is the canvas that is saved; each axes is one panel with its own coordinates, and everything is drawn by a method of an axes. `plt.subplots(1, 2)` creates this layout and returns the figure and the array of the two axes.

The gallery script `code/chT0/04_plot_gallery.py` draws one panel of each common kind from simulated data (Figure T0.18). A real figure of the book is the same calls in a slightly longer sequence. Here is the log-likelihood plot of the decay-time example (Section 3.12.4):

Listing T0.38: A complete figure: curve, theory, reference lines, labels, legend, save (`code/ch03/09_mle_decay.py`)

```
42 fig, ax = plt.subplots(figsize=(5.2, 3.2))
43 ss = np.linspace(0.6 * tau_hat, 1.6 * tau_hat, 400)
44 ax.plot(ss, lnL(ss, t) - lmax, color=SERIES[0],
       label=r"$\ln\mathcal{L}(\tau)-\ln\mathcal{L}_{\max}$")
45 theory_line(ax, ss, -0.5 * ((ss - tau_hat) / sig_curv) ** 2, label="parabola, curvature at
       max")
46 ax.axhline(-0.5, color=SERIES[1], lw=1, ls=":")
47 for v in (lo, hi):
48     ax.axvline(v, color=SERIES[1], lw=0.8, ls=":")
49 ax.set_ylim(-3, 0.3)
50 ax.set_xlabel(r"$\tau$"); ax.set_ylabel(r"$\Delta\ln\mathcal{L}$")
51 ax.legend(fontsize=7, loc="lower center")
52 fig.tight_layout()
53 savefig(fig, "ch03", "mle_decay_lnL")
```

Line 42 makes one axes of 5.2×3.2 inches. Line 44 plots $\ln \mathcal{L}(\tau) - \ln \mathcal{L}_{\max}$ on a grid of 400 values of τ, with a LaTeX label for the legend (the `r` before the string keeps the backslashes, Section T0.3.2). Line 45 overlays the parabola that has the same curvature at the maximum. Lines 46–48 add a dotted horizontal line at $-\frac{1}{2}$ and vertical lines at the two roots found by `brentq`. Lines 49–51 set the range, the axis labels and the legend; line 53 saves.

Example: why a histogram needs `density=True`. A histogram of 5000 draws in bins of width Δx has heights n_j that add up to 5000. A density $p(x)$ integrates to 1. To compare them, the heights must be divided by the number of draws and by the bin width:

$$\hat{p}_j \overset{(1)}{=} \frac{n_j}{N\,\Delta x}, \qquad \sum_j \hat{p}_j\,\Delta x \overset{(2)}{=} \frac{1}{N}\sum_j n_j \overset{(3)}{=} 1.$$

(1) The fraction of draws in bin j, n_j/N, estimates $\int_{\text{bin}} p\,\mathrm{d}x \approx p(x_j)\,\Delta x$; divide by Δx.
(2) Multiply each height by its bin width and add.
(3) Every draw falls in some bin (if the bins cover the data), so the counts add up to N.

`ax.hist(x, bins=40, density=True)` does exactly this division, which is why the histogram and the black dashed density in the third panel of Figure T0.18 lie on top of each other. Without

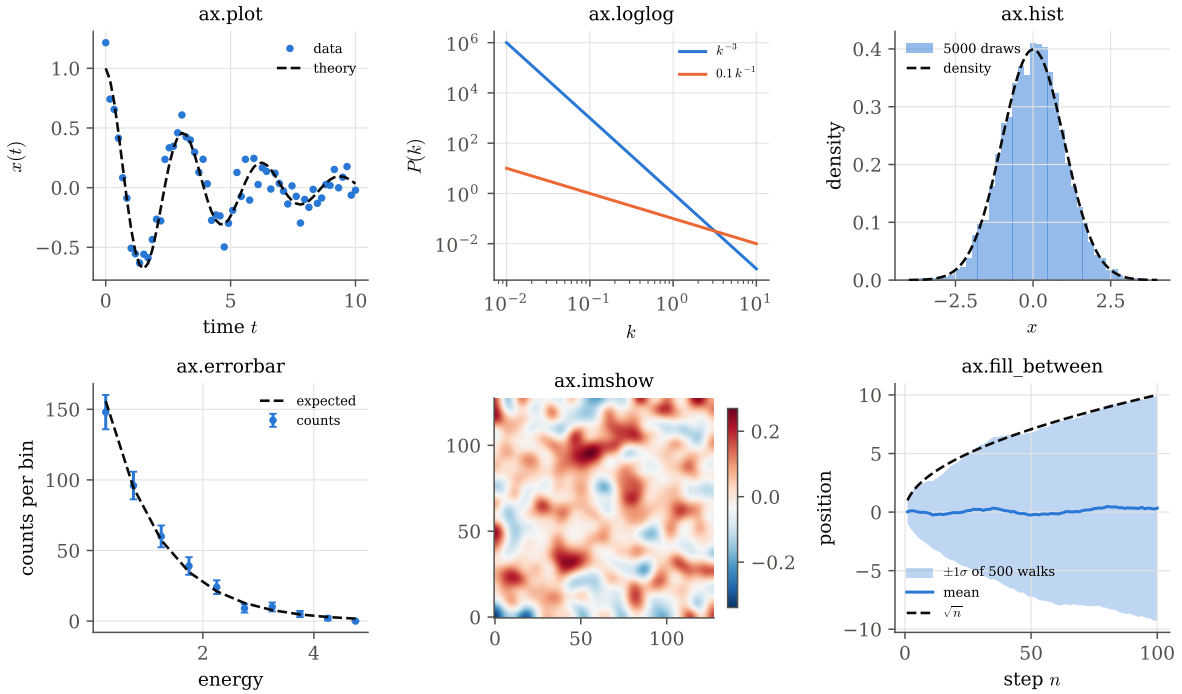


Figure T0.18: Six kinds of panel and the call that draws each, from simulated data: noisy samples of a damped oscillation with the curve they scatter around (`plot`), two power laws as straight lines on logarithmic axes (`loglog`), 5000 Gaussian draws against their density (`hist`), Poisson counts with $\sqrt{n}$ error bars (`errorbar`), a smoothed random field (`imshow`), and the $\pm1\sigma$ band of 500 random walks around their mean, against $\sqrt{n}$ (`fill_between`).

`density=True` the bars would be $N\,\Delta x \approx 5000 \times 0.2 = 1000$ times taller than the curve (for 40 bins over a range of about 8, $\Delta x \approx 0.2$).

Watch out: `plt` versus `ax`

Many examples on the internet draw with `plt.plot(...)`, which draws on "the current axes" that Matplotlib keeps in the background. With several panels, "current" is easily not the panel intended. The book always draws with `ax.plot(...)` on an axes it holds by name, and calls `plt` only to create the figure.

T0.7.2 Files, folders and caches

Some results cost minutes or hours: a thousand simulated CMB skies, a long Markov chain, a hundred calls of a Boltzmann code. They are computed once and kept in the folder `data/`, in NumPy's own file format. Two calls do it. `np.savez(path, name1=array1, name2=array2)` writes several arrays into one `.npz` file, each under a name. `z = np.load(path)` opens it, and `z["name1"]` gives back the first array, bit for bit.

Paths are handled by Python's `pathlib`. A `Path` is a file or folder name that knows its structure: `NOTES / "data" / "ch08"` joins names with the operator `/`, whatever the operating system's separator; `path.exists()` asks whether the file is there; `path.mkdir(parents=True, exist_ok=True)` creates a folder (and any missing folders above it) unless it already exists; `path.parents[1]` is two levels up (Section T0.2).

The two combine into the **cache** pattern: if the file exists, load it; otherwise compute, save, and return. The Monte Carlo library of Section 8.5 writes it once, as a function that takes the

computation itself as an argument:

Listing T0.39: Compute once, then load: a cache in eleven lines (`code/ch08/lib_cmbsim.py`)

```
252 def cached(name: str, make, force: bool = False):
253     """Load data/ch08/<name>.npz if it exists, else run make() -> dict, save it and return
        it."""
254     DATA.mkdir(parents=True, exist_ok=True)
255     path = DATA / f"{name}.npz"
256     if path.exists() and not force:
257         z = np.load(path)
258         return {k: z[k] for k in z.files}
259     d = make()
260     np.savez(path, **d)
261     print(f"[cache] wrote {path}")
262     return d
```

`make` is a function that returns a dictionary of arrays (line 252 receives it without calling it). Line 254 makes sure `data/ch08` exists. If the file is already there and recomputing was not forced (line 256), lines 257–258 load it and turn it back into a dictionary with a comprehension over `z.files`, the list of the names stored in it. Otherwise line 259 calls `make()`, line 260 saves every entry of the dictionary with `**d` (which hands each key–value pair to `np.savez` as a keyword argument), and line 262 returns the result. A thousand skies are simulated the first time; every later run of every script that needs them reads them in a fraction of a second (Figure T0.19).

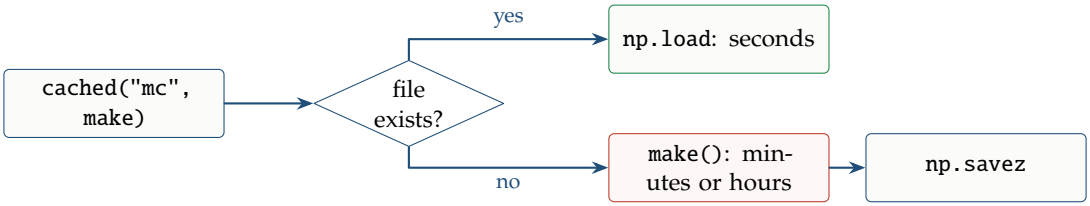


Figure T0.19: The cache pattern of `lib_cmbsim.cached`. The expensive branch (red) runs once and leaves a file behind; every later call takes the cheap branch (green). Deleting the file, or passing `force=True`, forces a fresh computation.

The fiducial CMB spectra shared by the whole book are cached the same way, in `code/camb_fiducial.py`:

Listing T0.40: Load the cached CAMB spectra, computing them on first use (`code/camb_fiducial.py`)

```
25 def load_fiducial(spec="TT"):
26     if not PATH.exists():
27         main()
28     z = np.load(PATH)
29     return z["ell"], z[spec]
30
31
32 def main():
33     ell, tot = compute()
34     PATH.parent.mkdir(exist_ok=True)
35     np.savez(PATH, ell=ell, TT=tot[:, 0], EE=tot[:, 1], BB=tot[:, 2], TE=tot[:, 3],
36              **{f"param_{k}": v for k, v in FIDUCIAL.items()})
37     print(f"wrote {PATH}")
```

`load_fiducial()` computes and saves the spectra only if `data/camb_fiducial.npz` is missing (lines 26–27), and otherwise loads them (line 28) and returns the multipoles and the requested spectrum (line 29). Line 36 shows one more use of `**`: it unpacks a dictionary built

on the spot, so that the parameters of the fiducial model are stored in the same file as `param_H0`, `param_ombh2`, and so on, and anyone opening the file can see which model produced it.

Watch out: A stale cache

A cache does not know that the code which produced it has changed. If a simulation is corrected but its cache file is left in place, every script keeps reading the old, wrong results. After changing a computation that feeds a cache, delete the file (or run once with `force=True`).

T0.8 Seen a few times: rare features and two black boxes

The census (Figure T0.4) found a sharp drop after the core of the language: a few constructions appear in fewer than one script in ten. They can be skipped on a first reading and looked up here when met. The same holds for the two physics libraries, healpy and CAMB, which the book uses as black boxes: we need to know what goes in and what comes out, not how they work inside.

T0.8.1 Classes and data classes

A **class** bundles data together with the functions that act on it. The book defines classes in only 2% of its scripts, always for the same reason: a group of settings that belong together and some quantities derived from them. The clearest example is the description of a CMB experiment in the Monte Carlo library of Section 8.2.1:

Listing T0.41: An experiment as a data class: settings, and quantities derived from them (`code/ch08/lib_cmbsim.py`)

```
68 @dataclass
69 class Experiment:
70     """A full-sky temperature experiment: resolution, depth and the analysis band limit."""
71     nside: int = 256
72     fwhm_arcmin: float = 30.0
73     depth_uK_arcmin: float = 200.0      # Delta_T: white-noise level per square arcminute
74     lmax: int = 512                     # analysis band limit, 2 nside (trusted range, ch.
       7)
75     pixwin: bool = True
76
77     @property
78     def lmax_sim(self) -> int:          # synthesis band limit, 3 nside - 1
79         return 3 * self.nside - 1
80
81     @property
82     def npix(self) -> int:
83         return hp.nside2npix(self.nside)
84
85     @property
86     def omega_pix(self) -> float:       # sr
87         return 4.0 * np.pi / self.npix
88
89     @property
90     def sigma_pix(self) -> float:       # muK per pixel: Delta_T / pixel side
91         return self.depth_uK_arcmin * ARCMIN / np.sqrt(self.omega_pix)
```

Read it as a form with blanks. Line 68, `@dataclass`, is a **decorator**: a line starting with `@` that modifies the definition below it, here by writing the routine parts of the class automatically. Lines 71–75 list the settings, each with a type hint and a default value: the HEALPix resolution,

the beam width, the noise level, the band limit, and whether to include the pixel window. Then `exp = Experiment(nside=128, fwhm_arcmin=10.0)` creates one experiment with two settings changed and the rest left at their defaults, and `exp.nside` reads a setting back. Lines 77–91 define quantities computed from the settings. The decorator `@property` makes each of them look like a setting: `exp.npix` or `exp.sigma_pix` without brackets, recomputed every time it is read, so it can never disagree with the settings. Inside the class, `self` means "this particular experiment". Lines 92 onwards are ordinary functions of the experiment, called as `exp.nl()`.

T0.8.2 Errors on purpose, and `with`

`try:` followed by `except SomeError:` runs a block and, if that particular error occurs, runs the second block instead of stopping. It appears in 3% of the scripts, typically to record that something failed. The SciPy tour uses it to show that 200! cannot be a float (Section T0.6.5), and Section T0.10.4 uses it to detect a matrix that is not positive definite. Catching errors to hide them is never done: a script that fails should fail loudly.

`with ... as name:` sets something up for the duration of an indented block and tidies it away afterwards, even if the block fails. It appears in 9% of the scripts, in three forms: opening a file, starting a pool of worker processes (below), and `with np.errstate(divide="ignore"):`, which silences NumPy's warning about a division by zero that the code is about to repair anyway (line 54 of `lib_fields.py`, where $P(k)$ is evaluated at $k = 0$ and then set to zero).

T0.8.3 Several cores at once

A Monte Carlo study runs the same function on many independent random seeds, and those runs can go on different processor cores at the same time. Python's `multiprocessing.Pool` does this with one call: `pool.map(f, jobs)` applies `f` to every item of the list `jobs`, spreads the items over the workers, and returns the results in the original order. The topology study of Section 12.23.1 is the clearest case:

Listing T0.42: Hundreds of independent simulations spread over four cores (`code/ch12/c02_mc_fnl.py`)

```
50 def one(args):
51     stream, eps_list = args
52     rng = rng_for("ch12", "c02_mc_fnl", stream)
53     g = model.gauss(rng)                                  # one Gaussian field per seed
54     res = []
55     for eps in eps_list:                                  # common random numbers across eps
56         u = model.draw(eps, g=g)
57         s = summaries(u, model)
58         s["sk"] = skew_params(u)
59         res.append(s)
60     return res
61
62
63 def run():
64     jobs = ([(i, [0.0]) for i in range(N_NULL)]
65             + [(10_000 + i, [0.0]) for i in range(N_CHECK)]
66             + [(20_000 + i, list(EPS_GRID)) for i in range(N_GRID)])
67     t0 = time.time()
68     with Pool(4) as pool:
69         out = pool.map(one, jobs, chunksize=8)
```

Each job (lines 64–66) is a tuple: a stream number and a list of non-Gaussianity amplitudes ϵ. The function `one` (lines 50–60) makes its own generator from the stream number (line 52), draws one Gaussian field, and measures it at every ϵ in the list. Line 68 starts four workers,

and line 69 runs all the jobs. Two details make this safe. Every job builds its generator from its own stream number, so the result does not depend on which worker ran which job or in what order (Section T0.5.1); a generator shared between workers would make the results depend on the scheduling. And the function given to `pool.map` is defined at the top level of the module, because each worker process imports the script to find it.

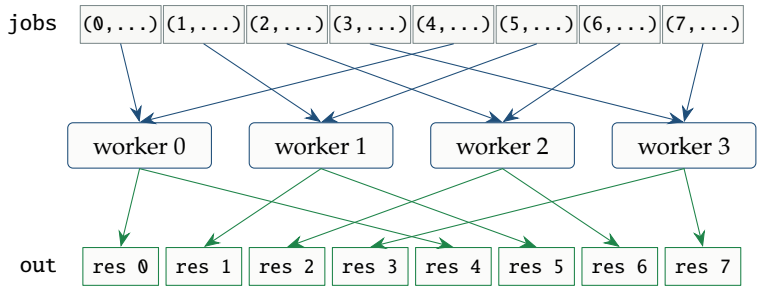


Figure T0.20: `pool.map(one, jobs)` with four workers. Each job carries its own stream number, so it produces the same result whichever worker runs it, and the list of results comes back in the order of the jobs.

T0.8.4 Compiled loops: numba

Vectorisation (Section T0.4.5) needs the elements to be independent, so that one operation can act on all of them. Some algorithms are sequential by nature: in a Metropolis sweep of the Ising model each spin's update depends on its neighbours' *current* values, which earlier steps of the same sweep may have just changed. Such a loop cannot be written as one array expression. The library numba solves this by compiling a Python function into machine code the first time it is called: the decorator `@njit` on top of the function is the only change. Five files use it (1%). The Ising sweep of Section 9.15.3 is one:

Listing T0.43: A sequential Metropolis sweep, compiled by numba (`code/ch09/20_ising.py`)

```
19 @njit(cache=True)
20 def sweep(s, beta, u):
21     """One sweep: visit every site once (in order) and propose to flip its spin.
22     dE = 2 J s_i (sum of the 4 neighbours), J = 1.  Accept with min(1, exp(-beta dE)).
23     Each single-site update satisfies detailed balance, so each preserves the Boltzmann
       law."""
24     L = s.shape[0]
25     acc = 0
26     k = 0
27     for i in range(L):
28         for j in range(L):
29             nb = s[(i + 1) % L, j] + s[(i - 1) % L, j] + s[i, (j + 1) % L] + s[i, (j - 1) %
       L]
30             dE = 2.0 * s[i, j] * nb
31             if dE <= 0.0 or u[k] < np.exp(-beta * dE):
32                 s[i, j] = -s[i, j]
33                 acc += 1
34             k += 1
35     return acc
```

Apart from line 19 this is plain Python with two nested loops: visit every site, sum its four neighbours with periodic wrap-around (the `% L` on line 28), compute the energy change, and accept the flip with the Metropolis probability. Uncompiled, the L^2 inner steps would run at Python's speed of a few hundred nanoseconds each; compiled, at a few nanoseconds. A compiled function accepts only NumPy arrays and plain numbers, which is why the random

numbers u are drawn outside, by the book's generator, and passed in.

T0.8.5 healpy and CAMB as black boxes

Two libraries do physics that the book does not reimplement. **healpy** handles maps on the sphere in the HEALPix pixelisation [7]: $12N_{\mathrm{side}}^2$ pixels of equal area, and the spherical-harmonic transforms between a map and its coefficients $a_{\ell m}$. **CAMB** [8] solves the linearised Boltzmann equations of a cosmological model and returns its predicted spectra. Neither needs to be understood inside. Each can be used as a box with known inputs and outputs (Figure T0.21), and checked from outside on a case with a known answer.

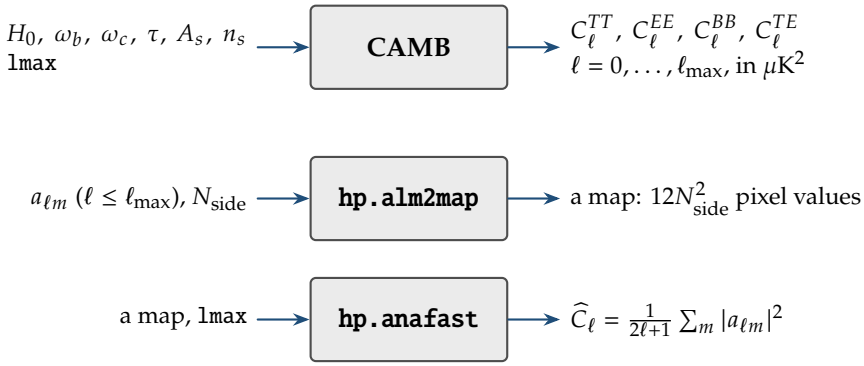


Figure T0.21: Three black boxes and what crosses their walls. CAMB turns six cosmological parameters into the predicted CMB spectra. healpy's `alm2map` turns harmonic coefficients into a pixelised map, and `anafast` turns a map back into the estimated power spectrum. A round trip through the last two must return the spectrum that went in, which is how the book checks them.

The whole CAMB interface the book uses is three calls in `code/camb_fiducial.py`:

Listing T0.44: Six parameters in, four spectra out (`code/camb_fiducial.py`)

```
16 def compute(params=FIDUCIAL, lmax=LMAX):
17     import camb
18     p = camb.set_params(**params, lmax=lmax + 200, lens_potential_accuracy=1)
19     r = camb.get_results(p)
20     cls = r.get_cmb_power_spectra(p, CMB_unit="muK", raw_cl=True)
21     tot = cls["total"][: lmax + 1]
22     return np.arange(lmax + 1), tot  # columns TT, EE, BB, TE
```

Line 18 collects the parameters into one object (the `**params` unpacks the dictionary `FIDUCIAL` of line 13 into keyword arguments such as `H0=67.36`); `lmax` is set 200 above the multipoles wanted, because the lensing calculation needs a margin. Line 19 runs the Boltzmann code, which takes several seconds. Line 20 asks for the spectra in $\mu\mathrm{K}^2$ as raw C_ℓ (not D_ℓ), and line 21 keeps the total (lensed) spectra, an array with one row per ℓ and four columns TT, EE, BB, TE. Because this costs seconds per call, the result is cached (Section T0.7.2) and CAMB is called again only by the few scripts that vary the parameters (Section T1.3, Section 10.5.2).

healpy calls look like this round trip, from Section 7.11.1:

Listing T0.45: From $a_{\ell m}$ to a map and back: healpy as two black boxes (`code/ch07/24_camb_healpix.py`)

```
61 def roundtrip(lm, it):
62     """Draw a_lm up to lm, make the N_side = 256 map, analyse it again; return
       |C_map/C_input - 1|."""
63     a = lh.synalm(cl, lm, rng)
64     sky = hp.alm2map(a, NS, lmax=lm)                # = synfast with our own random numbers
65     c_in = lh.alm2cl(a, lm)
```

```
66      c_map = hp.anafast(sky, lmax=lm, iter=it)
67      return sky, a, c_in, np.abs(c_map[2:] / c_in[2:] - 1)
```

Line 63 draws the coefficients with the book's own generator (`lh.synalm`, written from scratch in Section 7.8.2). Line 64 is `hp.alm2map`: coefficients in, map of $N_{\text{side}} = 256$ out. Line 66 is `hp.anafast`: map in, $\widehat{C}_\ell$ out; `iter` sets how many refinement passes it makes on the transform. Line 67 compares the result with the spectrum of the input coefficients. The few other healpy calls in the book are of the same kind: `hp.nside2npix` (pixels from resolution), `hp.map2alm` (map to coefficients), `hp.smoothing` (convolve with a beam), `hp.pixwin` (the pixel window), `hp.mollview` (draw the sphere on a page). Each is a documented box, and the book checks each once against a case with a known answer before relying on it.

T0.9 Seven traps that catch everyone once

A program does exactly what it says, which is not always what its author meant. Most mistakes stop the script with an error message, and those are the friendly ones: Python names the line, and the fix is usually obvious. The dangerous mistakes are the ones that run to the end and print a plausible number. Take the smallest example. A script estimates the variance of each of M simulated experiments, stored as the rows of an $M \times N$ array `x`, by subtracting each row's mean. If it writes `x - x.mean(axis=1)`, the means form an array of shape `(M,)`, and NumPy lines that shape up against the *columns* of `x`, not its rows (Section T0.4.4). When $M \neq N$ the shapes do not fit and the script stops; when by bad luck $M = N$, it subtracts the mean of experiment j from reading j of every experiment, runs to the end, and prints a variance that is simply wrong. The good version, `x - x.mean(axis=1, keepdims=True)`, keeps the means as a column of shape `(M, 1)` and is correct for every M and N; its price is eleven characters. The lesson: the traps worth knowing are the silent ones.

There are seven of them, and each is a fact about how Python or NumPy stores numbers. The script `code/chT0/05_pitfalls.py` runs each one in two or three lines and prints what happens, and the book's own scripts show how each is avoided.

T0.9.1 Comparing real numbers

A computer stores a real number in binary with a fixed number of digits, so most decimal fractions are stored only approximately. 0.1 in binary is 0.000110011001100 . . ., repeating for ever, and it is cut off after 53 binary digits. The first lines of the script show the consequence:

```
# 1. floats
0.1 + 0.2 == 0.3          -> False
0.1 + 0.2 - 0.3           -> 5.551e-17
np.isclose(0.1 + 0.2, 0.3) -> True
```

The stored 0.1 + 0.2 and the stored 0.3 differ by 5.551×10^{-17}, so the test == says they are different. Nothing is broken: each number is correct to about sixteen significant digits, and they disagree in the seventeenth. The consequence is a rule. Real numbers computed in two different ways are compared with a tolerance, never with ==. `np.isclose(a, b)` asks whether $|a - b| \leq 10^{-8} + 10^{-5}\,|b|$ (both tolerances can be changed with `atol=` and `rtol=`), and `np.allclose` asks the same of every element of two arrays.

The book's scripts compare in exactly this way when they check one method against another. The exact binomial confidence interval of Section 4.10.4 is found twice, once by solving two equations with `brentq` and once from a closed formula that uses the beta distribution:

Listing T0.46: Two routes to the same interval, compared with a tolerance (code/ch04/11_test_inversion.py)

```
40 lo = optimize.brentq(lambda t: p_upper_tail(t) - alpha / 2, 1e-6, z / N)
41 hi = optimize.brentq(lambda t: p_lower_tail(t) - alpha / 2, z / N, 1 - 1e-6)
42 # the same numbers from the binomial-beta identity (a check)
43 lo_beta = stats.beta.ppf(alpha / 2, z, N - z + 1)
44 hi_beta = stats.beta.ppf(1 - alpha / 2, z + 1, N - z)
45 assert abs(lo - lo_beta) < 1e-6 and abs(hi - hi_beta) < 1e-6
```

Line 45 is the comparison: the two lower ends must agree to within 10^{-6}, and so must the two upper ends. The tolerance is set by the less accurate of the two routes, here the root finder, whose default stopping rule pins the root to about 10^{-12}; 10^{-6} leaves a wide margin while still catching any real mistake, which would show up in the second or third digit. How large a tolerance must be in general, and why, is the subject of Section T0.10.1.

T0.9.2 Whole numbers and real numbers

Python and NumPy keep two kinds of number apart: integers (`int`, whole numbers, exact) and floating-point numbers (`float`, real numbers, approximate). Two consequences surprise newcomers:

```
# 2. integer arrays
a = np.array([1, 2, 3]); a[0] = 0.7  -> a = [0 2 3], a.dtype = int64
7 / 2 -> 3.5,   7 // 2 -> 3
```

The first line is the dangerous one. `np.array([1, 2, 3])` is made from integers, so its `dtype` is `int64`, and an integer array can only hold integers: writing 0.7 into it silently stores 0, the integer part. The second line is the reassuring one: in Python 3, `/` always divides as real numbers, `7 / 2` is 3.5, and only the explicit floor division `//` throws the fraction away.

Integer arrays arise naturally in the book, because counts are integers: `rng.poisson` and `rng.integers` return them, and so does `np.histogram` for its counts. The iterative unfolding of Section 6.14.5 works on simulated Poisson counts, and converts them to real numbers the moment they are drawn:

Listing T0.47: Counts converted to real numbers before any arithmetic, and a default of `None` (code/ch06/30_unfolding.py)

```
172 def ibu(n, iters, prior=None):
173     """D'Agostini iterations; n may be (..., M).  Returns the estimate after `iters`
        steps."""
174     p = np.full(M, 1.0 / M) if prior is None else prior / prior.sum()
175     m_hat = np.broadcast_to(p * n.sum(axis=-1, keepdims=True), n.shape).copy()
176     for _ in range(iters):
177         nu_hat = m_hat @ R.T                                  # sum_j R_ij mu_j
178         m_hat = m_hat / eff * ((n / nu_hat) @ R)              # mu_j/eps_j sum_i R_ij n_i / nu_i
179     return m_hat
180
181
182 iters_grid = np.array([1, 2, 3, 4, 5, 6, 8, 10, 15, 20, 30, 50, 100, 200, 500])
183 toys = rng.poisson(nu, size=(1000, M)).astype(float)
```

Line 183 draws 1000 toy data sets of Poisson counts and immediately applies `.astype(float)`, which returns a copy of the array with `dtype float64`. The update of line 178 then divides these counts by the predicted ones and multiplies by efficiencies; with real numbers there is no risk that an intermediate result is cut to an integer anywhere along the way. (Lines 172 and 174 show another habit, the default value `None`, which Section T0.9.6 explains.) The rule: when an array will hold real numbers, make it a float array from the start, with `np.zeros(n)` (which is float by default), `np.array(..., float)` or `.astype(float)`.

T0.9.3 Views and copies

A slice of a NumPy array is not a new array. It is a **view** [a second name for part of the same block of memory]: `y = x[1:3]` makes `y` a window onto elements 1 and 2 of `x`, and writing through the window changes `x` (Figure T0.22). NumPy does this on purpose, because slicing a large array would otherwise copy it every time. The script shows both behaviours:

```
# 3. views and copies
x = np.zeros(5); y = x[1:3]; y[:] = 9  -> x = [0. 9. 9. 0. 0.]
same with y = x[1:3].copy()            -> x = [0. 0. 0. 0. 0.]
```

With the view, the two nines appear inside `x`; with `.copy()`, `y` is an independent array and `x` stays zero. The same is true of plain assignment: `y = x` gives the same array a second name and copies nothing. Python lists behave the same way for assignment, though their slices *are* copies.

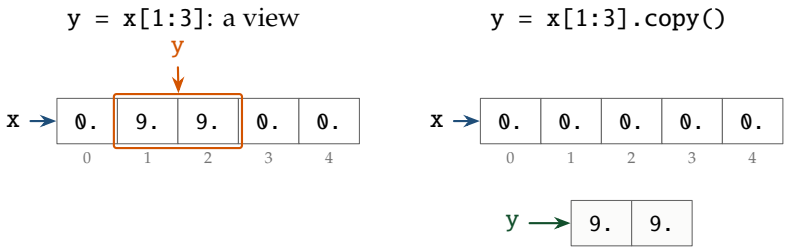


Figure T0.22: Two names for one block of memory, or two blocks. Left: the slice `y` is a window (orange) onto elements 1 and 2 of `x`, so `y[:] = 9` writes into `x`. Right: `.copy()` makes a new block (green) for `y`, and `x` is untouched.

Where does this matter in practice? When a loop records the history of something it keeps changing. Newton's method for the maximum-likelihood angle fit of Section 3.9.3 stores every step it takes, so that the path can be drawn:

Listing T0.48: Recording the path of an iteration: a copy at every step (`code/ch03/10_angular_ml.py`)

```
62 def newton(x, th0, tol=1e-10, maxit=50):
63     th, path = np.array(th0, float), [np.array(th0, float)]
64     for it in range(maxit):
65         g, H = grad_hess(th, x)
66         step = -np.linalg.solve(H, g)                # theta_{j+1} = theta_j - H^{-1} grad
67         th = th + step
68         path.append(th.copy())
69         if np.max(np.abs(step)) < tol:
70             break
71     return th, np.array(path)
```

Line 63 makes the starting point a float array (the previous trap) and starts the list `path` with a *separate* copy of it. Line 66 computes the Newton step, line 67 moves to the new point, and line 68 appends a copy of it to the list. Here line 67 happens to build a new array, because `th + step` always creates one, so the copy is a safeguard rather than a necessity. It becomes a necessity the moment someone "simplifies" line 67 to `th += step`: the operator += updates the existing array in place, and without the copy every entry of `path` would be the same array, so the plotted path would collapse to its final point. The rule: when an array is stored and then changed, store a `.copy()`.

T0.9.4 Where ranges and slices stop

Python counts from zero and stops *before* the end:

```
# 4. where ranges stop
list(range(5))            -> [0, 1, 2, 3, 4]
len(np.arange(0, 1, 0.1)) -> 10   (1.0 is not included)
len(np.linspace(0, 1, 11)) -> 11  (both ends included)
v = np.arange(10); v[2:5] -> [2 3 4],  v[-1] -> 9
```

`range(5)` is 0, 1, 2, 3, 4: five numbers, and 5 is not among them. A slice `v[2:5]` takes elements 2, 3, 4. `np.arange(0, 1, 0.1)` stops before 1 and has ten points, while `np.linspace(0, 1, 11)` includes both ends and has eleven. The convention has a reward: `v[a:b]` has exactly $b - a$ elements, and `v[:k]` and `v[k:]` split `v` into two pieces without overlap or gap. Its cost is that every formula written on paper with $i = 1, \ldots, N$ must be translated to `i = 0, ..., N-1`.

The commonest form of the trap in the book is the **fencepost** [n bins need $n + 1$ edges, as n fence panels need $n + 1$ posts] of every histogram (Figure T0.23). The density-histogram script of Section 1.9 builds its bins like this:

Listing T0.49: Bin edges with `arange`, and the centres between them (`code/ch01/11_density_histogram.py`)

```
34      E = E_all[:N]
35      bins = np.arange(mu - 4 * sigma, mu + 4 * sigma + w / 2, w)
36      counts, edges = np.histogram(E, bins=bins)
37      centres = 0.5 * (edges[1:] + edges[:-1])
```

Line 35 wants edges from $\mu - 4\sigma$ to $\mu + 4\sigma$ *inclusive*, in steps of the bin width w. Because `np.arange` stops before its end point, and because $8\sigma/w$ steps of a fractional w may land a hair above or below $\mu + 4\sigma$ by rounding, the end is given as $\mu + 4\sigma + w/2$: half a step beyond the last edge wanted, so that the last edge is included and nothing beyond it is. Line 36 returns the counts, one per bin, and the edges, one more than the counts. Line 37 makes the bin centres by averaging each edge with the next: `edges[1:]` drops the first edge and `edges[:-1]` drops the last, so the two arrays have the same length, the number of bins, and their average is the midpoint of every bin.

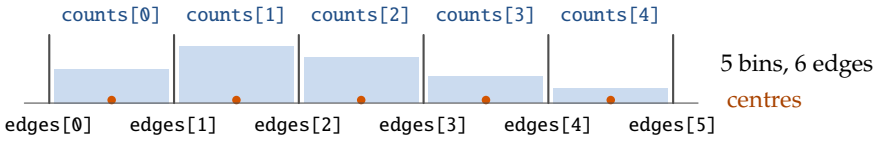


Figure T0.23: The fencepost. Five bins (blue) are bounded by six edges (black). `np.histogram` returns five counts and six edges; the five centres (orange dots) are the averages of `edges[:-1]` (the left edges) and `edges[1:]` (the right edges).

T0.9.5 Shapes that do not fit, or fit by accident

Broadcasting (Section T0.4.4) compares the shapes of two arrays from the right, and stretches any axis of length one. Two outcomes are printed by the script:

```
# 5. shapes
(r - r[:, None]).shape    -> (3, 3)
r + np.ones(4)            -> ValueError: operands could not be broadcast together with
    shapes (3,) (4,)
```

Subtracting a column (shape `(3, 1)`, made by `r[:, None]`) from a row (shape `(3,)`) gives the 3×3 table of all differences, which is often exactly what is wanted. Adding arrays of lengths 3 and 4 fails with a message that names both shapes. The failure is the friendly case. The unfriendly case is the one of the opening paragraph: two axes that happen to have the same length, so that the shapes fit even though the meaning does not.

The sample-variance study of Section 3.4.4 has exactly that structure, M simulated experiments of N readings each, stored as an $M \times N$ array, with N running over a list of sample sizes:

Listing T0.50: Each row minus its own mean: `keepdims=True` (code/ch03/04_sample_variance.py)

```
29 mse = {"N-1": [], "N": [], "N+1": []}
30 for N in Ns:
31     x = rng.standard_normal((M, N))                  # sigma = 1
32     ss = ((x - x.mean(axis=1, keepdims=True)) ** 2).sum(axis=1)
33     mean_biased.append((ss / N).mean())
```

Line 31 subtracts each row's mean from that row. `x.mean(axis=1)` alone would have shape `(M,)`, a row of M means, which broadcasting would line up against the N columns: an error if $M \neq N$, and a silently wrong answer if $M = N$. With `keepdims=True` the result keeps the averaged axis with length one, shape `(M, 1)`, a column, and broadcasting stretches it across each row. The habit that catches the rest: print `.shape` of every array in a new formula once, and check that it is what the paper formula says.

T0.9.6 A default value that remembers

A default value in a function definition (Section T0.3.5) is created *once*, when Python reads the `def` line, and shared by every call that does not supply that argument. For a number or a string this never matters, because they cannot be changed in place. For a list it does:

```
# 6. mutable default argument
append_bad(1); append_bad(2)   -> [1, 2]
append_good(1); append_good(2) -> [2]
```

`append_bad(value, store=[])` appends to its default list, and since that list is the same object at every call, the second call returns $[1, 2]$: it remembers the first. `append_good(value, store=None)` uses the conventional cure: the default is the special value `None`, and the first line of the function makes a fresh list whenever `store` is `None`. The unfolding function `ibu` of the excerpt in Section T0.9.2 uses the same pattern for an array: `prior=None` on line 172, and on line 174 a flat prior made fresh at each call when none is given. The book never uses a list, a dictionary or an array as a default value.

T0.9.7 Forgetting the seed

A generator made without a seed takes one from the computer's clock and its internal noise, so its numbers differ at every run:

```
# 7. seeds
default_rng() twice        -> equal: False
rng_for(same name) twice   -> equal: True  (0.858458)
```

Two generators made with `np.random.default_rng()` give different first draws; two made with `rng_for` and the same name give the same draw, 0.858458, at every run and on every machine (Section T0.5.1). The difference matters for three reasons. A number quoted in the text must be the number the script prints, every time it is rerun. A bug that appears for one random sample can be chased only if the same sample can be produced again. And two settings can be compared on the same random numbers, which makes their difference far more precise than either one (Section T0.10.5).

The census counts of Section T0.5.1 show how faithfully the book keeps the rule: of the 216 files that draw random numbers, 197 make their generator with `rng_for`, and nearly all of the

others are libraries that receive a generator from the script that calls them. The script above is one of the very few places where an unseeded generator is made on purpose, to show what it does.

Watch out: A trap that hides in an average

Each of the seven traps can hide inside a Monte Carlo average, where a wrong number looks like noise. The cure is the same for all of them: run the code once on a case whose answer is known, and compare with a tolerance (Section T0.10.7).

T0.10 Is this number real? Numerical noise and how to check

A script finishes and prints a number. It ran without an error, the number has the right size, and it carries twelve digits. How many of those digits mean anything? The question is not idle, because every number a computer produces has passed through three filters that each blur it. Every real number is stored with about sixteen significant digits and rounded after every operation. Every method that replaces calculus by arithmetic (a derivative by a difference, an integral by a sum, an equation by an iteration) makes an error of its own that depends on a step size or a tolerance. And every Monte Carlo number is an average of random draws, which a different seed would change.

The smallest example. The probability that a standard Gaussian variable exceeds 2 is $\mathrm{P}(Z > 2) = 0.02275$. A script estimates it from $N = 10^4$ draws as the fraction of draws above 2, and with the first seed prints 0.0226. Read naively, that is a number with three significant digits, and it does not match the truth to all three. Which of the three filters set the error? Rounding contributes about one part in 10^{16}, and the method (counting) is exact. The Monte Carlo filter is the whole story: the count of draws above 2 is binomial, so the fraction has standard deviation $\sqrt{q(1-q)/N} = 0.00149$ for $q = 0.02275$. The printed value should therefore be read as 0.0226 ± 0.00149: one and a half significant digits, and entirely consistent with the truth: the difference is -0.10 standard deviations. The naive reading would call a correct computation wrong; with only 10^2 draws it would just as readily call a wrong one right. The good habit, quoting every number with its error, costs one more line of code. *The lesson: a number without an error estimate has not yet been measured.*

The checks are a short list, one for each source of error: the floating-point facts that set the limit of every computation, the step-size balance of numerical derivatives, the reports and silences of SciPy's integrators, root finders and minimisers, the condition number of a matrix, the $1/\sqrt{N}$ of Monte Carlo, convergence studies, validation against known answers, and finally a single rule for deciding whether a difference between two numbers is real. Every one of them is already used somewhere in the book, and each is shown at work on a small problem by `code/chT0/08_trust.py`.

T0.10.1 Sixteen digits, and how to lose them

How a real number is stored. NumPy's `float64`, Python's `float`, is a number in binary scientific notation, packed into 64 bits (Figure T0.24):

$$x = (-1)^s \left(1 + \sum_{k=1}^{52} b_k\, 2^{-k}\right) 2^e, \qquad b_k \in \{0,1\}, \quad -1022 \le e \le 1023. \tag{T0.1}$$

One bit holds the sign s, eleven the exponent e, and fifty-two the binary digits b_k of the fraction, the **mantissa**. Between 1 and 2 (that is, $e = 0$) the numbers that can be stored are 1, $1 + 2^{-52}$, $1 + 2 \cdot 2^{-52}$, and so on: they are evenly spaced, and the gap is

$$\epsilon = 2^{-52} \approx 2.22 \times 10^{-16}, \tag{T0.2}$$

the **machine epsilon** [the gap between 1 and the next larger storable number]. Between 2^e and 2^{e+1} the same 2^{52} steps are stretched by 2^e, so the gap is $\epsilon\, 2^e$: the *relative* spacing is about ϵ everywhere. A real number is rounded to the nearest storable one, so its stored value is $x(1 + \delta)$ with $|\delta| \le \epsilon/2$, and so is the result of every arithmetic operation. That is what "sixteen significant digits" means: $\log_{10}(1/\epsilon) \approx 15.7$ [9]. NumPy reports these facts itself, through `np.finfo(np.float64)`:

```
eps = 2.2204e-16   largest = 1.7977e+308   smallest normal = 2.2251e-308
1.0 + eps == 1.0 -> False;   1.0 + eps/2 == 1.0 -> True
0.1 + 0.2 == 0.3 -> False;   np.isclose -> True
```

$1 + \epsilon$ is the next number after 1, so it differs from 1; $1 + \epsilon/2$ lies exactly halfway and is rounded back to 1. The largest storable number is about 1.798×10^{308} and the smallest with full precision about 2.225×10^{-308}, so overflow and underflow are rare in practice, though not impossible (Section T0.11).

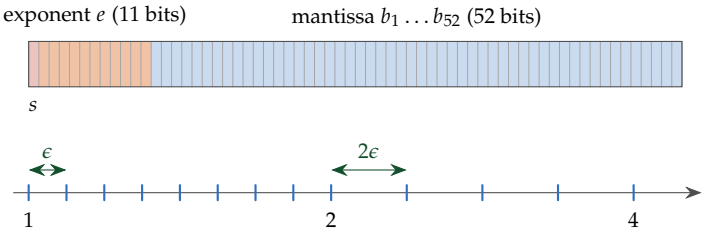


Figure T0.24: A `float64`. Top: one sign bit, eleven exponent bits and fifty-two mantissa bits. Bottom: the storable numbers near 1, with the spacing exaggerated (there are really 2^{52} of them between 1 and 2). The gap is ϵ between 1 and 2 and doubles at every power of two, so the relative gap stays near ϵ.

Comparing with a tolerance. Because every result carries a relative error of order ϵ, two routes to the same quantity agree to about ϵ times its size, not exactly, and the comparison must be relative. The dark-matter halo of Section T2.9 is built from its virial mass and radius, and the script checks that the mass enclosed by the profile at r_{200} really is the mass put in:

Listing T0.51: A relative comparison: the ratio minus one (`code/chT2/04_dm_profile.py`)

```
40 # check: the mean density inside r200 is 200 rho_crit
41 M_nfw = lambda r: 4 * np.pi * rho_s * rs**3 * (np.log(1 + r / rs) - (r / rs) / (1 + r / rs))
42 assert abs(M_nfw(r200) / M200 - 1) < 1e-10
```

Line 41 is the enclosed mass of the profile, from its closed form, and line 42 compares the ratio of the two masses with 1. Written as $|M/M_{200} - 1| < 10^{-10}$ the test is relative: it does not care whether masses are measured in grams or in solar masses, while `abs(M - M200) < 1e-10` would be meaningless for numbers of order 10^{12}. The tolerance sits well above the rounding level (a cube root, two logarithms and a few products, each adding about ϵ) and far below any error that would matter.

Cancellation. Rounding errors of order ϵ are harmless as long as they stay of order ϵ. One operation can magnify them enormously: the subtraction of two nearly equal numbers. Suppose

a and b are stored as $a(1+\delta_a)$ and $b(1+\delta_b)$, with $|\delta_a|, |\delta_b| \le \epsilon/2$. What is the relative error of the computed difference?

$$\frac{|a(1+\delta_a) - b(1+\delta_b) - (a-b)|}{|a-b|} \overset{(1)}{=} \frac{|a\delta_a - b\delta_b|}{|a-b|} \overset{(2)}{\le} \frac{\epsilon}{2}\,\frac{|a|+|b|}{|a-b|}. \tag{T0.3}$$

(1) Expand: the exact $a-b$ cancels, and the two rounding errors remain.
(2) The triangle inequality, $|a\delta_a - b\delta_b| \le |a|\,|\delta_a| + |b|\,|\delta_b|$, and $|\delta| \le \epsilon/2$.

The inputs' relative error $\epsilon/2$ is multiplied by $(|a|+|b|)/|a-b|$, which is about $2a/(a-b)$ when $a \approx b$. If a and b agree in their first ten digits, ten of the sixteen digits are lost: the result is the difference of two numbers whose leading digits cancelled, leaving the uncertain trailing ones. This is **catastrophic cancellation**, and it needs no long calculation: a single subtraction is enough.

Example: $1-\cos x$ for small x. The ratio $(1-\cos x)/x^2$ tends to $\frac{1}{2}$ as $x \to 0$. At $x = 10^{-5}$, $\cos x \approx 1 - 5\times10^{-11}$, so the subtraction $1-\cos x$ cancels ten digits. Since 1 is stored exactly and $\cos x$ with a relative error of at most $\epsilon/2$, (T0.3) bounds the relative error of the result by $(\epsilon/2)/(x^2/2) = \epsilon/x^2 \approx 2\times10^{-6}$. The script prints:

```
x = 1e-5: (1 - cos x)/x^2 = 0.5000000414,  2 sin^2(x/2)/x^2 = 0.5000000000
x = 1e-8: naive = 0.0,  rewritten = 0.5000000000
```

At $x = 10^{-5}$ the naive formula gives 0.5000000414, wrong in the eighth digit, within the bound. At $x = 10^{-8}$ it gives exactly 0. There $\cos x = 1 - 5\times10^{-17}$; just below 1 the storable numbers are spaced by $\epsilon/2 \approx 1.1\times10^{-16}$, and 5×10^{-17} is less than half of that spacing, so $\cos x$ is stored as exactly 1 and the difference vanishes. The cure is to rewrite the formula so that nothing nearly equal is subtracted. The double-angle identity $\cos x = 1 - 2\sin^2(x/2)$ gives $1-\cos x = 2\sin^2(x/2)$, a product of small numbers each known to sixteen digits, and the rewritten form returns 0.5000000000 at both values. Figure T0.25 (left) shows the relative error of both forms against x: the naive one follows the bound ϵ/x^2 up to order one, and the rewritten one stays at the rounding floor.

Example: the variance in one pass. The variance of a set of readings is $\langle x^2\rangle - \langle x\rangle^2$, and that formula needs only one pass through the data. Try it on 10^4 readings $10^8 + z_i$, with z_i standard Gaussian: times in nanoseconds measured from some epoch would look like this. Both $\langle x^2\rangle$ and $\langle x\rangle^2$ are about 10^{16}, and near 10^{16} the storable numbers are spaced by $\epsilon\cdot 2^{53} = 2$ (since $2^{53} \approx 9\times10^{15}$). Each of the two terms is therefore known only to the nearest even number, and their difference, which should be about 1, can only come out as a multiple of 2:

```
variance of 1e8 + N(0,1): one-pass = 2.0000,  two-pass = 0.9919,  np.var = 0.9919
```

The one-pass formula returns 2.00; the two-pass form, which first subtracts the mean and then squares, returns 0.9919, and so does `np.var`, which works the same way. By (T0.3) the amplification factor is about 2×10^{16}, enough to destroy all sixteen digits. The book always computes variances in two passes: the sample-variance excerpt of Section T0.9.5 subtracts each row's mean before squaring (line 31 there), and elsewhere `np.var` and `np.cov` do it internally. *The lesson: never subtract two large, nearly equal numbers; rewrite the formula, or subtract the large common part first.*

T0.10.2 Derivatives: the step that is neither too large nor too small

Fisher matrices (Chapter 10), emulators (Section 9.20) and Newton's method all need derivatives of functions that are known only as code. The obvious approach is the definition of the

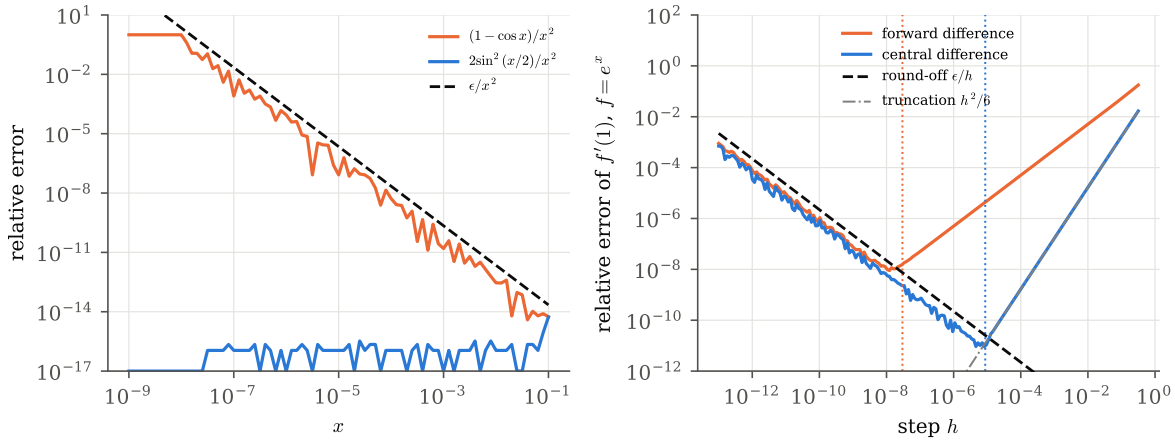


Figure T0.25: Where digits go. Left: relative error of $(1-\cos x)/x^2$ computed naively (orange) and as $2\sin^2(x/2)/x^2$ (blue). The naive form loses digits along the dashed line ϵ/x^2 and is useless below $x \approx 10^{-8}$; the rewritten form stays at the rounding floor. Right: relative error of the numerical derivative of e^x against the step h, for the forward and central differences, as the root-mean-square over 101 points between 0.5 and 1.5. To the right of each minimum the truncation error grows (grey line, $h^2/6$ for the central difference); to the left the round-off error grows as ϵ/h (dashed). Dotted verticals: the best steps predicted by (T0.5) and its forward analogue.

derivative with a small but finite step h. How small? Intuition says "as small as possible", and Figure T0.25 (right) shows that intuition to be wrong: the error falls as h shrinks, reaches a minimum, and then rises again. Two errors compete, and the best step is where they balance.

The truncation error. Take the **central difference** $D(h) = [f(x+h) - f(x-h)]/2h$. Taylor's theorem gives $f(x \pm h) = f \pm hf' + \frac{1}{2}h^2 f'' \pm \frac{1}{6}h^3 f''' + \dots$, all derivatives at x. Subtracting the two expansions cancels the even powers, and dividing by $2h$ leaves

$$D(h) = f'(x) + \frac{h^2}{6} f'''(x) + O(h^4). \tag{T0.4}$$

The error $h^2 f'''/6$ shrinks as h shrinks: this is the part intuition sees.

The round-off error. The two function values are computed with relative errors of order ϵ, so each is off by up to about $\epsilon\,|f|$, and in the worst case the two errors have opposite signs and add. Their difference is then off by $2\epsilon\,|f|$, and dividing by $2h$ leaves $\epsilon\,|f|\,/h$. This part *grows* as h shrinks: for a tiny h, $f(x+h)$ and $f(x-h)$ are nearly equal, and their difference is the cancellation of (T0.3).

The balance. The total error and its minimum follow in three lines:

$$\begin{aligned} E(h) &\overset{(1)}{=} \frac{\epsilon\,|f|}{h} + \frac{h^2}{6}\,|f'''|\,, \\ \frac{\mathrm{d}E}{\mathrm{d}h} &\overset{(2)}{=} -\frac{\epsilon\,|f|}{h^2} + \frac{h}{3}\,|f'''| = 0 \quad\Longrightarrow\quad h_{\mathrm{opt}} \overset{(3)}{=} \left(\frac{3\epsilon\,|f|}{|f'''|}\right)^{1/3} \sim \epsilon^{1/3}, \\ E(h_{\mathrm{opt}}) &\overset{(4)}{=} \frac{3}{2}\,\frac{\epsilon\,|f|}{h_{\mathrm{opt}}} \sim \epsilon^{2/3}. \end{aligned} \tag{T0.5}$$

(1) Round-off plus truncation, the two estimates above.
(2) The smallest error is where the slope of E vanishes.

(3) Multiply by $3h^2/|f'''|$: $h^3 = 3\epsilon\,|f|\,/|f'''|$, and take the cube root.

(4) At $h_{\rm opt}$, $\frac{h^2}{6}\,|f'''| = \frac{h^3}{6h}\,|f'''| = \frac{3\epsilon|f|}{6h} = \frac{1}{2}\,\frac{\epsilon|f|}{h}$: the truncation term is half the round-off term, and the sum is $\frac{3}{2}$ of it. Since $h_{\rm opt} \sim \epsilon^{1/3}$, this is $\epsilon/\epsilon^{1/3} = \epsilon^{2/3}$.

For $f = e^x$, where $f = f'''$, the best step is $(3\epsilon)^{1/3} = 8.733\times10^{-6}$ and the best relative error about $\epsilon^{2/3} \approx 3.814\times10^{-11}$: two-thirds of the sixteen digits, about ten, survive. The same reasoning for the **forward difference** $[f(x+h)-f(x)]/h = f' + \frac{1}{2}hf'' + \ldots$, whose truncation error is first order in h, gives $E = 2\epsilon\,|f|\,/h + \frac{1}{2}h\,|f''|$, $h_{\rm opt} = 2\sqrt{\epsilon\,|f|\,/|f''|} \sim \epsilon^{1/2}$ and an error of order $\epsilon^{1/2}$: only eight digits. The script scans h and finds:

```
best h: forward 1.2e-08 (predicted 3.0e-08),  central 8.2e-06 (predicted 8.7e-06)
smallest error: forward 9.0e-09 (predicted 3.0e-08),  central 9.0e-12 (predicted 3.8e-11)
```

The measured best steps sit where (T0.5) puts them. The measured errors are a few times smaller than predicted, because the prediction assumed the worst case, with the two rounding errors adding; on average they partly cancel.

The plateau check. For a function whose third derivative is unknown, (T0.5) cannot be evaluated. The practical test is to compute the derivative for steps a factor of ten apart and look for a **plateau**, a range of h over which the result does not change:

```
h = 1e-02:  central difference = 2.718327133383
h = 1e-03:  central difference = 2.718282281506
h = 1e-04:  central difference = 2.718281832990
h = 1e-05:  central difference = 2.718281828518
h = 1e-06:  central difference = 2.718281828296
h = 1e-08:  central difference = 2.718281821856
h = 1e-10:  central difference = 2.718281155722
h = 1e-12:  central difference = 2.718492098097
```

The exact value is $e = 2.718281828459$. From $h = 10^{-2}$ to 10^{-4} the digits settle (truncation shrinking a hundredfold per step); at $h = 10^{-5}$ and 10^{-6} the results agree with each other to ten digits; below that, round-off takes over, and at $h = 10^{-12}$ only four digits survive. The digits that stay fixed while h changes by a factor of ten are the ones to trust.

The book chooses its steps in exactly this spirit. The emulator of Section 9.20.2 replaces CAMB by a second-order Taylor expansion in the six cosmological parameters, and its first derivatives are central differences with a fixed step for each parameter:

Listing T0.52: Central differences of $\ln C_\ell$, one parameter at a time (`code/ch09/lib_cmbemu.py`)

```
57      E = np.eye(d) * STEPS
58      plus = [lnc(FID + E[i]) for i in range(d)]
59      minus = [lnc(FID - E[i]) for i in range(d)]
60      D = np.array([(plus[i] - minus[i]) / (2 * STEPS[i]) for i in range(d)])
```

Line 57 builds a diagonal matrix whose row i moves parameter i alone by its step (the array `STEPS`, set on line 29 of the same file to 0.02 in $\ln(10^{10}A_s)$, 0.01 in n_s, 0.5 in H_0, and so on). Lines 58–59 run CAMB at each shifted point, and line 60 is (T0.4). These steps are far larger than $\epsilon^{1/3}$, and rightly so. CAMB's output is not accurate to sixteen digits but to perhaps six, so its ϵ is about 10^{-6}, and (T0.5) then asks for steps of order 10^{-2} in relative terms. Section 10.5.1 redoes the balance with such a "code noise" and with the four-point rule, and Section 10.5.2 runs the plateau check on CAMB itself, multiplying all six steps by factors from $\frac{1}{64}$ to 4 and confirming that the forecast does not move over the middle of that range. *The lesson: a numerical derivative is trustworthy only on a plateau; find it by changing h by factors of ten.*

T0.10.3 Integrals, roots and minima: what the black boxes report

`quad`, `brentq` and `minimize` (Section T0.6) are iterations that stop when an internal test is satisfied. Each one reports something about its own reliability, and each can be fooled. The three ways they go wrong are different: `quad` can be confidently wrong, `brentq` refuses loudly, and `minimize` stops quietly.

quad is only as good as the points it looked at. `quad` returns a value and an error estimate (Section T0.6.2). The estimate is the disagreement between two integration rules applied to the same function values, and so it says nothing about regions where the function was never evaluated. The script integrates a Gaussian of width 0.01 centred at 37.3, whose integral is exactly 1, over $[0, 100]$, and then tries four remedies:

```
quad narrow peak on [0, 100]: 0.000e+00 (claimed error 0.0e+00);  tighter tolerance:
    0.000e+00;  points=[37.3]: 4.951e-14;  points=[37.25, 37.35]: 0.9999994267 (claimed
    8.7e-10);  on [37.2, 37.4]: 1.0000000000 (claimed 8.7e-10)
```

Over $[0, 100]$ `quad` first applies a 21-point rule to the whole interval. No point lands within the peak, every value is zero or astronomically small, both rules agree on zero, and `quad` reports 0 with a claimed error of 0. Tightening the tolerance changes nothing, because by `quad`'s own estimate the answer is already exact. Telling it that something happens at 37.3, with `points=[37.3]`, still fails, and instructively: the peak now sits on the boundary between two pieces, and the integration rules never evaluate the function at the ends of a piece. An interval exactly as wide as the peak, `points=[37.25, 37.35]`, nearly works, and yet the answer is too small by 5.7×10^{-7} while the claimed error is 10^{-9}: the missing amount is the probability beyond $\pm 5\sigma$ (twice the 2.867×10^{-7} of Section T0.6.1), which lives in the two long outer pieces where `quad` once again saw nothing. Only an interval that contains the whole peak with room to spare gives 1.0000000000.

The book's integrals are chosen with this in mind: the limits and break points are put where the function lives. The posterior mean of a location parameter with a Student-t likelihood, in the self-normalised importance-sampling study of Section 6.12.4, is one ratio of two integrals:

Listing T0.53: Integration limits wide enough for heavy tails, and a larger subdivision limit (`code/ch06/27_importance.py`)

```
114 xdat = np.array([-4.0, 0.5, 1.0])
115 like = lambda th: np.prod(stats.t.pdf(xdat[:, None] - th[None, :], df=3), axis=0)
116 num = integrate.quad(lambda t: t * like(np.array([t]))[0], -60, 60, limit=400)[0]
117 den = integrate.quad(lambda t: like(np.array([t]))[0], -60, 60, limit=400)[0]
```

Line 115 is the likelihood of three data points at -4, 0.5 and 1, a product of three t densities with three degrees of freedom. The range $[-60, 60]$ is far wider than the data, because t tails fall only as a power of the distance and a short range would cut off a noticeable part of the integral; `limit=400` lets `quad` subdivide up to 400 times instead of the default 50, so that it can resolve the structure near the data within that wide range. The checks for any `quad` result follow from the example: plot the integrand once; tighten `epsabs` and `epsrel` and see whether the value moves; split the range or change its ends and see whether the value moves. A value that moves was never as accurate as claimed.

brentq needs a bracket, and says so. `brentq(f, a, b)` requires $f(a)$ and $f(b)$ to have opposite signs (Section T0.6.3). If they do not, it stops at once:

```
brentq on [0, 1] for x^2 - 2: ValueError: f(a) and f(b) must have different signs
brentq on [0, 2]: 1.414213562373136   (sqrt 2 = 1.414213562373095)
```

On $[0, 1]$ the function $x^2 - 2$ is negative at both ends, so there is no guaranteed root and `brentq` refuses; on $[0, 2]$ it finds $\sqrt{2}$ to fourteen digits. This failure is friendly. The care goes into choosing the bracket so that it contains exactly the root wanted: the exact binomial interval of Section T0.9.1 (lines 40–41 of that excerpt) brackets the lower end between 10^{-6} and the estimate z/N, and the upper end between z/N and $1 - 10^{-6}$, because each tail probability crosses its target exactly once on its own side of the estimate.

`minimize` can stop early, and reports it only if asked. A minimiser stops when its steps become small, when it runs out of iterations, or when it can no longer find a way downhill. Only the first is success. The script minimises the five-dimensional Rosenbrock function, a long curved valley whose minimum is at $(1, 1, 1, 1, 1)$, twice:

Listing T0.54: A minimiser stopped early, and one allowed to finish (`code/chT0/08_trust.py`)

```
105 rosen = optimize.rosen                                    # a curved valley, minimum at (1, 1,
        1, 1, 1)
106 x_start = np.zeros(5)
107 short = optimize.minimize(rosen, x_start, method="Nelder-Mead", options=dict(maxiter=200))
108 full = optimize.minimize(rosen, x_start, method="BFGS", jac=optimize.rosen_der)
109 say(f"Nelder-Mead, 200 iterations: success = {short.success}, message = '{short.message}', "
110     f"x = {np.round(short.x, 3)}, |grad| =
        {np.linalg.norm(optimize.rosen_der(short.x)):.2e}")
111 say(f"BFGS with gradient: success = {full.success}, x = {np.round(full.x, 6)}, "
112     f"|grad| = {np.linalg.norm(optimize.rosen_der(full.x)):.2e}")
```

Line 107 runs Nelder–Mead with at most 200 iterations, line 108 runs BFGS with the exact gradient, and lines 109–112 print, for each, the flag `success`, the point found, and the length of the gradient there:

```
Nelder-Mead, 200 iterations: success = False, message = 'Maximum number of iterations has
    been exceeded.', x = [ 0.714  0.497  0.248  0.065 -0.023], |grad| = 7.60e+00
BFGS with gradient: success = True, x = [1. 1. 1. 1. 1.], |grad| = 5.00e-06
```

The first result looks like an answer, a point with a function value of 1.868, but `success` is `False`, the message says the iteration limit was reached, and the gradient there has length 7.598, far from zero. At a true minimum the gradient vanishes, as it does for BFGS, 5.001×10^{-6}. Three checks catch an early stop: read `res.success` and `res.message`; evaluate the gradient (numerically if need be) at `res.x`; and restart from `res.x`, or from a different starting point, to see whether the answer moves. For a likelihood, also check that the curvature there gives a sensible error bar (Chapter 3).

T0.10.4 Linear algebra: the condition number

The question. Solving $A\boldsymbol{x} = \boldsymbol{b}$ is the most common linear-algebra task in the book: every χ^2, every Gaussian likelihood and every Fisher matrix contains one. The data $\boldsymbol{b}$ is never exact: it carries at least rounding error, usually measurement noise. How much does a small relative change of $\boldsymbol{b}$ change $\boldsymbol{x}$?

The smallest example. Take $A = \mathrm{diag}(1,\ 10^{-6})$. The equation splits into $x_1 = b_1$ and $x_2 = 10^6\, b_2$. A change of $\boldsymbol{b}$ by 10^{-16} in its second component changes x_2 by 10^{-10}, while $\boldsymbol{x}$ itself may be of order 1: a relative change 10^6 times that of $\boldsymbol{b}$. The matrix stretches some directions a million times more than others, and inverting it magnifies whatever lies along the weak direction. That ratio of stretches is the whole story.

The general bound. For any perturbation $\delta\boldsymbol{b}$ the solution changes by $\delta\boldsymbol{x} = A^{-1}\delta\boldsymbol{b}$. Write $\|A\|$ for the largest factor by which A stretches any vector (its largest singular value, which for a symmetric positive-definite matrix is its largest eigenvalue), so that $\left\|A^{-1}\right\|$ is one over the smallest. Then

$$\frac{\|\delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \overset{(1)}{=} \frac{\left\|A^{-1}\delta\boldsymbol{b}\right\|}{\|\boldsymbol{x}\|} \overset{(2)}{\le} \frac{\left\|A^{-1}\right\|\,\|\delta\boldsymbol{b}\|}{\|\boldsymbol{x}\|} \overset{(3)}{\le} \left\|A^{-1}\right\|\,\|A\|\,\frac{\|\delta\boldsymbol{b}\|}{\|\boldsymbol{b}\|} \equiv \kappa(A)\,\frac{\|\delta\boldsymbol{b}\|}{\|\boldsymbol{b}\|}. \tag{T0.6}$$

(1) $A(\boldsymbol{x}+\delta\boldsymbol{x}) = \boldsymbol{b}+\delta\boldsymbol{b}$ and $A\boldsymbol{x} = \boldsymbol{b}$ give $\delta\boldsymbol{x} = A^{-1}\delta\boldsymbol{b}$.
(2) No vector is stretched by A^{-1} more than $\left\|A^{-1}\right\|$.
(3) $\|\boldsymbol{b}\| = \|A\boldsymbol{x}\| \le \|A\|\,\|\boldsymbol{x}\|$, so $1/\|\boldsymbol{x}\| \le \|A\|\,/\|\boldsymbol{b}\|$.

The **condition number** $\kappa(A) = \|A\|\,\left\|A^{-1}\right\|$ is the ratio of the largest to the smallest stretch, $\lambda_{\max}/\lambda_{\min}$ for a symmetric positive-definite matrix, and 10^6 in the example. NumPy computes it as `np.linalg.cond(A)`. Since $\delta\boldsymbol{b}$ is at least rounding, of relative size ϵ, the solution can carry a relative error up to $\kappa\epsilon$. In digits: *about* $\log_{10}\kappa$ *of the sixteen digits are lost*. The tour of Section T0.5.2 solved a 10×10 Hilbert system, whose condition number is 1.602×10^{13}: thirteen digits lost, three left. The recovered $\boldsymbol{x}$ was indeed wrong by 2.272×10^{-5}, inside the bound $\kappa\epsilon/2 \approx 2\times10^{-3}$, whether computed with `solve` or with `inv`. No algorithm can do better than the problem allows. What `solve` does achieve is a tiny residual, 8.502×10^{-17}: its $\boldsymbol{x}$ solves a problem extremely close to the one posed, which is the most any method can promise. `inv` followed by a product does not even achieve that (3.584×10^{-5}), the second reason, after speed, for the rule "solve, do not invert".

The book measures condition numbers where it matters. The mode-coupling matrix of the pseudo-C_ℓ method (Section 8.11.2) must be inverted for every mask, and the script that builds it records its condition number for each:

Listing T0.55: The condition number of the coupling matrix, measured (`code/ch08/22_coupling.py`)

```
74      sub = M[2:, 2:]
75      cond = np.linalg.cond(sub)
```

Line 74 keeps the multipoles $\ell \ge 2$ (the monopole and dipole are removed from the maps), and line 75 computes κ. For the full sky it is 1.1, for the 70% cap 1.8: harmless. For the 30% and 10% caps it is 1.3×10^{16} and 1.3×10^{17}, beyond $1/\epsilon$: all sixteen digits lost, the matrix singular to machine precision. Section 8.11.2 explains why, and what is done instead (binning, which keeps only the well-determined combinations).

Cholesky failing is a message. `np.linalg.cholesky` succeeds only for a symmetric matrix whose eigenvalues are all positive, and stops with `LinAlgError` otherwise. When a covariance matrix fails it, the matrix is telling us something. The script estimates the covariance of a 50-bin data vector from 30 and from 200 simulations:

```
30 simulations of 50 bins: smallest eigenvalue -1.3e-15, rank 29, cholesky fails
    (LinAlgError)
200 simulations of 50 bins: smallest eigenvalue 2.7e-01, rank 50, cholesky succeeds
```

Why does 30 fail? The sample covariance $\hat{\mathbf{C}} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^{\mathsf{T}}$ is a sum of n matrices of the form $\boldsymbol{u}\boldsymbol{u}^{\mathsf{T}}$, each of which stretches only along its own vector $\boldsymbol{u}$. And the n vectors $\boldsymbol{x}_i-\bar{\boldsymbol{x}}$ are not independent: they sum to zero, so any one is minus the sum of the others. So $\hat{\mathbf{C}}$ stretches at most $n-1$ directions and is exactly zero along the remaining $p-(n-1)$. With $n=30$ and $p=50$ the rank is 29, twenty-one eigenvalues are zero (computed as $\pm10^{-15}$, rounding around zero), and the Cholesky factor does not exist. With $n=200$ all 50 eigenvalues are positive and it succeeds. The failure says: not enough simulations to estimate this covariance. Even when

$n > p$ the inverse of $\hat{\mathbf{C}}$ is biased, and the Hartlap factor $(n - p - 2)/(n - 1)$ of Section 8.13.2 corrects it; Section 8.13 asks how many simulations are enough.

A script that might meet such a matrix catches the error instead of crashing. The bootstrap of Hubble's 1929 galaxies (Section 11.14.3) refits a model with a solar-motion term to each resampled set of twenty-four galaxies:

Listing T0.56: A resample whose fit is singular is recorded as missing, not hidden (`code/ch11/08_hubble1929.py`)

```
119     for i in range(nb):
120         j = rng.integers(0, N, N)
121         Kb_a[i] = fit_origin(r[j], v[j])[0]
122         try:
123             Kb_b[i] = fit_solar(r[j], v[j], ra[j], dec[j])[0][0]
124         except np.linalg.LinAlgError:
125             Kb_b[i] = np.nan
126     Kb_b = Kb_b[np.isfinite(Kb_b)]
```

Line 120 draws N galaxy indices with replacement, and line 121 refits the simple model. A resample can repeat a few galaxies so often that their directions on the sky no longer determine the four parameters of the fit, and the linear solve inside `fit_solar` (`np.linalg.solve` on the normal equations) then fails. Lines 122–125 catch exactly that error and record the resample as `np.nan`; line 126 removes those before any statistic is formed, and the number removed can be printed. *The lesson:* $\log_{10}\kappa$ *digits are lost to the matrix; a failed Cholesky factor is a diagnosis, not a nuisance.*

T0.10.5 Monte Carlo noise: every simulated number has an error bar

A Monte Carlo estimate is an average of N independent draws, so its standard deviation is $\sigma/\sqrt{N}$, with σ the standard deviation of a single draw: the variance of an average is the single-draw variance divided by N, derived in Section 6.11.2. The rule that follows is simple: every simulated number is quoted with its $\sigma/\sqrt{N}$, which is estimated from the same draws. The first Monte Carlo of the book, the estimate of π from random points in a square (Section 0.6), does exactly that:

Listing T0.57: A Monte Carlo estimate and its standard error, side by side (`code/ch00/06_reproducible_seeds.py`)

```
21 # a one-line Monte Carlo: fraction of darts in the unit quarter-disc -> pi/4
22 N = 1_000_000
23 x, y = r1.random(N), r1.random(N)
24 frac = np.mean(x**2 + y**2 < 1.0)
25 pi_hat = 4 * frac
26 pi_err = 4 * np.sqrt(frac * (1 - frac) / N)       # binomial error of a fraction, times 4
```

Line 24 is the fraction of points inside the quarter disc, line 25 the estimate $\hat{\pi} = 4\times$fraction, and line 26 its standard error: a fraction of N independent yes/no draws has variance $q(1-q)/N$, and multiplying by 4 multiplies the standard deviation by 4. With $N = 10^6$ the script finds $\hat{\pi} = 3.143 \pm 0.001641$, which is within one standard error of π.

Rerunning with a different seed. The error bar is a prediction: other seeds should scatter by about that much. That prediction can be tested directly. The script estimates $\mathrm{P}(Z > 2)$ with $N = 10^4$ draws for eight different streams:

```
P(Z>2) exact 0.02275;  N = 10^4, eight seeds: [0.0226 0.0203 0.0229 0.0196 0.0251 0.0229
    0.0227 0.0223];  predicted spread 0.00149, observed 0.00169
```

The eight estimates scatter by 0.00169 against the predicted 0.00149, the agreement expected for eight values, whose own standard deviation is uncertain by about $1/\sqrt{2\times 7} \approx 27\%$ (Section 7.12.3). Figure T0.26 (left) shows the same for N from 10^2 to 10^6: the estimates converge onto the exact value inside a band that narrows as $1/\sqrt{N}$. Each factor of ten in accuracy costs a factor of a hundred in draws.

Common random numbers. Often the question is not a number but a difference: how much does a probability change when a parameter changes? Two independent Monte Carlo runs each carry their own noise, and the noise of the difference is larger than either. The cure is to run both settings on the *same* random draws. Why that helps follows from the variance of a difference of two estimates A and B:

$$\operatorname{Var}(A-B) \overset{(1)}{=} \operatorname{Var} A + \operatorname{Var} B - 2\operatorname{Cov}(A,B). \tag{T0.7}$$

(1) Expand $\mathbb{E}[(A - B - \mathbb{E}\,A + \mathbb{E}\,B)^2]$: the two squares give the variances, and the cross term gives $-2\operatorname{Cov}(A,B)$.

With independent draws, $\operatorname{Cov}(A,B) = 0$ and the variances add. With the same draws, the two estimates rise and fall together, $\operatorname{Cov}(A,B)$ is positive and nearly as large as the variances, and most of the noise cancels in the difference. This is the method of **common random numbers**. The script compares $\mathrm{P}(X > 2)$ for $X \sim \mathcal{N}(\mu, 1)$ at $\mu = 0$ and at $\mu = 0.1$, repeating the comparison 400 times with 10000 draws each:

Listing T0.58: The same change estimated with fresh draws, and with the same draws shifted (`code/chT0/08_trust.py`)

```
147 delta, M, R = 0.1, 10_000, 400
148 diff_ind, diff_crn = [], []
149 r5 = rng_for("chT0", "08_trust", 99)
150 for _ in range(R):
151     z1, z2 = r5.standard_normal(M), r5.standard_normal(M)
152     diff_ind.append(np.mean(z2 + delta > 2) - np.mean(z1 > 2))    # fresh draws for the
        second setting
153     diff_crn.append(np.mean(z1 + delta > 2) - np.mean(z1 > 2))    # the same draws, shifted
```

Line 151 draws two independent sets of standard Gaussians. Line 152 estimates the change with fresh draws for the second setting; line 153 uses the first set for both settings, shifting it by $\delta = 0.1$ for the second. The exact change is 0.00597. Its scatter over the repetitions is 0.00224 with independent draws and 0.00076 with common ones, a variance 9 times smaller, worth 9 times as many draws (Figure T0.26, right). With independent draws a single comparison measures the change at only 2.7 standard deviations; with common random numbers, at 7.9.

The book uses common random numbers wherever it compares settings. The CMB pipeline of Section 8.5 computes seven different spectra from the *same* simulated sky, so that the bias of one estimator relative to another is measured to a fraction of a per cent. The simulated Fisher matrices of Section 12.20.2 and Section 12.26.1 take each derivative from maps that share their random seed, so that the maps differ only through the change of cosmology. *The lesson: quote $\sigma/\sqrt{N}$ with every simulated number, and compare settings on the same random numbers.*

T0.10.6 Convergence: change the resolution until nothing changes

Every computation in the book has a resolution: a grid spacing, a number of points per axis, a HEALPix N_{side}, a maximum multipole $\ell_{\max}$, a number of simulations or of chain steps. The answer at finite resolution differs from the answer the resolution is meant to approximate. The test is a **convergence study**: compute at two or three resolutions, and accept the answer only when it changes by less than its other errors.

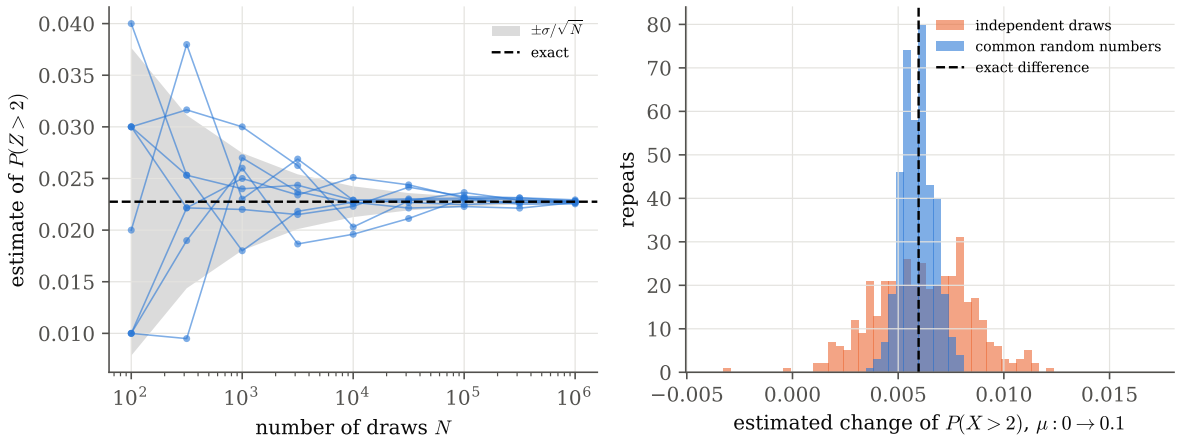


Figure T0.26: Monte Carlo noise. Left: estimates of $\mathrm{P}(Z > 2)$ from eight independent streams (blue lines) against the number of draws; the grey band is the predicted $\pm\sigma/\sqrt{N}$ around the exact value (dashed). Right: the estimated change of $\mathrm{P}(X > 2)$ when the mean moves from 0 to 0.1, over 400 repetitions of 10^4 draws each. Independent draws for the two settings (orange) scatter about three times more widely than common random numbers (blue) around the exact change (dashed).

Example: the trapezoid rule. The integral $\int_0^\pi \sin x \, \mathrm{d}x = 2$ computed by the trapezoid rule with n intervals:

```
trapezoid n =  8: 1.9742316019   error -2.58e-02
trapezoid n = 16: 1.9935703438   error -6.43e-03
trapezoid n = 32: 1.9983933610   error -1.61e-03
```

Each halving of the step divides the error by 4.008, almost exactly 4. That number is itself information: an error that falls as h^2 is what the trapezoid rule should give (for a smooth function the trapezoid value exceeds the integral by $\frac{(b-a)h^2}{12}\overline{f''}$, with $\overline{f''}$ an average of the second derivative, here $-\sin x < 0$, hence the negative errors), so the code behaves as its theory says. An error ratio far from the expected one is the sign of a bug or of a function that is not smooth.

Richardson extrapolation. Knowing that the error falls as h^2 allows it to be removed. Write the result with step h as its limit plus the leading error, and do the same for $h/2$:

$$T(h) \overset{(1)}{=} I + ch^2 + O(h^4), \qquad T(h/2) \overset{(2)}{=} I + \tfrac{1}{4}ch^2 + O(h^4),$$
$$\frac{4T(h/2) - T(h)}{3} \overset{(3)}{=} \frac{4I + ch^2 - I - ch^2}{3} + O(h^4) = I + O(h^4). \tag{T0.8}$$

(1) The trapezoid error expanded in powers of the step. For a smooth function only even powers appear.
(2) The same with h replaced by $h/2$: the coefficient c is the same, so the h^2 term is a quarter as large.
(3) Four times the second line minus the first: the h^2 terms cancel and $3I$ remains; divide by 3.

The script prints the result:

```
Richardson (4 T16 - T8)/3 = 2.0000165910   error 1.66e-05
```

From the two cheap values at $n = 8$ and 16, whose errors are -0.02577 and -0.00643, the combination gives an error of 1.659×10^{-5}: about four hundred times smaller than at $n = 16$, for no new function evaluations. The four-point derivative of Section 10.5.1 is the same trick applied to (T0.4).

The book's convergence studies follow this pattern. The brute-force evidence of a four-parameter model in Section 5.19.1 is computed with n points per axis for seven values of n:

Listing T0.59: The same integral at seven resolutions, each timed (`code/ch05/27_grid_cost.py`)

```
50 ns = [8, 12, 16, 24, 32, 48, 64]
51 times, lnBs = [], []
52 for n in ns:
53     t0 = time.perf_counter()
54     lnZ = ln_Z_bump_bruteforce(y, n)
55     times.append(time.perf_counter() - t0)
56     lnBs.append(lnZ - lnZ0)
57     print(n, times[-1], lnBs[-1])
```

Line 50 lists the resolutions, and the loop of lines 52–57 computes the log Bayes factor at each, timing it with `time.perf_counter` (Section T0.12). The answer moves from 4.558 at $n = 12$ to 2.379 at $n = 24$ and 2.1 at $n = 64$, against the exact 2.096: only the finest grids have converged. Section 8.6.5 does the same with the number of simulated skies, watching the estimated bias and error bar settle inside the band $\pm\sigma/\sqrt{N}$. *The lesson: an answer is converged when doubling the resolution changes it by less than its other errors.*

T0.10.7 Validation: test the code where the answer is known

The checks so far estimate the size of an error. None of them catches a plain mistake in the code: a sign, a factor of two, a wrong index. Those are caught by one habit, used throughout the book: before a method is trusted on a problem with no known answer, it is run on a problem with one. The Monte Carlo variance of simulated $\widehat{C}_\ell$ is compared with the cosmic variance $2C_\ell^2/(2\ell+1)$ before simulations are used for masked skies (Section 7.12.3); the exact binomial interval found by root finding is compared with the closed formula in terms of the beta distribution (Section T0.9.1); the healpy round trip of Section T0.8.5 must return the spectrum that went in; the Fisher forecast is checked against an MCMC run on the same likelihood (Chapter 10).

A comparison can be written into the script as an `assert`: `assert condition` does nothing when the condition is true and stops the script with an error when it is false. The tiny-image study of Section 12.22.1 enumerates every one of the $2^9 = 512$ black-and-white 3×3 images and checks three facts about each:

Listing T0.60: Three identities checked on every one of 512 images (`code/ch12/c01_tiny.py`)

```
29 configs = []
30 for bits in itertools.product([0, 1], repeat=9):
31     m = np.array(bits, bool).reshape(3, 3)
32     b0, b1 = betti_plane(m)
33     chi = euler_cubical(m)
34     assert chi == b0 - b1                                  # the Euler relation, checked
       512 times
35     # scikit-image: 8-connected pieces, Euler number with 8-connectivity (connectivity=2)
36     assert measure.label(m, connectivity=2).max() == b0
37     assert measure.euler_number(np.pad(m, 1), connectivity=2) == chi
38     configs.append((b0, b1, chi, m.sum()))
```

Line 30 runs through all 512 patterns of nine bits, line 31 turns each into a 3×3 image, and lines 32–33 count its pieces and holes and compute its Euler characteristic with the book's own functions. Line 34 asserts the Euler relation $\chi = \beta_0 - \beta_1$; lines 36–37 assert that scikit-image's independent implementation finds the same number of pieces and the same χ. If any of the 1536 checks failed, the script would stop at that image. These are exact integer comparisons, so == is right here; between real numbers the comparison is made with a tolerance (Section T0.10.1).

The census of Section T0.1 counts how often the book writes such checks into its code: 19 `assert` statements in 15 files, and 13 calls of `np.isclose`, `np.allclose` and similar comparison functions in 9. That is a small number, and it understates the validation, because most of the book's checks are not asserts. They are a figure in which the simulation is drawn over the analytic curve (the black dashed `theory_line` of Section T0.7), or a pair of numbers saved by the script and quoted side by side in the text. A comparison in a figure or in the text is read by a person, who notices a discrepancy that no one thought to assert. *The lesson: a method is trusted where its answer is unknown only after it has reproduced an answer that is known.*

How this is used in real research: Verification first, then research

The same simulation code plays two roles. Run on an ideal full-sky Gaussian field, it *verifies*: its variance of $\widehat{C}_\ell$ must reproduce $2C_\ell^2/(2\ell+1)$ within its Monte Carlo error. Run with a mask, a beam and inhomogeneous noise, where no formula exists, it is the *research instrument*: the covariance it produces feeds a likelihood (Chapter 8). Research codes follow this order as a rule, and publish their validation tests next to their results.

T0.10.8 The decision rule: is a difference real?

Every check above ends with a comparison of two numbers: a simulation and a formula, two resolutions, two methods, two settings. They never agree exactly. When is a difference real, and when is it noise? The rule is the one every physicist uses for measurements. Collect the uncertainty of each number from every source, numerical and Monte Carlo, add the independent ones in quadrature, and divide the difference by the result:

$$z = \frac{|x_1 - x_2|}{\sqrt{\sigma_1^2 + \sigma_2^2}}, \qquad \sigma_i^2 = \sigma_{i,\mathrm{MC}}^2 + \sigma_{i,\mathrm{num}}^2. \tag{T0.9}$$

If the two numbers really estimate the same quantity and the errors are roughly Gaussian, z is the absolute value of a standard Gaussian variable: below 1 two times in three, below 2 nineteen times in twenty, above 3 about once in four hundred. So $z \lesssim 2$ means "consistent"; z between 2 and 3 means "look again", with more draws or another seed; $z \gtrsim 3$, in a single planned comparison, means that the difference is real, which may mean a bug. When many comparisons are made at once, a few values of z near 2 to 3 are expected by chance alone, exactly the look-elsewhere problem of Section 4.14.

Example: the Monte Carlo variance at the quadrupole. Section 7.12.3 estimates the variance of $\widehat{C}_\ell$ from 20,000 simulated skies and compares it with the cosmic variance $2C_\ell^2/(2\ell+1)$. At $\ell = 2$ the ratio of the two is 1.021, a 2% excess. Is the formula wrong, or the simulation, or neither? The analytic answer has no numerical error worth mentioning, and the simulated one has the Monte Carlo error derived there, 0.01483 in units of the ratio. So

$$z = \frac{1.021 - 1}{0.01483} \approx 1.4,$$

well inside the range of chance: neither is wrong. Had the excess been 10%, z would have been about 7, and the code would have been searched for the bug.

Example: a change seen with and without common random numbers. In Section T0.10.5, a single comparison of $\mathrm{P}(X > 2)$ at $\mu = 0$ and $\mu = 0.1$ with independent draws measures the true change 0.00597 against a noise of 0.00224: $z \approx 2.7$, a "look again". Over the 400 repeats, 60% of

such single comparisons fall short of 3σ, so one of them alone often would not establish that the probability changed at all. With common random numbers the noise is 0.00076 and $z \approx 7.9$, and 0% fall short: a difference beyond doubt, from the same number of draws.

Key idea: When a computed number can be trusted

A computed number is trusted to the extent that its error budget is known: rounding (ϵ times the amplification of cancellation or of the condition number), the method (step size, tolerance, resolution, checked by changing them), and Monte Carlo noise ($\sigma/\sqrt{N}$, checked by changing the seed). A difference is real when it is large compared with the combined budget, $z \gtrsim 3$; and a method is trusted on an unknown answer only after it has reproduced a known one.

Table 1 collects the checks in the form in which they are needed: a symptom, the check that diagnoses it, the fix, and a place in the book where it is done.

T0.11 When sixteen digits are not enough: mpmath

A double-precision number carries about sixteen significant digits and spans magnitudes from about 10^{-308} to 10^{308} (Section T0.10.1). For almost everything in the book that is ample. Occasionally it is not, and the failure has a recognisable shape: a probability so small that it underflows to zero, a sum whose terms cancel so completely that no digit of the answer survives, a factorial or a Gamma function so large that it overflows. The library **mpmath** computes with as many digits as we ask for, and the question is when to reach for it, how many digits to ask for, and what it costs.

The smallest example: the significance of a large excess. The one-sided p-value of a $z\sigma$ excess is the Gaussian tail $P(Z > z)$ (Chapter 4). There are three ways to compute it in SciPy: `1 - norm.cdf(z)`, `norm.sf(z)` (the survival function, which computes the tail directly) and `norm.logsf(z)`, its natural logarithm. The script `code/chT0/09_mpmath.py` compares all three with mpmath at fifty digits:

```
z =  5: 1-cdf = 2.8665e-07   sf = 2.8665e-07   log10 sf =   -6.5426   mpmath log10 =
    -6.5426
z =  8: 1-cdf = 6.6613e-16   sf = 6.2210e-16   log10 sf =  -15.2061   mpmath log10 =
    -15.2061
z =  9: 1-cdf = 0.0000e+00   sf = 1.1286e-19   log10 sf =  -18.9475   mpmath log10 =
    -18.9475
z = 20: 1-cdf = 0.0000e+00   sf = 2.7536e-89   log10 sf =  -88.5601   mpmath log10 =
    -88.5601
z = 37: 1-cdf = 0.0000e+00   sf = 5.7256e-300   log10 sf = -299.2422   mpmath log10 =
    -299.2422
z = 38: 1-cdf = 0.0000e+00   sf = 0.0000e+00   log10 sf = -315.5398   mpmath log10 =
    -315.5398
z = 39: 1-cdf = 0.0000e+00   sf = 0.0000e+00   log10 sf = -332.2714   mpmath log10 =
    -332.2714
z = 40: 1-cdf = 0.0000e+00   sf = 0.0000e+00   log10 sf = -349.4370   mpmath log10 =
    -349.4370
```

At 5σ, the discovery threshold of particle physics, all three are fine: even the naive `1 - cdf` is accurate to a relative 1.547×10^{-10}. At 8σ the naive form is already 7% wrong, and from 8.5σ on it returns exactly 0. The reason is the spacing of storable numbers just below 1, which is $\epsilon/2 \approx 1.1 \times 10^{-16}$: the cdf is stored as the nearest of these, so a tail much smaller than 10^{-16} cannot be represented as "1 minus something" at all. `sf` avoids the subtraction and stays accurate down to about 10^{-300}, but at 38.0σ the tail itself, about 10^{-316}, is below the smallest

symptom	check	fix	in the book
== between real numbers fails	print the difference; compare it with ϵ times the size	`np.isclose`, or a relative tolerance	Sections T0.9.1 and T2.9
digits lost in a small difference	count the leading digits that cancel, (T0.3)	rewrite the formula; subtract the large common part first	Sections T0.10.1 and 3.4.4
derivative changes with the step	scan h by factors of ten; look for a plateau	a step on the plateau; four-point rule	Sections 9.20.2 and 10.5.2
integral with a tiny claimed error	plot the integrand; tighten the tolerance; change the range	limits and `points` where the function lives	Section 6.12.4
`brentq` raises `ValueError`	evaluate f at both ends	a bracket with a sign change around the wanted root	Section 4.10.4
minimiser result looks plausible	`res.success`, `res.message`, gradient at `res.x`	more iterations; restart from `res.x`; a gradient-based method	Section T0.6.4
solution of $A\boldsymbol{x} = \boldsymbol{b}$ unstable	`np.linalg.cond(A)`; $\log_{10}\kappa$ digits are lost	`solve`, not `inv`; bin or regularise	Section 8.11.2
`cholesky` raises `LinAlgError`	eigenvalues of the matrix; rank against the number of simulations	more simulations; Hartlap factor	Sections 8.13 and 8.13.2
simulated number without an error bar	$\sigma/\sqrt{N}$ from the same draws; another seed	quote it; increase N	Sections 6.11.2 and 7.12.3
noisy difference of two simulations	scatter of the difference over seeds	common random numbers	Sections 8.5 and 12.20.2
answer depends on grid, N_{side}, $\ell_{\max}$	halve the step or double the size	finer resolution; Richardson extrapolation	Sections 5.19.1 and 8.6.5
a new method on a new problem	run it where the answer is known	find the bug before trusting the result	Sections 7.12.3 and 12.22.1
two numbers disagree	z of (T0.9) with all errors	$z \lesssim 2$: noise; $z \gtrsim 3$: real, or a bug	Section 7.12.3

Table 1: Is the number real? Thirteen symptoms, the check that diagnoses each, the usual fix, and a place in the book where it is done.

number with full precision, 2.225×10^{-308}, and the function returns 0. `logsf` never forms the tiny number: it computes $\ln \mathrm{P}(Z > z)$ directly, a modest negative number, and agrees with mpmath to every printed digit all the way to 40σ, where $\log_{10} \mathrm{P} = -349.44$ (Figure T0.27, left). The naive choice fails silently at 8σ; the good choice, `sf` or `logsf`, costs nothing. *The lesson: rewrite first; reach for more digits only when no rewriting will do, and then mostly to check.*

T0.11.1 Log space first

Most overflows and underflows in statistics come from products of many factors (likelihoods), from factorials and Gamma functions (binomial and Poisson probabilities, normalising constants), and from exponentials of large numbers (evidences, Boltzmann weights). All of them disappear when the computation is done with logarithms, which turn products into sums and keep every intermediate number of modest size. SciPy supplies the two tools that make this

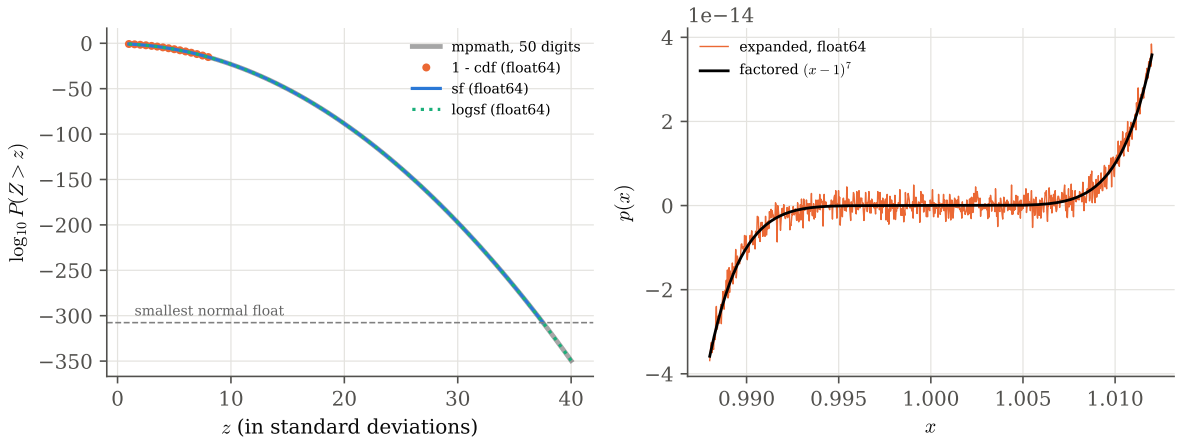


Figure T0.27: Where double precision breaks. Left: $\log_{10} \mathrm{P}(Z > z)$ from mpmath at fifty digits (thick grey) and from three float64 routes. `1 - cdf` (orange points) stops near 8.5σ, `sf` (blue) at 38σ, where the tail falls below the smallest normal float (dashed); `logsf` (dotted) follows mpmath to the end. Right: the polynomial $(x-1)^7$ near its root, evaluated from its expanded coefficients in float64 (orange) and in factored form (black). The expanded form returns rounding noise of size about 10^{-14} instead of a smooth curve.

painless.

`gammaln`. `scipy.special.gammaln(x)` returns $\ln\Gamma(x)$ without forming $\Gamma(x)$. Since $n! = \Gamma(n+1)$, $\ln 200!$ is `gammaln(201)`, which is 863.232, while 200! itself, about 10^{375}, overflows (Section T0.6.5). A ratio of Gamma functions is the exponential of a difference of `gammaln`s. The bias factor $c_4(N) = \sqrt{2/(N-1)}\,\Gamma(N/2)/\Gamma\big((N-1)/2\big)$ of the sample standard deviation (Section 3.4.1) is computed this way:

Listing T0.61: A ratio of Gamma functions as the exponential of a difference of logarithms (`code/ch03/04_sample_variance.py`)

For $N = 5$ the direct ratio would be harmless, but $\Gamma(N/2)$ overflows above $N \approx 344$, while the line as written works for any N: the logarithms stay of order $N \ln N$, and their difference, the logarithm of c_4, is close to zero.

`logsumexp`. An evidence, a mixture density or a normalisation is a sum of exponentials, $\ln \sum_i e^{a_i}$, and the a_i are log-likelihoods, often in the hundreds or thousands. For $a_1 = -1000$ and $a_2 = -1001$, `np.exp` underflows both to 0 and the logarithm of the sum is $-\infty$. Factoring out the largest term $m = \max_i a_i$ removes the problem:

$$\ln \sum_i e^{a_i} \overset{(1)}{=} \ln\Big(e^m \sum_i e^{a_i - m}\Big) \overset{(2)}{=} m + \ln \sum_i e^{a_i - m}. \tag{T0.10}$$

(1) $e^{a_i} = e^m e^{a_i - m}$ for every i; take the common factor out of the sum.
(2) The logarithm of a product is the sum of the logarithms, and $\ln e^m = m$.

Now every exponent $a_i - m$ is at most 0, the largest term is exactly 1, nothing overflows, and the terms that underflow are those too small to matter. For the example, $-1000 + \ln(1 + e^{-1}) = -999.687$. `scipy.special.logsumexp` does exactly this. The posterior probabilities of three competing models in Section 5.18.3 are their evidences normalised to sum to one, computed in log space throughout:

Listing T0.62: Posterior model probabilities from log-evidences (`code/ch05/25_how_many_bumps.py`)

```
66      lnZ = np.array([lnZ0, lnZ0 + lnB1, lnZ0 + lnB2])
67      lnp[key] = lnZ - logsumexp(lnZ)
```

Line 66 collects the three log-evidences, for no bump, one bump and two bumps. Line 67 subtracts $\ln \sum_k e^{\ln Z_k}$ from each, which is the logarithm of $Z_k / \sum_j Z_j$, the posterior probability of model k when the three are equally probable a priori. The evidences themselves are never formed. The tail probability above is the same idea in SciPy's distributions: `logpdf`, `logcdf` and `logsf` exist for every distribution, and a likelihood of many data points is built as a sum of `logpdf` values rather than a product of `pdf` values, which underflows to zero once there are a few hundred points.

T0.11.2 Cancellation that cannot be rewritten away

Sometimes no rewriting is available, because the formula arrives as given. A polynomial from a series expansion or a fit comes as its coefficients, not in factored form. The script takes the polynomial $(x-1)^7$ in its expanded form, $x^7 - 7x^6 + 21x^5 - 35x^4 + 35x^3 - 21x^2 + 7x - 1$, evaluates it near $x = 1$ by Horner's rule (multiply by x, add the next coefficient, repeat), and compares with the factored form:

```
x = 1.01: float64 expanded = 7.993606e-15   factored = 1.000000e-14
          mpmath 15 digits = 7.9936058e-15   30 digits = 1.0e-14   exact (x = 1.01 as
    decimal) = 1.0e-14
```

At $x = 1.01$ the exact value is 10^{-14}; float64 returns 7.994×10^{-15}, wrong by 20%. Figure T0.27 (right) shows the expanded form near the root: not a curve but a band of rounding noise. mpmath with fifteen digits reproduces the float64 answer digit for digit, because it is making the same rounding errors with the same number of digits; with thirty digits it returns 1.000×10^{-14}.

How many digits survive? The question is the one answered for a single subtraction by (T0.3), now for a sum of many terms. Suppose each term t_k of a sum $p = \sum_k t_k$ is computed with relative error at most u, the working precision: $u \approx \epsilon/2$ in float64, and $u \approx 10^{-d}$ when mpmath carries d decimal digits. Then

$$\frac{|\hat{p} - p|}{|p|} \overset{(1)}{=} \frac{|\sum_k t_k \delta_k|}{|\sum_k t_k|} \overset{(2)}{\le} u \, \frac{\sum_k |t_k|}{|\sum_k t_k|} \equiv u\,\kappa,$$

$$\text{digits trusted} \overset{(3)}{=} -\log_{10}(u\kappa) \approx d - \log_{10}\kappa. \tag{T0.11}$$

(1) Each computed term is $t_k(1+\delta_k)$, so the error of the sum is $\sum_k t_k \delta_k$.
(2) The triangle inequality and $|\delta_k| \le u$.
(3) A relative error 10^{-j} means j correct digits; with $u = 10^{-d}$, $-\log_{10} u = d$.

The **condition number** of the sum, $\kappa = \sum_k |t_k| \,/ |\sum_k t_k|$, is the ratio of the size of the terms to the size of the result. For two terms, a and $-b$, it is $(|a| + |b|)/|a - b|$, the amplification factor of (T0.3); for a linear system it is the $\kappa(A)$ of (T0.6). The rule reads the same each time: *working digits minus* $\log_{10}\kappa$. (Rounding inside Horner's rule adds a factor of order the number of terms, a fraction of a digit to one digit; the rule is a guide, not a guarantee.)

For the expanded polynomial the terms are the binomial coefficients times powers of x, so $\sum_k |t_k| = (x+1)^7$ by the binomial theorem, and

$$\kappa = \frac{(x+1)^7}{|x-1|^7} = \left(\frac{2.01}{0.01}\right)^7 \approx 1.3 \times 10^{16} \quad \text{at } x = 1.01,$$

which the script confirms, $\kappa = 1.33e + 16$ ($\log_{10}\kappa = 16.1$). With $d \approx 16$ working digits, (T0.11) predicts that *no* digit survives, and indeed none of the float64 answer is right. To keep ten digits we need $d \approx 16.1 + 10 \approx 26$; asking for thirty leaves a margin.

The practical test: d against $2d$. The condition number is rarely known in advance. The practical rule is to compute the same quantity twice, at d and at $2d$ digits, and keep the digits on which the two agree. Why does this work? The error of the $2d$ result is about 10^{-d} times smaller than that of the d result, so the difference between the two is, to a good approximation, the error of the d result itself:

Listing T0.63: The same sum at d and at $2d$ digits, and the number of digits on which they agree (`code/chT0/09_mpmath.py`)

```
77 xs_dec = "1.01"
78 rows = []
79 for dps in (10, 15, 20, 30):
80     mp.mp.dps = dps
81     a = expanded(mp.mpf(xs_dec))
82     mp.mp.dps = 2 * dps
83     b = expanded(mp.mpf(xs_dec))
84     mp.mp.dps = 4 * dps
85     agree = max(float(-mp.log10(abs((a - b) / b))), 0.0) if a != b else float(4 * dps)
86     rows.append((dps, a, b, agree))
87     say(f"dps = {dps:2d}: {mp.nstr(a, 12):>18}   dps = {2 * dps:2d}: {mp.nstr(b, 12):>18}
       agree to "
88         f"{agree:5.1f} digits")
```

Line 80 sets the working precision to `dps` decimal digits, and line 81 evaluates the polynomial at 1.01; line 82 doubles the precision and line 83 evaluates again. The input is given as the string `"1.01"`, so mpmath reads the decimal number exactly to the working precision; `mp.mpf(1.01)` would inherit the binary rounding of the float 1.01. Line 85 converts the relative difference into a number of agreeing digits. The result:

```
dps = 10:  2.18278728426e-10   dps = 20:  9.99999982452e-15   agree to   0.0 digits
dps = 15:   7.9936057773e-15   dps = 30:            1.0e-14   agree to   0.7 digits
dps = 20:  9.99999982452e-15   dps = 40:            1.0e-14   agree to   7.8 digits
dps = 30:            1.0e-14   dps = 60:            1.0e-14   agree to  16.4 digits
condition number of the expanded sum at x = 1.01: 1.325e+16  (log10 = 16.1)
```

At $d = 10$ and $d = 15$ the two results disagree completely: no digit of the d-digit answer can be trusted, as (T0.11) predicted for $d < 16$. At $d = 20$ they agree on 7.8 digits, and at $d = 30$ on 16.4; the rule, $d - 16$, predicted four and fourteen, so it is conservative, because it assumes that all rounding errors line up. *The lesson: the digits that survive a doubling of the precision are the digits to keep.*

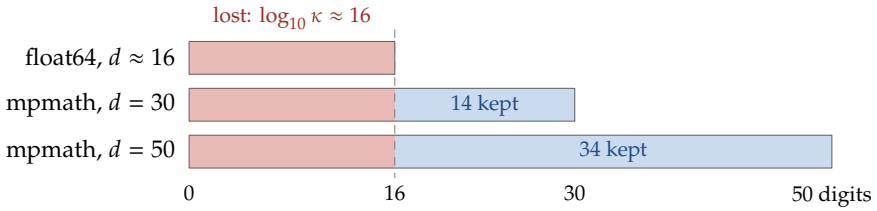


Figure T0.28: The digit budget of the expanded polynomial at $x = 1.01$. Each bar is the working precision; cancellation consumes the first $\log_{10}\kappa \approx 16$ digits (red), whatever the precision, and only the rest (blue) can be trusted. In float64 nothing is left. The kept digits are the guaranteed minimum of (T0.11); the d-against-$2d$ test usually finds a few more.

T0.11.3 What it costs, and when to use it

An mpmath number is a Python object, and every operation on it runs Python code rather than a single machine instruction; nor does mpmath work on whole arrays. The script times one exponential each way:

```
time per exp: numpy 13.0 ns, math 332 ns, mpmath (15 digits) 11330 ns, mpmath (50 digits)
    21097 ns
```

NumPy, working on a whole array at once, needs 13.0 ns per exponential; Python's `math.exp`, one number at a time, 332 ns; mpmath, at the same fifteen digits, 11330 ns, about 870 times slower than NumPy, and slower again at fifty digits. A Monte Carlo study of 10^8 draws that takes a second in NumPy would take more than a day in mpmath. So mpmath is a tool for checking, not for running.

It is the right tool in four situations, in order of how often they arise.

1. *As a reference.* A float64 routine whose accuracy is in doubt is compared, at a few points, with the same formula in mpmath at thirty or fifty digits: if the two agree to the digits that matter, the float64 routine is trusted everywhere it is fast. This is how Figure T0.27 judges SciPy's tail functions.
2. *Extreme tails*, beyond about 38σ, where even `sf` underflows. In practice `logsf` already covers this, and mpmath is needed only for the tail of a distribution that has no log-survival function.
3. *Cancellation that no rewriting removes*, as in the expanded polynomial, with d chosen by (T0.11) and checked by the d-against-$2d$ test.
4. *Huge or tiny intermediate numbers that log space cannot handle*, such as a sum of terms of alternating sign each larger than 10^{308} (logarithms cannot represent the signs that make such a sum cancel).

No script of the book needs mpmath for its results; the census of Section T0.1 finds it only here. Every case the book meets is handled by a rewrite or by log space, and those are what to try first.

T0.12 How long will it run? Counting operations

Before a long computation is started, two questions deserve an answer: how long will it take, and will it fit in memory? A script that finishes in a second for a test case may need a day, or a terabyte, for the real one, and it is far better to know that before pressing the key than after a night of waiting.

The smallest example: counting close pairs. The galaxy correlation function of Section 13.17 needs the number of pairs of points closer than some distance r. The obvious program compares every point with every other. The script `code/chT0/10_cost.py` times that brute-force count for N random points in a cube: for $N = 16\,000$ it takes 16.89 s. How long for a catalogue of a million galaxies? The naive extrapolation scales the time with the number of points: $10^6/16\,000 = 62.5$ times longer, about 18 minutes. That is wrong, because the work is not proportional to the number of points but to the number of *pairs*, $N(N-1)/2 \approx N^2/2$. Sixty-two and a half times more points means $62.5^2 \approx 3900$ times more pairs, and the honest estimate is about 18 hours; the script's fit to its measured times gives 22.6 hours. The good choice is to count first: how does the work grow with N? Its price is a minute of thought, and the same

thought shows the way out. A k-d tree (Section 13.17.1) groups the points into boxes and counts whole boxes of pairs at once, and it counts 256 000 points in 1.36 s. *The lesson: how the cost grows with the size matters far more than how fast the computer is.*

T0.12.1 Counting how the work grows

The cost of a computation, its run time or the number of arithmetic operations, depends on the size N of the problem: the number of data points, pixels, grid points, simulations. For large N it is usually close to a power law, $T(N) \approx cN^p$, possibly with a logarithm. The exponent p comes from the structure of the algorithm and can be found by counting; the constant c depends on the computer and the implementation and is found by timing. One writes $T(N) = O(N^p)$, read "of order N^p", to say that the cost grows no faster than a constant times N^p. Only the fastest-growing term is kept, and constant factors are dropped: $3N^2 + 100N$ is $O(N^2)$, because for large N the first term dominates, whatever the constants.

The exponent answers the most useful practical question directly: *if the problem gets ten times bigger, how much longer does it take?* For $O(N)$, ten times; for $O(N^2)$, a hundred times; for $O(N^3)$, a thousand times. The book's heavy computations fall into a handful of classes, each found by a count that fits in a line or two.

$O(N)$: anything done once per element. Adding two vectors, evaluating a density on a grid, summing a χ^2, drawing N random numbers: one operation, or a fixed few, per element. Every vectorised NumPy expression of Section T0.4.5 is of this kind.

$O(N \log N)$: halving, again and again. The fast Fourier transform (Section T0.5.3) splits a transform of length N into two transforms of length $N/2$, one for the even-numbered and one for the odd-numbered points, and combines them with about cN further operations. How many operations does that make in all? Apply the split repeatedly:

$$T(N) \overset{(1)}{=} 2\,T(N/2) + cN \overset{(2)}{=} 4\,T(N/4) + 2cN \overset{(3)}{=} 2^k\,T(N/2^k) + k\,cN \overset{(4)}{=} N\,T(1) + cN\log_2 N. \quad \text{(T0.12)}$$

(1) One split: two half-size transforms plus the combination.
(2) Split each half again: $2[2T(N/4) + cN/2] + cN$.
(3) After k splits there are 2^k transforms of length $N/2^k$, and each level of splitting has added cN.
(4) The splitting stops at length one, after $k = \log_2 N$ levels.

A direct sum would need N^2 operations (every output frequency from every input point). For $N = 10^6$ the difference between N^2 and $N\log_2 N \approx 2\times 10^7$ is a factor of fifty thousand. Sorting a list (`np.sort`, used for quantiles and empirical distribution functions) is $O(N \log N)$ for the same reason.

$O(N^2)$: all pairs. N points have $N(N-1)/2$ pairs, so anything that visits every pair, such as the brute-force pair count, the pairwise-distance matrix of the correlated field in Section 6.9.2, or the Vietoris–Rips complex of Chapter 12, is $O(N^2)$ in time, and $O(N^2)$ in memory if the pairs are stored. Trees and grids bring pair counting down by never visiting pairs that are obviously too far apart or that can be counted as a block.

$O(N^3)$: dense linear algebra. Solving $Ax = b$, the Cholesky factor, the determinant, the inverse and the eigenvalues of an $N \times N$ matrix all cost $O(N^3)$. For the Cholesky factor the

count is short. Column k of L is found from the entries to its left, after which every entry of the remaining $(N-k)\times(N-k)$ block below and to the right is updated once, with one multiplication and one addition; the block is symmetric, so only the entries on and below its diagonal are computed:

$$\text{operations} \overset{(1)}{=} \sum_{k=1}^{N} 2\cdot\frac{(N-k)(N-k+1)}{2} \overset{(2)}{=} \sum_{j=0}^{N-1}\left(j^2+j\right) \overset{(3)}{=} \frac{(N-1)N(2N-1)}{6}+\frac{(N-1)N}{2} \approx \frac{N^3}{3}. \tag{T0.13}$$

(1) An $m\times m$ symmetric block has $m(m+1)/2$ entries on and below its diagonal, with $m=N-k$; each costs two operations.
(2) Rename $j=N-k$, which runs from $N-1$ down to 0.
(3) The sum of the first n squares is $n(n+1)(2n+1)/6$ and of the first n integers $n(n+1)/2$, with $n=N-1$; for large N only the $N^3/3$ of the first term matters.

Every covariance matrix, every Gaussian likelihood in pixel space and every Fisher matrix of many parameters lives in this class.

$O(\ell_{\max}^3)$: spherical-harmonic transforms. A HEALPix map at resolution N_{side} has $N_{\text{pix}}=12N_{\text{side}}^2$ pixels on about $4N_{\text{side}}$ rings of constant latitude, and is analysed up to $\ell_{\max}\approx 3N_{\text{side}}$. On each ring an FFT separates the azimuthal modes m, which is cheap; then for every ring and every m the transform sums over all $\ell\geq m$:

$$\underbrace{4N_{\text{side}}}_{\text{rings}}\times\underbrace{\ell_{\max}}_{m}\times\underbrace{\ell_{\max}}_{\ell}\sim N_{\text{side}}^3\sim N_{\text{pix}}^{3/2}.$$

Doubling N_{side} makes `map2alm` and `alm2map` eight times slower [7].

Simulations and chains. A Monte Carlo study costs the number of simulations times the cost of one; an MCMC run costs the number of steps times the cost of one likelihood evaluation. These multiply. A chain of 10^5 steps with a Boltzmann code that takes a second per call takes more than a day; with the emulator of Section 9.20, which evaluates in microseconds, it takes seconds. That is the whole reason the emulator exists.

Memory. Time is not the only limit. A `float64` occupies 8 bytes, so an array of N numbers takes $8N$ bytes and an $N\times N$ matrix $8N^2$. A covariance matrix of 10^4 bins takes 800 MB; one of 4×10^4 bins, 12.8 GB, more than a laptop can spare. The pixel covariance of a map at $N_{\text{side}}=256$, with 786 432 pixels, would take $8\times786\,432^2\approx4.9\times10^{12}$ bytes, almost five terabytes. Memory can also be estimated before running: count the largest arrays, multiply by eight bytes per element, and remember that NumPy expressions create temporary arrays of the same size as their inputs.

T0.12.2 Estimating a run time: time small, fit, extrapolate

Counting gives the exponent; timing gives the constant. The recipe has four steps: time the computation at a few small sizes; fit a straight line to $\log T$ against $\log N$, whose slope is the measured exponent; extrapolate the line to the size wanted; and multiply by a safety factor. The script does the first two steps with two small functions:

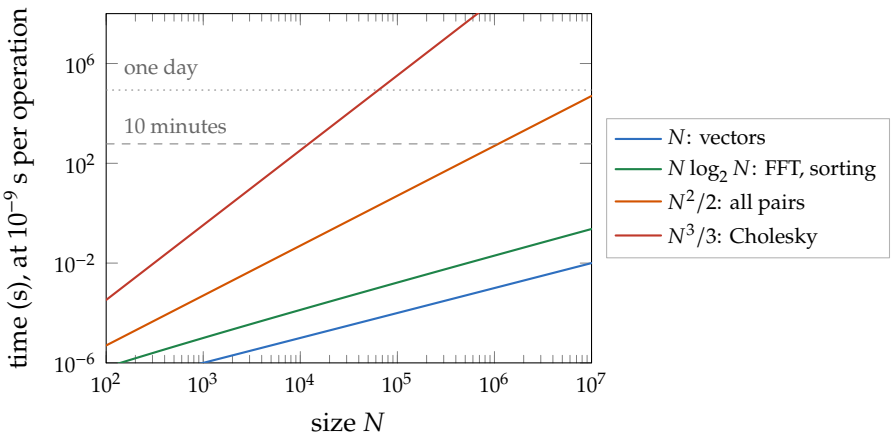


Figure T0.29: How the four commonest costs grow, at an illustrative 10^{-9} s per operation. At $N = 10^4$ all four are affordable; at $N = 10^6$ an N^2 computation takes about ten minutes and an N^3 one about ten years. The horizontal lines mark the ten-minute budget of a laptop script and one day.

Listing T0.64: Timing: the best of a few runs, and the slope on log-log axes (`code/chT0/10_cost.py`)

```
36 def best_time(f, repeats=3):
37     """Shortest of a few runs: the least disturbed by whatever else the machine is doing."""
38     t = []
39     for _ in range(repeats):
40         t0 = time.perf_counter()
41         f()
42         t.append(time.perf_counter() - t0)
43     return min(t)
44
45
46 def slope(sizes, times, last=4):
47     """Least-squares slope and intercept of log10(time) against log10(size), from the
       largest sizes."""
48     b, a = np.polyfit(np.log10(sizes[-last:]), np.log10(times[-last:]), 1)
49     return b, a
```

`time.perf_counter()` (lines 40 and 42) reads a clock with sub-microsecond resolution, and the difference of two readings is the time spent in between. `best_time` runs the computation three times and keeps the shortest time (line 43), the run least disturbed by whatever else the computer was doing. `slope` fits a straight line to the logarithms with `np.polyfit` (line 48); for $T = cN^p$, $\log T = \log c + p \log N$, so the slope is p. Only the largest sizes are used (`last=4`), because at small sizes fixed overheads, such as the call itself, distort the power law. The census finds such timing in 11% of the scripts, with 32 calls of `time.perf_counter` and 98 of the coarser `time.time`.

The script times four representative operations on a single processor core: the FFT, the Cholesky factor, pair counting by brute force and with a tree, and `healpy.map2alm`. For two of them it then measures a size beyond the fitted range, to see whether the prediction holds. Figure T0.30 shows the result, and the printed summary is:

```
FFT:       slope 1.35   (N log N alone would give 1.07)
Cholesky:  slope 2.45;  n = 6000: predicted 3.25 s, measured 3.21 s
pairs:     brute-force slope 2.04, k-d tree slope 1.69;  same counts: True
           brute force for 10^6 points, predicted 22.6 h;  tree at 256000 points 1.36 s
map2alm:   slope in N_side 3.06;  N_side = 512: predicted 2.01 s, measured 1.43 s
Cholesky of the 786432 x 786432 pixel covariance at N_side = 256: fitted slope 6 days,
    exponent 3 84 days, memory 4.9 TB
```

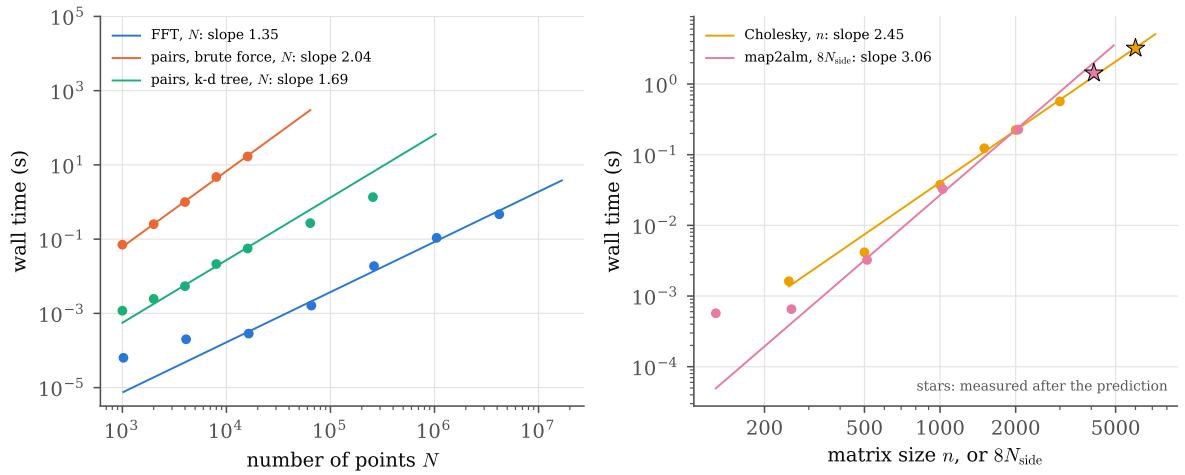


Figure T0.30: Measured run times on one core (points) and power-law fits (lines), with the fitted exponents in the legends. Left: FFT, brute-force pair counting and k-d-tree pair counting against the number of points. Right: the Cholesky factor of an $n \times n$ matrix, and `map2alm` against $8N_{\text{side}}$ (a rescaling that puts both on one axis). Stars: sizes beyond the fitted range, measured after the prediction was made.

What do the measured exponents say? Brute-force pair counting has slope 2.04, the counted 2. `map2alm` has slope 3.06 in N_{side}, the counted 3, and its prediction for $N_{\text{side}} = 512$, 2.01 s, compares with a measured 1.43 s. The Cholesky fit predicts 3.25 s for $n = 6000$ and the measurement is 3.21 s. These are successes. Two exponents differ from the count, and both differences teach something. The FFT's slope is 1.35, steeper than the 1.07 that $N \log N$ alone would give over the same range: at the largest sizes the arrays no longer fit in the processor's fast cache memory, and every operation waits longer for its data. The Cholesky slope, 2.45, is shallower than the counted 3: for small matrices the fixed overheads are a large part of the time, and the optimised linear-algebra library also runs more efficiently on larger blocks, so the time grows more slowly than the operation count until the matrix is large. A fitted slope describes the range it was fitted on, and nothing guarantees it beyond.

Example: recalling the likelihood that cannot be evaluated. Section 8.1.1 asked how long one evaluation of the exact Gaussian likelihood of a CMB map would take. It needs the Cholesky factor of the $N_{\text{pix}} \times N_{\text{pix}}$ pixel covariance, and its script times that factor up to $n = 6000$:

Listing T0.65: The cost study of Section 8.1.1: time at the largest size, exponent fixed by counting (`code/ch08/03_cost_scaling.py`)

```
# dense likelihood: Cholesky of a positive-definite N x N matrix
Ns = np.array([1000, 2000, 3000, 4500, 6000])
t_chol = []
for n in Ns:
    A = rng.standard_normal((n, n)); C = A @ A.T / n + np.eye(n)
    t_chol.append(best_time(lambda: np.linalg.cholesky(C)))
t_chol = np.array(t_chol)
a3 = t_chol[-1] / Ns[-1] ** 3          # t = a3 N^3, normalised at the largest N (closest to asymptotic)
```

Lines 30–35 time the Cholesky factor of random positive-definite matrices, with the same best-of-three function as above. Line 36 is the instructive one: rather than fitting a slope, it fixes the exponent at the counted 3 and uses only the largest, most nearly asymptotic, measurement to set the constant. From 0.80 s at $n = 6000$ it predicted about 21 days for the 786 432 pixels of $N_{\text{side}} = 256$, and 4.9 TB of memory. The script of this section, on a single core, measured 3.21 s

at $n = 6000$, about 4 times longer, because the earlier study let the linear-algebra library use up to four cores. With the counted exponent it predicts 84 days, about 4 times longer for the same reason: the two studies agree on the exponent and differ only in the constant. With its own fitted slope of 2.45, however, it would predict only 6 days, an underestimate by more than a factor of ten, because a slope fitted where overheads still matter is extrapolated far beyond the range in which it was measured. When the asymptotic exponent is known by counting, use it, and take the constant from the largest size measured. Either way the conclusion of Section 8.1.1 stands: the exact pixel likelihood is out of reach, while a spherical-harmonic transform of the same map takes a fraction of a second.

Example: a Monte Carlo before it is run. Suppose a study needs 1000 simulated skies at $N_{\mathrm{side}} = 256$, each analysed with three `map2alm` calls (temperature and two noise halves, say). One call took 0.227 s on one core, so the whole study needs about $3000 \times 0.227\,\mathrm{s} \approx 11$ minutes of transforms. That is within a laptop's budget on a few cores; at $N_{\mathrm{side}} = 512$ it would be eight times longer, and at $N_{\mathrm{side}} = 2048$, the resolution of the Planck maps, 512 times longer, which is why the book's simulations stay at $N_{\mathrm{side}} \le 512$ and say what the full-resolution version would cost. The grid evidence of Section 5.19.1 is the same kind of estimate in a more dramatic form: an n^4 grid that takes 9.43 s at $n = 64$ would take $1.784e+04$ years in six dimensions with a thousand points per axis.

The safety factor. Extrapolation is uncertain in both directions: the exponent may change beyond the fitted range (the FFT and the Cholesky factor above), and a large run shares the computer with other work. A factor of two to three on the predicted time is prudent, and a prediction that is still uncomfortable after it is a prediction that the computation should not be run as planned. It should be made cheaper (a tree, an FFT, a coarser grid, an emulator), split over cores (Section T0.8.3), or sent to a bigger machine.

Key idea: Estimate before you run

Count how the work and the memory grow with the size of the problem; time the computation at a few small sizes; extrapolate with the counted exponent, the time measured at the largest size, and a safety factor of two or three. Anything that the estimate puts beyond about ten minutes or a few gigabytes is redesigned or moved to a bigger computer before it is started.

T0.13 The whole toolkit on one page

Table 2 is the toolkit in the order a script uses it: each concept, the smallest piece of syntax that does it, the subsection that explains it, and a place in the book where it does real work. The syntax column is meant to be read as a reminder of something already understood, not learned from.

The census, once more. The table has thirty-six rows. The census of Section T0.1 explains why that is enough. The book's 345 scripts make 27297 calls to 759 distinct library functions, but the calls are very unevenly shared: 25 names make half of them, 83 make 80%, and 159 make 90%. The other 362 names are each used at most three times and together make only 2.1% of the calls; each of those can be looked up in the documentation on the one occasion it is met. The syntax is more concentrated still: loops, function definitions, f-strings, slices, dictionaries and keyword arguments appear in most scripts, and classes, `try`, `while` and decorators in a

concept	syntax	here	at work in
A script and plain Python			
run a script	`python3 code/ch02/02_binomial.py`	§T0.2	§2.2
docstring, helpers	`"""Question: ..."""; setup(), save_numbers(...)`	§T0.2	§0.6
f-string	`f"{x:.3f} {n:,} {p:.2e}"`	§T0.3.2	§1.1
list, tuple	`xs = []; xs.append(v); a, b = f(x)`	§T0.3.3	§3.9.3
dictionary	`d = {"mean": m}; for k, v in d.items():`	§T0.3.3	§3.1.4
loops, conditions	`for i, v in enumerate(xs):; if ...: elif ...: else:`	§T0.3.4	§1.1
comprehension	`[f(v) for v in xs if v > 0]`	§T0.3.4	§1.10.2
function	`def f(x, n=10): return a, b; f(x, n=20)`	§T0.3.5	§3.13.2
lambda	`lambda t: -loglike(t, data)`	§T0.3.5	§3.13.2
import	`import numpy as np; from scipy import stats`	§T0.3.6	every script
NumPy			
array, shape, type	`x = np.array([1.0, 2.0]); x.shape, x.dtype`	§T0.4.1	§1.1
grids of numbers	`np.linspace(a, b, n); np.logspace(p, q, n)`	§T0.4.1	§1.9
index, slice	`x[0], x[-1], x[a:b], X[:, 0], x[::-1]`	§T0.4.2	§3.1.4
mask	`x[(x > 0) & (x < 1)]; np.mean(x > 2)`	§T0.4.3	§1.1
broadcasting	`r[:, None] - r[None, :]; keepdims=True`	§T0.4.4	§6.9.2
reduction	`X.mean(axis=1), X.std(axis=0, ddof=1)`	§T0.4.6	§3.3.5
grids in d dimensions	`np.meshgrid(x, y, indexing="ij")`	§T0.4.7	§7.7
random numbers	`rng = rng_for("ch06", "name"); rng.standard_normal(n)`	§T0.5.1	§0.6
linear algebra	`A @ x; np.linalg.solve(C, d); cholesky; eigh`	§T0.5.2	§6.9.2
Fourier transform	`np.fft.rfft(x); np.fft.rfftfreq(n, dt)`	§T0.5.3	§9.23.2
SciPy, Matplotlib, files			
distributions	`stats.norm.pdf, .cdf, .sf, .ppf, .logsf, .rvs`	§T0.6.1	§4.9.1
integral	`val, err = integrate.quad(f, a, b)`	§T0.6.2	§1.10.2
root	`optimize.brentq(f, a, b)`	§T0.6.3	§4.10.4
minimum	`res = optimize.minimize(f, x0); res.x, res.success`	§T0.6.4	§3.11.4
log space	`special.gammaln(n + 1); special.logsumexp(a)`	§T0.11.1	§5.18.3
figure	`fig, ax = plt.subplots(); ax.plot(x, y); savefig(...)`	§T0.7.1	§2.2
files, cache	`np.savez(path, a=a); np.load(path); Path(...).exists()`	§T0.7.2	§7.11.1
Seen a few times			
data class	`@dataclass class Experiment: nside: int = 256`	§T0.8.1	§8.2.1
catching an error	`try: ... except np.linalg.LinAlgError: ...`	§T0.8.2	§11.14.3
several cores	`with Pool(4) as pool: out = pool.map(f, jobs)`	§T0.8.3	§12.23.1
compiled loop	`@njit` above a plain loop	§T0.8.4	§9.15.3
healpy, CAMB	`hp.alm2map`, `hp.anafast`; `camb.get_results`	§T0.8.5	§7.11.1
Is it real? How long?			
compare reals	`np.isclose(a, b); abs(a / b - 1) < 1e-10`	§T0.10.1	§T2.9
Monte Carlo error	`x.std(ddof=1) / np.sqrt(len(x))`	§T0.10.5	§0.6
condition number	`np.linalg.cond(A)`	§T0.10.4	§8.11.2
more digits	`mp.mp.dps = 30; mp.mpf("1.01")`	§T0.11	checks only
timing	`t0 = time.perf_counter()`	§T0.12.2	§5.19.1

Table 2: The toolkit on one page. "Here" points to the subsection that explains it; "at work in" to a place where the book uses it on a real problem.

few percent of them. A reader who knows the rows of Table 2, and the habits of Sections T0.10 to T0.12, can read every script in the book line by line, change it, and judge whether what it prints can be believed.

Chapter summary

Every number and figure in the book comes from a short script, and the scripts are written in a small, repetitive part of Python: a fixed anatomy, plain Python for the bookkeeping, NumPy for all arithmetic on arrays, a few SciPy tools, and a few Matplotlib calls. In the book's pipeline,

$$\text{physical model} \longrightarrow \text{field or dataset} \longrightarrow \text{summary } S \longrightarrow P(S \mid \theta) \longrightarrow \text{infer physics},$$

the code is the instrument that carries out every arrow, and the second half of the chapter is the instrument's calibration: what limits a computed number (rounding, the method, Monte Carlo noise), how to check each, when sixteen digits are not enough, and how to predict the cost of a computation before running it.

concept	operational meaning	used later in
the census	read every script's syntax tree without running it; count features, imports and calls; a few dozen names make most of the calls	every chapter
anatomy of a script	docstring (question, computes, writes), imports, settings, functions, computation and figure, numbers saved for the text	every chapter
vectorisation	one array expression instead of a Python loop; 137 times faster for a million elements	every chapter
broadcasting	arrays of different shapes combined by stretching axes of length one, compared from the right	Chapters 6, 7 and 13
seeded generator	`rng_for(chapter, name)`: the same numbers on every run; streams for independent sequences	Chapters 6, 8, 9 and 12
solve, do not invert	$A^{-1}\boldsymbol{b}$ by `solve` or Cholesky; `inv` only when the inverse itself is wanted	Chapters 8 and 10
the seven traps	float equality, integer arrays, views, ranges, shapes, mutable defaults, missing seeds	every chapter
machine epsilon	$\epsilon = 2^{-52} \approx 2.2 \times 10^{-16}$, the relative spacing of `float64` numbers; compare with tolerances	Chapters 4 and 10
cancellation	subtracting nearly equal numbers multiplies the relative error by $(\lvert a\rvert + \lvert b\rvert)/\lvert a - b\rvert$; rewrite the formula	Chapters 3 and 4
finite-difference step	central difference: best $h \sim \epsilon^{1/3}$, error $\sim \epsilon^{2/3}$; trust the plateau	Chapters 9, 10 and 12
condition number	$\kappa = \lambda_{\max}/\lambda_{\min}$; $\log_{10}\kappa$ digits lost; failed Cholesky means too few simulations or a singular matrix	Chapters 8 and 10
Monte Carlo error	$\sigma/\sqrt{N}$ quoted with every simulated number; common random numbers for differences	Chapters 6 to 8 and 12
convergence and validation	change the resolution until the answer stops moving; reproduce a known answer before trusting an unknown one	every chapter
decision rule	$z = \lvert\Delta\rvert / \sqrt{\sigma_1^2 + \sigma_2^2}$ with all errors; $z \gtrsim 3$ is real (or a bug)	Chapters 4 and 7
log space and mpmath	`logsf`, `gammaln`, `logsumexp` first; mpmath at d and $2d$ digits to check; trusted digits $\approx d - \log_{10}\kappa$	Chapters 4 and 5
cost $O(N^p)$	count the exponent, time small sizes for the constant, extrapolate with a safety factor; $8N^2$ bytes for an $N \times N$ matrix	Chapters 5, 8 and 13

Recap: Python for this book

- The book's code uses a small vocabulary: 25 names make half of all library calls and 159 make 90%. Learn those; look up the rest when they appear.
- A script reads like a derivation: question, imports, settings, computation, figure, numbers. Its numbers reach the text through `save_numbers`, never by hand.
- NumPy does the arithmetic on whole arrays: shapes, slices, masks, broadcasting, reductions along an axis. Loops are kept for what is truly sequential.
- Randomness always comes from a seeded generator; matrices are solved, not inverted; real numbers are compared with tolerances.
- Every computed number has an error budget: rounding, the method, Monte Carlo noise. Check each by changing what controls it, validate on a known answer, and call a difference real only when it is large compared with the budget.
- Rewrite and use log space before asking for more digits; when asking, compare d with $2d$.
- Count how the cost grows before running; time small sizes and extrapolate with a safety factor.

Reading the literature

- **The libraries.** The NumPy paper [10] explains the array model (shapes, strides, views, broadcasting, vectorised operations) on which everything here rests; the NumPy user guide, especially its pages on indexing, broadcasting and random generators, is the reference for exact behaviour. The SciPy paper [11] describes the algorithms behind `quad` (QUADPACK), `brentq`, `minimize` and the distributions, and the SciPy reference guide gives every option and every returned field. Matplotlib is described in [12]; its gallery is the quickest way to find how a particular kind of plot is drawn.
- **The physics libraries.** HEALPix is described in [7] and its Python interface healpy in [13]; CAMB in [8], with the CAMB Python documentation for the call pattern. numba is described in [14].
- **Floating point.** Goldberg's survey [9] is the classic account of how real numbers are stored and rounded, why cancellation happens, and how to rewrite formulas to avoid it. The mpmath documentation explains working precision (`mp.dps`) and the functions available at arbitrary precision.

Chapter 1

Probability: The Language of Uncertain Events

Where this chapter is going

What is the probability that a muon passing through a detector fires the trigger, or that a background fluctuation produces a bump at least as large as the one we saw? Given a CMB map, we might instead ask for the probability that Ω_m, the fraction of the cosmic energy density in matter, lies between 0.30 and 0.32. To make sense of each question we need to specify the event whose probability we want, the rules that probability obeys and the information on which it is conditioned.

The answers come in two steps. The first step works with *events*: yes-or-no statements about the outcome of an experiment, such as "the track was tagged as a kaon". The second step, in the next part, turns outcomes into numbers, called *random variables*, and brings in densities, marginals and moments.

One picture runs through both steps: probability is area inside a box that holds every possible outcome. Learning that something happened (conditioning) shrinks the box. Conditioning several times in a row draws a tree. Adding the branches of the tree that end in an observation gives the total probability of that observation. Reading the tree backwards, from the observation to the branch that produced it, is Bayes' theorem. Bayes' theorem links the two directions of probability of Chapter 0: prediction, $\mathrm{P}(E \mid H)$, which goes from a hypothesis H to the event E it produces, and inference, $\mathrm{P}(H \mid E)$, which goes from the event we saw back to the hypothesis that produced it.

Events are the language in which the last two steps of the book's pipeline are written. The pipeline (Chapter 0) runs: a theory with parameters θ; the data it produces; a statistic S that compresses the data; how S scatters from one dataset to the next, $P(S \mid \theta)$; and what the observed S says about θ. A *sampling distribution* is a list of probabilities of events ("S lands in this bin"), each computed assuming a hypothesis is true. That is the forward direction. *Inference* reads the same joint structure backwards. Every later chapter, from p-values (Chapter 4) to Metropolis–Hastings (Chapter 9) and Fisher forecasts (Chapter 10), is built from three rules about events: the axioms, the definition of conditional probability, and the law of total probability.

Questions about events.

1. What is the thing that has a probability? (An event, a subset of all possible outcomes.)
2. What must any assignment of probabilities satisfy, and what follows from that alone?

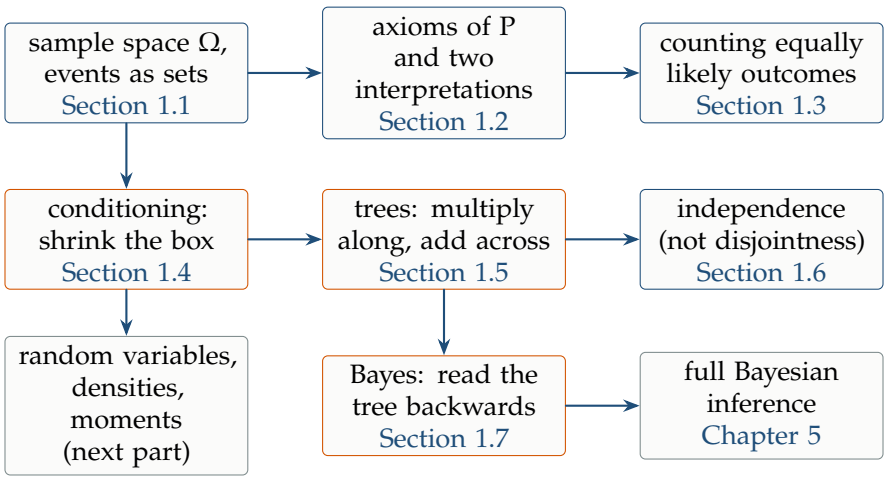


Figure 1.1: The route through the ideas. The top row builds the objects (events, the rules for their probabilities, and the special case in which probability is counting). The orange boxes are the core machine: conditioning shrinks the space of possibilities, trees chain conditionings together, and Bayes' theorem runs a tree in reverse. Grey boxes are where the ideas go next.

3. When every outcome is equally likely, how do we count?
4. How does learning "B happened" change the probability of A?
5. How do we compute the probability of an observation that can be produced in several ways?
6. When does learning B tell us nothing about A?
7. Having seen the observation, which of the ways produced it?

1.1 Sample spaces and events: probability needs a list of possibilities

A cosmic-ray muon telescope is three flat scintillator paddles stacked above each other. When a charged particle crosses a paddle, the paddle may or may not give a signal. For each passage we record which paddles fired: a pattern like $(1, 0, 1)$ means "top fired, middle did not, bottom fired". The trigger logic writes the passage to disk if *at least two* paddles fired.

Before we can ask "how often does the trigger fire?", we must say two things precisely: what could have happened, and which of those possibilities count as "the trigger fired". The first is the sample space, the second is an event. Once both are written down, every trigger condition, every selection cut and every hypothesis in the book is a *subset* of one fixed list. Combining conditions then becomes set algebra, which we can check by drawing pictures and by counting in code.

A box of possibilities. Picture a box. Every point in it is one complete description of how the experiment could turn out. Nature picks one point, the outcome ω. A yes-or-no statement about the result ("at least two paddles fired") is a region drawn in the box. The statement is true exactly when the chosen point lands inside the region. Probability will be the area of a region, and conditioning will mean throwing away part of the box.

Setting up any probability problem. Every problem in this book starts with the same four questions, asked in words before any symbol appears.

1. What is one run of the experiment? (One muon passage, one pair of die rolls, one collision.)
2. What is the complete list of its possible results, described finely enough that everything we will ask about can be read off?
3. Which statement about the result do we care about?
4. Which results make that statement true? That collection is the event.

Working definition 1.1 (sample space, outcome, event)

The **sample space** Ω is the set of all possible outcomes of an experiment [one entry per distinguishable result of one run, and exactly one entry is realised each time]. A point $\omega \in \Omega$ is called a **sample outcome**, **realization** or **element** [one complete description of how this run turned out]. A subset $A \subseteq \Omega$ is an **event** [a collection of outcomes, for example all patterns with at least two paddles fired]. We say "A happened" or "A occurred" when the realised outcome satisfies $\omega \in A$. *Operational test:* write down Ω as a list (or a formula), then write the event as $\{\omega \in \Omega : \text{ condition on } \omega\}$ and check that you can decide the condition for every ω.

The notation $\{\omega \in \Omega : \text{ condition}\}$ is **set-builder notation**. Read it as "the set of all ω in Ω such that the condition holds".

Example 1.1 (The telescope). One run is one muon passage. Each paddle either fires (1) or not (0), so

$$\Omega = \{000,\ 001,\ 010,\ 011,\ 100,\ 101,\ 110,\ 111\}, \qquad |\Omega| = 8,$$

where the digits are (top, middle, bottom) and $|\Omega|$ is the number of elements. The trigger event is

$$T = \{\omega : \text{ at least two digits of } \omega \text{ are } 1\} = \{011,\ 101,\ 110,\ 111\}.$$

The event "the top paddle fired" is $U = \{100, 101, 110, 111\}$. Both are subsets of the same Ω, which is what will let us combine them.

Example 1.2 (Three examples from Wasserman). *(a) Two coin tosses.* $\Omega = \{HH, HT, TH, TT\}$. "The first toss is heads" is $A = \{HH, HT\}$.

(b) A temperature measurement. The outcome is a real number, so we may take $\Omega = \mathbb{R} = (-\infty, \infty)$. Temperature is bounded below (by absolute zero), so $\mathbb{R}$ is larger than needed, but a sample space that is too big does no harm: the extra outcomes will simply get probability zero. "The reading is larger than 10 but at most 23" is the interval $A = (10, 23]$.

(c) Tossing a coin forever. An outcome is an infinite sequence,

$$\Omega = \big\{\omega = (\omega_1, \omega_2, \omega_3, \dots) : \ \omega_i \in \{H, T\}\big\}.$$

"The first head appears on the third toss" is

$$E = \big\{(\omega_1, \omega_2, \omega_3, \dots) : \ \omega_1 = T,\ \omega_2 = T,\ \omega_3 = H,\ \omega_i \in \{H, T\} \text{ for } i > 3\big\}.$$

E contains infinitely many outcomes, since tosses $4, 5, \dots$ are unrestricted. An event is a condition, not a single sequence.

Outcome versus event, and a sample space that is too coarse. Two cautions. First, an outcome and an event are different objects. The outcome 4 of a die is a point; the event $\{4\}$ is a set containing that point. Probabilities belong to events. Second, Ω must be fine enough to answer every question we will ask. Suppose we roll two dice but record only their *sum*, so

$\Omega = \{2, \dots, 12\}$. Then we can no longer ask whether the first die was even, because the sum does not say. The eleven sums are also not equally likely. The finer space of 36 ordered pairs (first die, second die) answers every question and makes each point equally likely. The sample space is fixed by the measurement we make [Kruschke §4.1].

1.1.1 Operations on events

Statements combine with "not", "or" and "and". Events combine in exactly the same way, through set operations.

Working definition 1.2 (complement, union, intersection, difference)

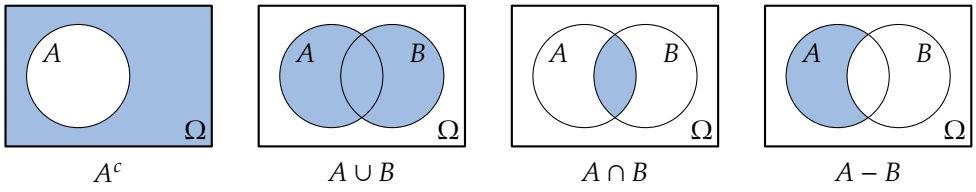
For events $A, B \subseteq \Omega$:

- the **complement** $A^c = \{\omega \in \Omega : \ \omega \notin A\}$ ["not A", every outcome outside A]. Note $\Omega^c = \varnothing$, the **empty set**.
- the **union** $A \cup B = \{\omega : \ \omega \in A$ or $\omega \in B$ or both$\}$ ["A or B", the inclusive or].
- the **intersection** $A \cap B = \{\omega : \ \omega \in A$ and $\omega \in B\}$ ["A and B", both conditions true for the same outcome]. It is also written AB or (A, B).
- the **set difference** $A - B = \{\omega : \ \omega \in A, \ \omega \notin B\} = A \cap B^c$ [in A but not in B].
- **inclusion** $A \subset B$ [every outcome in A is also in B, so "A happened" forces "B happened"]. Equivalently $B \supset A$.

For a sequence $A_1, A_2, \dots$ the countable union and intersection are

$$\bigcup_{i=1}^{\infty} A_i = \{\omega : \ \omega \in A_i \text{ for at least one } i\}, \qquad \bigcap_{i=1}^{\infty} A_i = \{\omega : \ \omega \in A_i \text{ for all } i\}.$$

The shorthand AB for $A \cap B$ will be used constantly, as in $\mathrm{P}(AB)$ for "the probability that A and B both happen". It is the same convention as the comma in $p(x, y)$ for a joint density in the next part.



A^c $A \cup B$ $A \cap B$ $A - B$

Figure 1.2: Venn diagrams: the rectangle is the sample space Ω, the discs are the events A and B, and the shaded region is the event named underneath. Shading the complement leaves a hole where A is. The union is everything touched by either disc, the intersection only the lens where they overlap, the difference the part of A outside B.

Four languages for the same thing.

sets	logic	detector / trigger	numpy mask
Ω	always true	every passage	`np.ones(N, bool)`
$\varnothing$	always false	impossible pattern	`np.zeros(N, bool)`
A^c	not A	veto on A	`~A`
$A \cup B$	A or B	OR of two trigger bits	`A | B`
$A \cap B$, AB	A and B	coincidence (AND)	`A & B`
$A - B$	A and not B	A with B as veto	`A & ~B`
$A \subset B$	A implies B	tighter cut	`np.all(B[A])`
$A \cap B = \varnothing$	cannot both hold	mutually exclusive patterns	`not np.any(A & B)`

Example 1.3 (Operations in the telescope). With T (trigger) and U (top fired) from Example 1.1:

$$
\begin{aligned}
T \cap U &= \{101, 110, 111\} && \text{(trigger fired and top fired),}\\
T \cup U &= \{011, 100, 101, 110, 111\} && \text{(trigger fired or top fired),}\\
T - U &= \{011\} && \text{(trigger fired without the top paddle),}\\
T^c &= \{000, 001, 010, 100\} && \text{(at most one paddle fired).}
\end{aligned}
$$

Each line is read off by going through the eight patterns and checking the condition. The event $T - U = \{011\}$ is useful: it is the only way the trigger can fire if the top paddle is dead.

1.1.2 The rules of set algebra

Set algebra looks like ordinary algebra, with $\cup$ playing $+$ and $\cap$ playing $\times$ [SeeingTheory p. 20]. So $A \cap (B \cup C)$ should expand like $a(b + c) = ab + ac$. The analogy is not perfect: $A \cup A^c = \Omega$, whereas $a + (-a) = 0$. So each rule has to be proved, and the method is always the same. Take one outcome and show that it is in the left side if and only if it is in the right side.

Key idea: De Morgan's laws and the distributive laws

For any events A, B, C,

$$
(A \cup B)^c = A^c \cap B^c, \qquad (A \cap B)^c = A^c \cup B^c, \tag{1.1}
$$

$$
A \cap (B \cup C) = (A \cap B) \cup (A \cap C), \qquad A \cup (B \cap C) = (A \cup B) \cap (A \cup C). \tag{1.2}
$$

In words: "not (A or B)" is "not A and not B". A veto on an OR trigger is the coincidence of the two vetoes.

Take any outcome $\omega \in \Omega$. Then

$$
\begin{aligned}
\omega \in (A \cup B)^c &\overset{(1)}{\iff} \omega \notin A \cup B\\
&\overset{(2)}{\iff} \text{it is false that } (\omega \in A \text{ or } \omega \in B)\\
&\overset{(3)}{\iff} \omega \notin A \text{ and } \omega \notin B\\
&\overset{(4)}{\iff} \omega \in A^c \text{ and } \omega \in B^c\\
&\overset{(5)}{\iff} \omega \in A^c \cap B^c.
\end{aligned}
$$

(1) Definition of the complement.
(2) Definition of the union.
(3) An "or" statement is false exactly when both parts are false.
(4) Definition of the complement, applied to A and to B.

(5) Definition of the intersection.

Every step is an "if and only if", so the two sets contain exactly the same outcomes. The second De Morgan law follows from the first. Apply the first law to the events A^c and B^c; since taking the complement twice gives back the original event, this reads $(A^c \cup B^c)^c = A \cap B$. Taking the complement of both sides gives $A^c \cup B^c = (A \cap B)^c$.

For the distributive law,

$$\begin{aligned} \omega \in A \cap (B \cup C) &\overset{(1)}{\iff} \omega \in A \text{ and } (\omega \in B \text{ or } \omega \in C) \\ &\overset{(2)}{\iff} (\omega \in A \text{ and } \omega \in B) \text{ or } (\omega \in A \text{ and } \omega \in C) \\ &\overset{(3)}{\iff} \omega \in (A \cap B) \cup (A \cap C). \end{aligned}$$

(1) Definitions of intersection and union.
(2) "p and (q or r)" is true exactly when "p and q" or "p and r" is true. Check the eight combinations of truth values if in doubt.
(3) Definitions of intersection and union again.

The other distributive law is proved the same way with "and" and "or" exchanged. All four identities extend to any collection $\{A_i : i \in I\}$, for example $\left(\bigcup_{i\in I} A_i\right)^c = \bigcap_{i\in I} A_i^c$: the chain above goes through word for word with "for at least one i" in place of "or" and "for all i" in place of "and".

1.1.3 Disjoint events, partitions and indicator functions

Working definition 1.3 (disjoint, partition, indicator)

Events $A_1, A_2, \ldots$ are **disjoint** or **mutually exclusive** if $A_i \cap A_j = \varnothing$ whenever $i \neq j$ [no single outcome belongs to two of them, so at most one can happen]. A **partition** of Ω is a sequence of disjoint events with $\bigcup_i A_i = \Omega$ [every outcome lies in exactly one piece]. The **indicator function** of A is

$$I_A(\omega) = I(\omega \in A) = \begin{cases} 1 & \omega \in A, \\ 0 & \omega \notin A. \end{cases}$$

For example $[0,1), [1,2), [2,3), \ldots$ are disjoint, and together with $[-1,0), [-2,-1), \ldots$ they partition $\mathbb{R}$. The histogram bins of any analysis are a partition of the range of the measured quantity. In the telescope, "exactly k paddles fired" for $k = 0,1,2,3$ is a partition of Ω into pieces of sizes $1,3,3,1$.

Indicators turn set operations into arithmetic, which is how a computer does set algebra. For any outcome,

$$\begin{aligned} &I_{A\cap B} = I_A\, I_B, \qquad I_{A^c} = 1 - I_A, \\ &I_{A\cup B} \overset{(1)}{=} 1 - I_{(A\cup B)^c} \overset{(2)}{=} 1 - I_{A^c\cap B^c} \overset{(3)}{=} 1 - (1-I_A)(1-I_B) \overset{(4)}{=} I_A + I_B - I_A I_B. \end{aligned} \tag{1.3}$$

(1) The complement rule $I_{C^c} = 1 - I_C$ with $C = A \cup B$.
(2) De Morgan (1.1).
(3) The product rule for intersections, then the complement rule for each factor.
(4) Expand the product.

The product rule holds because $I_A I_B$ is one only if both factors are one. Now sum (1.3) over all the outcomes of a finite Ω (or over all rows of a simulation). The sum of I_A over outcomes

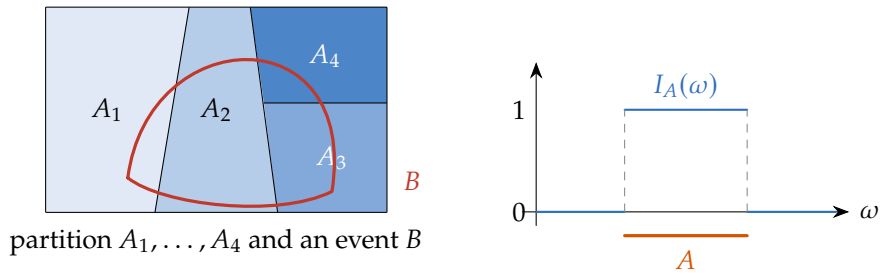


Figure 1.3: Left: a partition cuts Ω into non-overlapping pieces that cover it, so any event B (red outline) is cut into the disjoint pieces $B \cap A_i$. This picture is the law of total probability of Section 1.5. Right: the indicator of an event is a function on outcomes that equals one on the event and zero elsewhere.

counts the outcomes in A, which is $|A|$, the number of elements of A. So the sum gives the counting identity $|A \cup B| = |A| + |B| - |A \cap B|$. This is the seed of the inclusion–exclusion rule of Section 1.2.

1.1.4 Sequences of events and their limits

Some events are naturally the end of a sequence of events: "the detector is still alive after n hours" for $n = 1, 2, \dots$, or "the energy lies within $1/n$ of the peak". We need to say what the limit of such a sequence is.

Monotone sequences. A sequence $A_1, A_2, \dots$ is **monotone increasing** if $A_1 \subset A_2 \subset \cdots$ [each event is a looser condition than the one before], and then $\lim_{n\to\infty} A_n := \bigcup_{i=1}^{\infty} A_i$. It is **monotone decreasing** if $A_1 \supset A_2 \supset \cdots$ [each condition is tighter], and then $\lim_{n\to\infty} A_n := \bigcap_{i=1}^{\infty} A_i$. In either case we write $A_n \to A$.

Example 1.4 (Shrinking intervals: the endpoint matters). Let $\Omega = \mathbb{R}$ and $A_i = [0, 1/i)$. The sets shrink, $[0,1) \supset [0, \frac{1}{2}) \supset \cdots$. Their union is the biggest one, $\bigcup_i A_i = [0,1)$. For the intersection,

$$x \in \bigcap_{i=1}^{\infty} [0, 1/i) \overset{(1)}{\iff} 0 \le x < 1/i \text{ for every } i \ge 1$$
$$\overset{(2)}{\iff} 0 \le x \le 0$$
$$\overset{(3)}{\iff} x = 0,$$

(1) Definition of the intersection.
(2) $1/i$ can be made smaller than any positive number, so $x < 1/i$ for all i holds exactly when $x \le 0$.
(3) Combine $x \ge 0$ and $x \le 0$.

so $\bigcap_i A_i = \{0\}$. If instead $A_i = (0, 1/i)$, the same argument gives $\bigcup_i A_i = (0,1)$ and $\bigcap_i A_i = \varnothing$, because now $x = 0$ was never allowed. The limit of a decreasing sequence can be a single point or nothing at all, depending on one endpoint.

Try it: Infinitely often. Let $B_n = \bigcup_{i \ge n} A_i$ and $C_n = \bigcap_{i \ge n} A_i$. Show that $B_1 \supset B_2 \supset \cdots$ and $C_1 \subset C_2 \subset \cdots$; that $\omega \in \bigcap_n B_n$ exactly when ω lies in infinitely many of the A_i; and that $\omega \in \bigcup_n C_n$ exactly when ω lies in all but finitely many of them [Wasserman p. 14, Ex. 3]. *Answer*

sketch: $\omega \in \bigcap_n B_n$ means "for every n there is an $i \geq n$ with $\omega \in A_i$", which is "there is no last i with $\omega \in A_i$". If A_i is "a spark in the detector during hour i", $\bigcap_n B_n$ is "sparks never stop".

1.1.5 Events in code

In a simulation each row is one run of the experiment, so an event is simply a boolean array with one entry per row: `True` where the outcome lies in the event. The script below rolls two dice 100 000 times and builds A = "the first die is even" and B = "the sum is at least 9".

Listing 1.1: Events as boolean masks (`code/ch01/01_events_as_masks.py`)

```
20 rng = rng_for("ch01", "1a_events_masks")
21 N = 100_000
22
23 # --- one row per repetition of the experiment: the outcome omega = (d1, d2)
24 d1 = rng.integers(1, 7, size=N)
25 d2 = rng.integers(1, 7, size=N)
26
27 # --- events are boolean arrays: A[i] is True when outcome i lies in A
28 A = (d1 % 2 == 0)            # first die even
29 B = (d1 + d2 >= 9)           # sum at least 9
30
31 # --- set operations are logical operations on the masks
32 A_and_B = A & B              # intersection
33 A_or_B = A | B               # union
34 not_A = ~A                   # complement
35
36 # De Morgan: (A u B)^c == A^c n B^c must hold outcome by outcome
37 demorgan_violations = np.count_nonzero(~A_or_B != (~A & ~B))
38
39 # inclusion-exclusion on counts: |A u B| = |A| + |B| - |A n B|  (exact identity)
40 ie_gap = A_or_B.sum() - (A.sum() + B.sum() - A_and_B.sum())
41
42 # --- exact probabilities by counting the 36 equally likely outcomes
43 grid1, grid2 = np.meshgrid(np.arange(1, 7), np.arange(1, 7), indexing="ij")
44 Aex = (grid1 % 2 == 0)
45 Bex = (grid1 + grid2 >= 9)
46 pA, pB, pAB = Aex.mean(), Bex.mean(), (Aex & Bex).mean()
47 pAorB = (Aex | Bex).mean()
```

Lines 24–25 generate the outcomes, one pair (d_1, d_2) per row, where d_1 and d_2 are the two dice. Lines 28–29 turn the two conditions into masks. Lines 32–34 are the last column of the four-languages table above: `&`, `|` and `~` are intersection, union and complement. Line 37 checks De Morgan's law row by row and finds 0 violations, as it must, since (1.1) is an identity about every single outcome, not a statement that holds on average. Line 40 checks the counting identity $|A \cup B| = |A| + |B| - |A \cap B|$ and finds a gap of 0. Lines 43–47 compute exact probabilities by listing all 36 outcomes, which is the counting method of Section 1.3.

The simulated frequencies are 0.5004 for A (exact 0.5000), 0.2757 for B (exact $0.2778 = 10/36$), 0.1663 for $A \cap B$ (exact $0.1667 = 6/36$) and 0.6099 for $A \cup B$ (exact $0.6111 = 22/36$). The identities hold exactly, because they are true outcome by outcome. The frequencies only scatter around the exact probabilities, because they depend on which outcomes the simulation happened to draw. This difference, exact identity versus noisy estimate, returns throughout the book.

What this bought us. An event is a condition on outcomes, and in code it is a mask. Set identities such as De Morgan's hold outcome by outcome, so they hold exactly for simulated counts. Probabilities, which we define next, are what the frequencies of those masks approach.

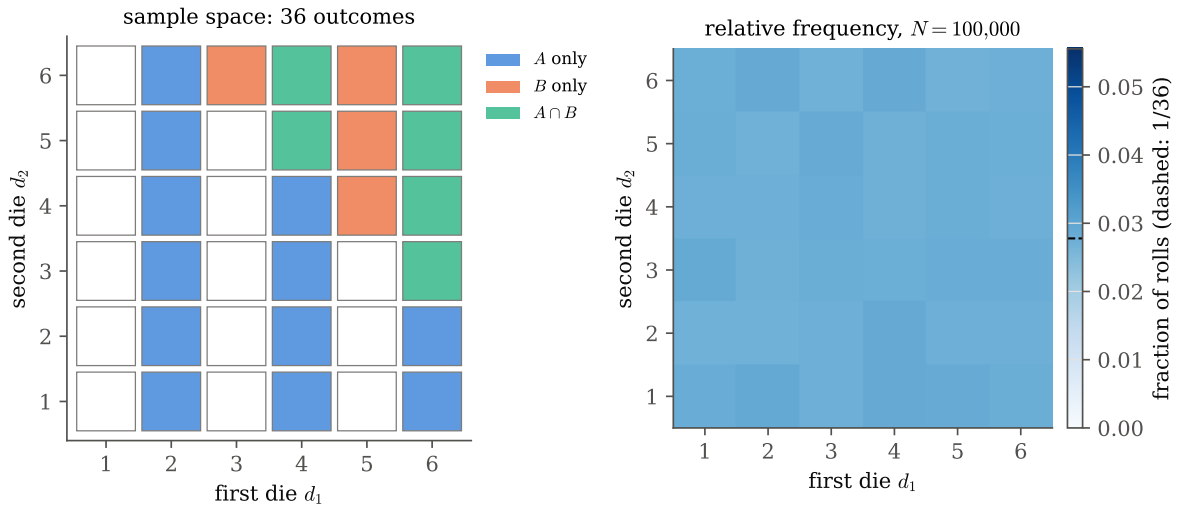


Figure 1.4: Left: the 36 outcomes of two dice as a grid, with the event A (first die even, blue), the event B (sum at least 9, orange) and their overlap (green). Counting cells gives the exact probabilities. Right: how often each cell was hit in the simulation. Every cell is close to the dashed value 1/36 on the colour bar, with small random differences.

1.2 The axioms of probability, and what they buy us

Return to the three-paddle muon telescope of Example 1.1. Analyst one has recorded a million muon passages and assigns to each of the eight patterns the fraction of passages that produced it. Analyst two has just installed a new middle paddle, has no data yet, and assigns numbers that express how strongly she expects each pattern, based on the data sheet of the photomultiplier. One analyst starts from measured frequencies and the other from expectations about a new detector. Nevertheless, both use the same rules to combine pattern probabilities into the probability that the trigger fires.

Whatever the numbers mean, any assignment of probabilities must obey three rules, the axioms. Every rule we compute with follows from these three: the complement rule, the inclusion–exclusion rule, the union bound that controls the look-elsewhere effect, and the continuity that lets us take limits. Only after the rules are in place do we ask what the numbers mean, and where each meaning is used.

Probability is a unit of sand spread over the box. Take the box of possibilities from Section 1.1 and pour exactly one kilogram of sand over it, as unevenly as you like. The probability of an event is the mass of sand lying inside its region. Mass is never negative. The whole box holds all of it. If two regions do not overlap, the sand inside their union is the sand in one plus the sand in the other. You would accept every one of these statements about sand without proof. The proofs below only check that three of them imply all the others.

What an assignment of probabilities must do. Stated in words, the demands are these.

1. To every event we can ask about, attach a number.
2. The number for "something happens" (Ω) must be the largest possible, and we fix the scale by calling it one.
3. No event may get a negative number.

4. Numbers for events that cannot happen together must add, also for an infinite list of such events.
5. Everything else (complements, overlaps, limits) must then follow, so that we never need another rule.

Working definition 1.4 (probability distribution)

A **probability distribution** or **probability measure** P is a rule that assigns a real number $\mathrm{P}(A)$ to each event A [each subset of Ω we are allowed to ask about, see Section 1.2.4] such that

Axiom 1: $\mathrm{P}(A) \geq 0$ for every event A.

Axiom 2: $\mathrm{P}(\Omega) = 1$.

Axiom 3: if $A_1, A_2, \ldots$ are disjoint [no outcome lies in two of them], then

$$\mathrm{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \mathrm{P}(A_i).$$

Axiom 3 is called **countable additivity** [additivity for an infinite sequence of disjoint events, one for each $i = 1, 2, 3, \ldots$]. *Operational test:* a table of numbers $\mathrm{P}(\{\omega\})$ on a finite or countable Ω is a valid distribution exactly when every entry is nonnegative and the entries sum to one.

These three lines are Kolmogorov's axioms (1933). They say nothing about *how* to find the numbers. That is the job of physics, of data and of judgement. They only say which lists of numbers are internally consistent.

1.2.1 Consequences of the axioms

Five properties follow at once [Wasserman p. 6, (1.1)]:

$$\begin{aligned} &\mathrm{P}(\varnothing) = 0, \\ &A \subset B \implies \mathrm{P}(A) \leq \mathrm{P}(B), \\ &0 \leq \mathrm{P}(A) \leq 1, \\ &\mathrm{P}(A^c) = 1 - \mathrm{P}(A), \\ &A \cap B = \varnothing \implies \mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B). \end{aligned} \tag{1.4}$$

Each takes a few lines to prove, and each proof is one idea that we will use again, so we prove them all.

The empty event has probability zero. Take the sequence $A_1 = \Omega$, $A_2 = A_3 = \cdots = \varnothing$. These are disjoint (the empty set meets nothing) and their union is Ω. Then

$$1 \overset{(1)}{=} \mathrm{P}(\Omega) \overset{(2)}{=} \mathrm{P}\big(\Omega \cup \varnothing \cup \varnothing \cup \cdots\big) \overset{(3)}{=} \mathrm{P}(\Omega) + \sum_{i=2}^{\infty} \mathrm{P}(\varnothing) \overset{(4)}{=} 1 + \sum_{i=2}^{\infty} \mathrm{P}(\varnothing).$$

(1) Axiom 2.
(2) Adding empty sets to a union changes nothing.
(3) Axiom 3 for the disjoint sequence $\Omega, \varnothing, \varnothing, \ldots$.
(4) Axiom 2 again.

So $\sum_{i\geq 2} \mathrm{P}(\varnothing) = 0$. This is a sum of infinitely many copies of the same number, which is nonnegative by Axiom 1. It is zero only if that number is zero: $\mathrm{P}(\varnothing) = 0$.

Finite additivity. For disjoint $A_1, \ldots, A_n$, pad the list with empty sets, $A_{n+1} = A_{n+2} = \cdots = \varnothing$. Axiom 3 then gives $\mathrm{P}(A_1 \cup \cdots \cup A_n) = \sum_{i=1}^{n} \mathrm{P}(A_i) + \sum_{i>n} 0$, which is the last line of (1.4) for $n = 2$.

Complement rule. A and A^c are disjoint and $A \cup A^c = \Omega$, so

$$1 \overset{(1)}{=} \mathrm{P}(\Omega) \overset{(2)}{=} \mathrm{P}(A \cup A^c) \overset{(3)}{=} \mathrm{P}(A) + \mathrm{P}(A^c). \tag{1.5}$$

(1) Axiom 2.
(2) Every outcome is either in A or not.
(3) Finite additivity.

This is the most used identity in the book: "at least one" is always computed as one minus "none".

Monotonicity, and the difference rule. If $A \subset B$, cut B into the part inside A and the part outside, $B = A \cup (B - A)$, which are disjoint. Then

$$\mathrm{P}(B) \overset{(1)}{=} \mathrm{P}(A) + \mathrm{P}(B - A) \overset{(2)}{\geq} \mathrm{P}(A). \tag{1.6}$$

(1) Finite additivity for the disjoint pieces A and $B - A$.
(2) Axiom 1 for the event $B - A$.

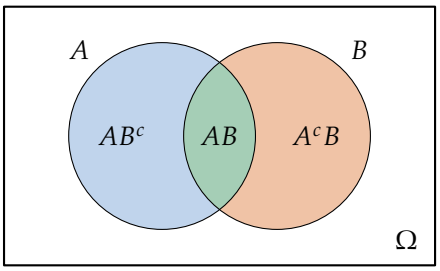
The first equality also says $\mathrm{P}(B - A) = \mathrm{P}(B) - \mathrm{P}(A)$ when $A \subset B$. In detector language: a tighter cut A inside a looser cut B can only lose outcomes, never gain them, so its probability cannot be larger. Finally $0 \leq \mathrm{P}(A)$ is Axiom 1, and $\mathrm{P}(A) \leq \mathrm{P}(\Omega) = 1$ is monotonicity with $B = \Omega$.

Key idea: Inclusion–exclusion

For any events A and B,

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(AB). \tag{1.7}$$

Adding $\mathrm{P}(A)$ and $\mathrm{P}(B)$ counts the overlap twice, so we subtract it once.



$\mathrm{P}(A) = \mathrm{P}(AB^c) + \mathrm{P}(AB)$

$\mathrm{P}(B) = \mathrm{P}(AB) + \mathrm{P}(A^cB)$

sum: the green lens twice

$\mathrm{P}(A \cup B)$: each piece once

Figure 1.5: The union of two events is cut into three disjoint pieces: the part of A outside B (blue), the overlap (green) and the part of B outside A (orange). $\mathrm{P}(A)$ and $\mathrm{P}(B)$ between them contain the green overlap twice, and the union contains it once.

To prove (1.7), use the three disjoint pieces of Figure 1.5: $A \cup B = (AB^c) \cup (AB) \cup (A^cB)$. Then

$$\begin{aligned} \mathrm{P}(A \cup B) &\overset{(1)}{=} \mathrm{P}(AB^c) + \mathrm{P}(AB) + \mathrm{P}(A^cB) \\ &\overset{(2)}{=} \mathrm{P}(AB^c) + \mathrm{P}(AB) + \mathrm{P}(A^cB) + \mathrm{P}(AB) - \mathrm{P}(AB) \\ &\overset{(3)}{=} \mathrm{P}\big((AB^c) \cup (AB)\big) + \mathrm{P}\big((A^cB) \cup (AB)\big) - \mathrm{P}(AB) \\ &\overset{(4)}{=} \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(AB). \end{aligned}$$

(1) Finite additivity for three disjoint pieces.
(2) Add and subtract the same number.
(3) Finite additivity backwards, for the disjoint pairs AB^c, AB and A^cB, AB.
(4) An outcome of A is either outside B or inside it, so $(AB^c) \cup (AB) = A$. Likewise $(A^cB) \cup (AB) = B$.

Example 1.5 (Two coin tosses). Let H_1 be "heads on toss 1" and H_2 "heads on toss 2". With the four outcomes HH, HT, TH, TT equally likely, $\mathrm{P}(H_1) = \mathrm{P}(\{HH, HT\}) = \frac{1}{2}$, $\mathrm{P}(H_2) = \mathrm{P}(\{HH, TH\}) = \frac{1}{2}$ and $\mathrm{P}(H_1H_2) = \mathrm{P}(\{HH\}) = \frac{1}{4}$. So

$$\mathrm{P}(H_1 \cup H_2) \overset{(1)}{=} \mathrm{P}(H_1) + \mathrm{P}(H_2) - \mathrm{P}(H_1H_2) = \tfrac{1}{2} + \tfrac{1}{2} - \tfrac{1}{4} = \tfrac{3}{4}.$$

(1) Inclusion–exclusion (1.7).

Check by the complement rule: "at least one head" fails only for TT, so the answer is $1 - \frac{1}{4} = \frac{3}{4}$. The two routes must agree, and they do.

Example 1.6 (A probability table for the telescope). On a finite sample space, finite additivity over single outcomes gives

$$\mathrm{P}(A) = \sum_{\omega \in A} \mathrm{P}(\{\omega\}), \tag{1.8}$$

so a distribution is just a table of nonnegative numbers that sum to one. Suppose the telescope patterns have the probabilities

pattern	000	001, 010, 100	011, 101, 110	111
probability (each)	0.001	0.009	0.081	0.729

They sum to $0.001 + 3(0.009) + 3(0.081) + 0.729 = 1$, so the table is legitimate. (Where such numbers come from is the subject of Section 1.6: they are what one gets if each paddle fires with probability 0.9 on its own.) For the trigger T and the event U (top fired) of Example 1.3,

$$\begin{aligned} \mathrm{P}(T) &= 3(0.081) + 0.729 = 0.972, \qquad \mathrm{P}(U) = 0.009 + 2(0.081) + 0.729 = 0.9, \\ \mathrm{P}(TU) &= \mathrm{P}(\{101, 110, 111\}) = 2(0.081) + 0.729 = 0.891, \\ \mathrm{P}(T \cup U) &\overset{(1)}{=} 0.972 + 0.9 - 0.891 = 0.981. \end{aligned}$$

(1) Inclusion–exclusion. Directly, $T \cup U = \{011, 100, 101, 110, 111\}$ has probability $0.081 + 0.009 + 0.081 + 0.081 + 0.729 = 0.981$, the same.

The trigger misses a muon with probability $\mathrm{P}(T^c) = 1 - 0.972 = 0.028$, by the complement rule.

Try it: Three events. Show that $\mathrm{P}(A\cup B\cup C) = \mathrm{P}(A)+\mathrm{P}(B)+\mathrm{P}(C)-\mathrm{P}(AB)-\mathrm{P}(AC)-\mathrm{P}(BC)+\mathrm{P}(ABC)$. *Hint:* apply (1.7) to $A\cup B$ and C, then use the distributive law $(A\cup B)\cap C = (AC)\cup(BC)$ and (1.7) once more, noting $(AC)(BC) = ABC$. In the telescope, the three events "paddle j fired" give $3(0.9) - 3(0.81) + 0.729 = 0.999$, which is $1 - \mathrm{P}(\{000\})$.

1.2.2 Two inequalities: Bonferroni and Boole

Often we cannot compute an overlap $\mathrm{P}(AB)$, but we can still bound it from below. Rearrange (1.7) and use $\mathrm{P}(A\cup B)\le 1$:

$$\mathrm{P}(AB) = \mathrm{P}(A)+\mathrm{P}(B)-\mathrm{P}(A\cup B) \;\ge\; \mathrm{P}(A)+\mathrm{P}(B)-1. \tag{1.9}$$

This is the simplest **Bonferroni inequality** [CasellaBerger (1.2.9)]. *Example.* The tracker and the calorimeter are each fully operational in 95% of the luminosity blocks. Then both are operational in at least $95\% + 95\% - 100\% = 90\%$ of them, however their failures are correlated. When $\mathrm{P}(A)+\mathrm{P}(B) < 1$ the bound is negative: true, but useless.

Key idea: Boole's inequality, the union bound

For any events $A_1, A_2, \ldots$ (disjoint or not),

$$\mathrm{P}\Big(\bigcup_{n=1}^{\infty} A_n\Big) \le \sum_{n=1}^{\infty} \mathrm{P}(A_n). \tag{1.10}$$

The probability that at least one thing goes wrong is at most the sum of the probabilities of each thing going wrong.

Make the events disjoint by keeping only the part of each A_n that is new: $B_1 = A_1$ and $B_n = A_n - \bigcup_{i<n} A_i$ [Wasserman p. 14, Ex. 7]. An outcome in $\bigcup_n A_n$ has a *first* index n with $\omega\in A_n$, and it lies in B_n for that n only. So the B_n are disjoint, $\bigcup_n B_n = \bigcup_n A_n$, and $B_n\subset A_n$. Then

$$\mathrm{P}\Big(\bigcup_n A_n\Big) \overset{(1)}{=} \mathrm{P}\Big(\bigcup_n B_n\Big) \overset{(2)}{=} \sum_n \mathrm{P}(B_n) \overset{(3)}{\le} \sum_n \mathrm{P}(A_n).$$

(1) The two unions contain the same outcomes.
(2) Axiom 3 for the disjoint B_n.
(3) Monotonicity (1.6), term by term, since $B_n\subset A_n$.

In research: the look-elsewhere effect in one line. A bump hunt scans m mass bins. If background alone makes a given bin fluctuate up to the observed size with probability p_{local}, the probability that *some* bin does so is at most $m\,p_{\text{local}}$ by (1.10). This is the Bonferroni correction for multiple tests (Chapter 4). It is a bound, and it is loose when neighbouring bins are correlated. In real searches the actual "global" probability is obtained by simulating many background-only experiments and counting how often the largest fluctuation anywhere exceeds the observed one [15]. The union bound is what tells us in advance roughly how bad the effect can be.

Example 1.7 (Events of probability one). If $\mathrm{P}(A_i) = 1$ for every i, then $\mathrm{P}\big(\bigcap_i A_i\big) = 1$ [Wasserman p. 14, Ex. 8]:

$$\mathrm{P}\Big(\big(\textstyle\bigcap_i A_i\big)^c\Big) \overset{(1)}{=} \mathrm{P}\Big(\textstyle\bigcup_i A_i^c\Big) \overset{(2)}{\le} \sum_i \mathrm{P}(A_i^c) \overset{(3)}{=} \sum_i 0 = 0.$$

(1) De Morgan's law (1.1) for a countable family.

(2) Boole's inequality (1.10).
(3) Complement rule: $\mathrm{P}(A_i^c) = 1 - 1 = 0$.

A probability is nonnegative, so the complement has probability exactly zero and the intersection has probability one. Countably many "almost sure" statements hold almost surely all at once.

1.2.3 Continuity: probabilities of limits are limits of probabilities

Continuity of probability. If $A_n \to A$ (a monotone sequence, Section 1.1), then $\mathrm{P}(A_n) \to \mathrm{P}(A)$ as $n \to \infty$ [Wasserman p. 7, Thm 1.8].

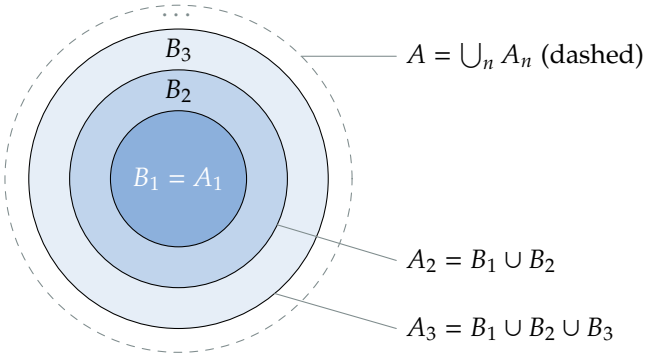


Figure 1.6: An increasing sequence of events drawn as nested discs. The rings B_n (the part of A_n that is new at step n) do not overlap, and the first n rings together make up A_n. The probability of the limit is the total mass of all rings, which is the limit of the partial sums.

Suppose $A_1 \subset A_2 \subset \cdots$ with $A = \bigcup_i A_i$. Define the rings of Figure 1.6: $B_1 = A_1$ and $B_n = \{\omega : \omega \in A_n,\ \omega \notin A_{n-1}\}$ for $n \geq 2$. Because the sets are nested, "not in A_{n-1}" already means "not in any earlier A_i", so this is the construction used for Boole's inequality. The B_n are disjoint, $A_n = \bigcup_{i=1}^{n} B_i$ and $\bigcup_{i=1}^{\infty} B_i = A$. Then

$$\lim_{n\to\infty} \mathrm{P}(A_n) \overset{(1)}{=} \lim_{n\to\infty} \mathrm{P}\Big(\bigcup_{i=1}^{n} B_i\Big) \overset{(2)}{=} \lim_{n\to\infty} \sum_{i=1}^{n} \mathrm{P}(B_i) \overset{(3)}{=} \sum_{i=1}^{\infty} \mathrm{P}(B_i) \overset{(4)}{=} \mathrm{P}\Big(\bigcup_{i=1}^{\infty} B_i\Big) \overset{(5)}{=} \mathrm{P}(A).$$

(1) A_n is the union of the first n rings.
(2) Finite additivity.
(3) An infinite sum is by definition the limit of its partial sums. It converges because the partial sums increase and never exceed one.
(4) Axiom 3, which is where *countable* additivity is essential.
(5) The rings together make up A.

For a decreasing sequence $A_1 \supset A_2 \supset \cdots$ with $A = \bigcap_i A_i$, the complements increase, $A_1^c \subset A_2^c \subset \cdots$, and by De Morgan $\bigcup_i A_i^c = \big(\bigcap_i A_i\big)^c = A^c$. The increasing case gives $\mathrm{P}(A_n^c) \to \mathrm{P}(A^c)$, so $\mathrm{P}(A_n) = 1 - \mathrm{P}(A_n^c) \to 1 - \mathrm{P}(A^c) = \mathrm{P}(A)$ [Wasserman p. 13, Ex. 1].

Example 1.8 (A detector that is alive forever). Let A_n be "the readout board has not failed during the first n hours". The events decrease, $A_1 \supset A_2 \supset \cdots$, and their intersection A is "the board never fails". Suppose a reliability model gives $\mathrm{P}(A_n) = e^{-n/\tau}$ with a mean lifetime of $\tau = 10^4$ hours. Continuity gives

$$\mathrm{P}(\text{never fails}) \overset{(1)}{=} \lim_{n\to\infty} \mathrm{P}(A_n) = \lim_{n\to\infty} e^{-n/\tau} = 0.$$

(1) Continuity for the decreasing sequence $A_n \to A$.

We could not have computed $\mathrm{P}(A)$ directly, since A is defined by an infinite number of conditions. Continuity turns it into an ordinary limit.

Example 1.9 (There is no uniform distribution on the integers). Can we pick a whole number "completely at random", so that every $n \in \{0, 1, 2, \dots\}$ has the same probability c [Wasserman p. 14, Ex. 6]? The single-number events $\{0\}, \{1\}, \dots$ are disjoint with union Ω, so

$$1 \overset{(1)}{=} \mathrm{P}(\Omega) \overset{(2)}{=} \sum_{n=0}^{\infty} \mathrm{P}(\{n\}) = \sum_{n=0}^{\infty} c = \begin{cases} 0 & c = 0, \\ \infty & c > 0. \end{cases}$$

(1) Axiom 2.
(2) Axiom 3.

Neither case equals one, so no such P exists. The same obstruction appears for a "flat prior" on an unbounded parameter range, such as a uniform distribution for a mass on $(0, \infty)$: it cannot be normalised. Such an *improper* prior can still be used as a device when the posterior turns out normalisable, and Chapter 5 explains when.

1.2.4 Which events are allowed? A word on σ-fields

On a finite or countable Ω every subset can be an event. On $\mathbb{R}$ this fails: one can prove that there is no way to give a "length" to *every* subset of $[0, 1]$ that is countably additive and unchanged by shifts. So probabilities are assigned only to a large, well-behaved class of subsets [Wasserman p. 13, §1.9].

σ-field, probability space. A **σ-field** (or σ-algebra) $\mathcal{A}$ is a collection of subsets of Ω with (i) $\varnothing \in \mathcal{A}$, (ii) if $A_1, A_2, \dots \in \mathcal{A}$ then $\bigcup_i A_i \in \mathcal{A}$, and (iii) if $A \in \mathcal{A}$ then $A^c \in \mathcal{A}$ [the class of events is closed under not, and countable or]. Its members are called **measurable**. The pair $(\Omega, \mathcal{A})$ is a **measurable space**, and with a probability measure P on $\mathcal{A}$ the triple $(\Omega, \mathcal{A}, \mathrm{P})$ is a **probability space**. On $\mathbb{R}$ one uses the **Borel σ-field**, the smallest σ-field containing all open intervals [everything you can build from intervals by countably many unions, intersections and complements].

In practice this restriction never bites. Every event in this book (intervals, histogram bins, excursion sets of a smooth field, limits of monotone sequences) is Borel. A σ-field is also closed under countable intersections, by De Morgan: $\bigcap_i A_i = \left(\bigcup_i A_i^c\right)^c$, and the right side uses only complements and unions. So the limits of decreasing sequences from Section 1.1 stay inside it too.

1.2.5 What the numbers mean: frequency and degree of belief

The axioms fix the arithmetic. They do not say what $\mathrm{P}(A) = 0.3$ *means*. There are two standard answers [Wasserman p. 6], [Cowan §1.2], [Kruschke §4.2].

Working definition 1.5 (frequency and degree-of-belief interpretations)

In the **frequency interpretation**, $\mathrm{P}(A)$ is the long-run proportion of repetitions in which A happens,

$$\mathrm{P}(A) = \lim_{n \to \infty} \frac{n_A}{n}, \tag{1.11}$$

where n_A is the **count** of A [the number of runs, out of n identical and independent runs, in which A occurred]. In the **degree-of-belief** (subjective, Bayesian) interpretation, $\mathrm{P}(A)$ measures how strongly a given observer, with given information, believes that the

statement A is true. Here Ω is often called the **hypothesis space** [a list of mutually exclusive statements, exactly one of which is true].

Both readings obey the same axioms, for different reasons. For frequencies the axioms are automatic. A frequency is never negative, the frequency of "some outcome" is one, and the frequency of one of several disjoint outcomes is the sum of their frequencies [Cowan p. 5].

For beliefs the axioms are a consistency requirement, and a bet shows why. Suppose you would pay 0.6 for a ticket that pays 1 if A happens, and also 0.6 for a ticket that pays 1 if A does not. Buying both costs 1.2. Exactly one of the tickets wins, so you get back exactly 1 whatever happens. Your beliefs $\mathrm{P}(A) = \mathrm{P}(A^c) = 0.6$ break the complement rule, and they lose money for sure. Beliefs that cannot be exploited this way must obey the axioms.

The same kind of comparison pins a belief down. Suppose you prefer a ticket on A to a ticket on drawing red from a bag with 1 red and 9 white marbles. Then your $\mathrm{P}(A)$ is above 0.1. Suppose you also prefer a ticket on a bag with 5 red and 5 white to the ticket on A. Then your $\mathrm{P}(A)$ is below 0.5. So it lies between 0.1 and 0.5 [Kruschke §4.2.2.1].

The frequency reading needs repetition, so it fails for a unique object. It is natural for anything that is repeated: collisions at a collider, photons in a detector, the simulated universes of a Monte Carlo. Now take the electron mass m_e. It is either inside the interval $[0.510, 0.512]$ MeV or not, and repeating the experiment does not change it. So the frequency of "m_e is in the interval" is either one (always inside) or zero (never), and we do not know which. A statement like "95% probability that m_e lies in this interval" only makes sense as a degree of belief [Cowan p. 6]. The same is true of "the probability that Ω_m lies between 0.30 and 0.32": there is one universe.

Taxonomy: Which interpretation, where in the book

Frequency	sampling distributions of data and statistics (Chapters 2 and 3), p-values and the coverage of confidence intervals (Chapter 4), Monte Carlo error bars (Chapter 6), cosmic variance of C_ℓ as scatter over hypothetical skies (Chapter 7), the transfer function and covariance of a masked pipeline (Chapter 8).
Degree of belief	priors and posteriors for parameters $p(\boldsymbol{\theta} \mid \boldsymbol{d})$ (Chapter 5), credible intervals from MCMC (Chapter 9), Fisher forecasts read as posterior widths (Chapter 10), systematic uncertainties.
Both, same rules	everything in the theory of probability. The calculus is shared, and in a Bayesian analysis the likelihood $\mathrm{P}(\boldsymbol{d} \mid \boldsymbol{\theta})$ is itself a frequency statement about repeated data.

1.2.6 A simulated long-run frequency

The frequency definition (1.11) is a limit, and a computer lets us watch the limit being approached [Wasserman p. 16, Ex. 21]. We flip a coin with P(heads) = p a thousand times and plot the proportion of heads so far.

Listing 1.2: Running proportion of heads (`code/ch01/02_running_frequency.py`)

```
for ax, p, tag in zip(axes, (0.3, 0.03), ("A", "B")):
    rng = rng_for("ch01", "1a_running_frequency", stream=int(1000 * p))
    ends = []
```

```
    for r in range(runs):
        flips = rng.random(n_max) < p          # True = heads, with probability p
        running = np.cumsum(flips) / n         # proportion of heads after n flips
        ax.plot(n, running, color=SERIES[r], lw=1.0, alpha=0.9)
        ends.append(running[-1])
    band = 2 * np.sqrt(p * (1 - p) / n)
    theory_line(ax, n, np.full_like(n, p, dtype=float), label=f"$p={p}$")
    ax.fill_between(n, p - band, p + band, color="0.85", zorder=0,
                    label=r"$p\pm2\sqrt{p(1-p)/n}$")
```

Line 27 is the whole simulation: `rng.random(n_max)` draws uniform numbers in $[0, 1)$, and a number lands below p with probability p, so the comparison is a coin with heads probability p. Line 28 divides the cumulative count of heads by the number of flips so far, which is n_A/n in (1.11). Line 31 draws the band $p \pm 2\sqrt{p(1-p)/n}$. Where that width comes from is the variance of the binomial distribution in Chapter 2. For now it is simply a rule of thumb: at any given n, a run lies outside the band only about one time in twenty.

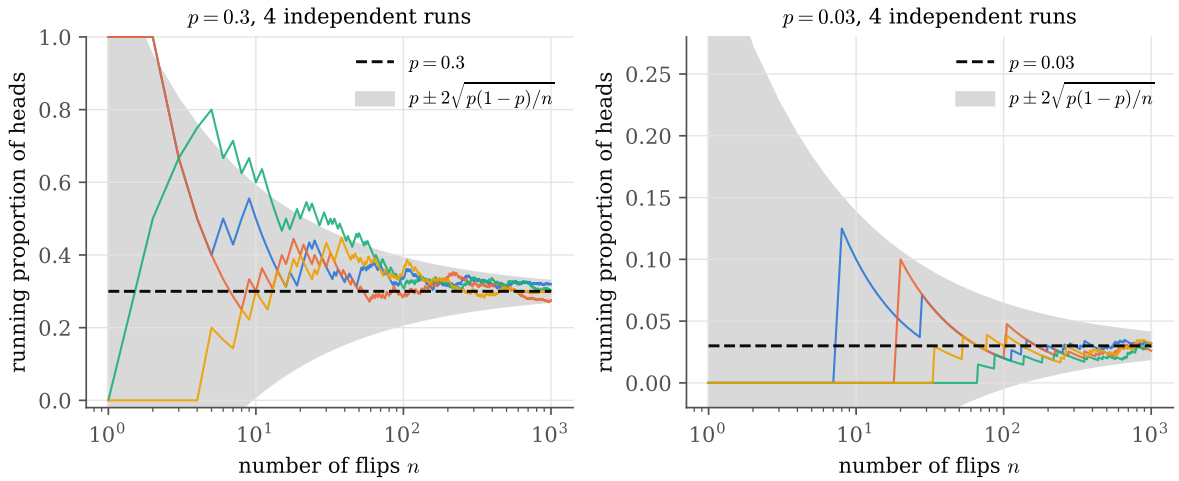


Figure 1.7: Running proportion of heads for four independent runs of 1000 flips, with $p = 0.3$ (left) and $p = 0.03$ (right). Early on the proportion jumps around wildly. As flips accumulate the runs settle towards the dashed value, and the shaded band $p \pm 2\sqrt{p(1-p)/n}$ narrows like $1/\sqrt{n}$.

A rare event is harder to pin down than a common one. After 1000 flips with $p = 0.3$, the first two runs end at 0.320 and 0.275, and 4 of the 4 runs end inside the band 0.3 ± 0.029. With $p = 0.03$ the first two end at 0.032 and 0.026, and 4 of the 4 end inside 0.03 ± 0.011. The band is narrower for the rare event in absolute terms. Relative to p, though, it is much wider: about 10% of p for $p = 0.3$, and almost 40% of p for $p = 0.03$. Dividing the half-width $2\sqrt{p(1-p)/n}$ by p gives the relative width $2\sqrt{(1-p)/(np)}$. It depends on np, the expected number of times the event occurs. So estimating a small probability to a given relative precision needs a number of trials proportional to $1/p$.

What this bought us. A finite run gives a frequency, not the probability. The frequency scatters around the probability, and the scatter shrinks like $1/\sqrt{(\text{number of occurrences})}$. This is the first appearance of the Monte Carlo error, which we will meet again every time a simulation estimates a number.

How this is used in real research: Verification here, instrument later

Here we *knew* p and the simulation only verified that frequencies approach it. In research the direction is reversed. Nobody can write down the probability that a pp collision at the LHC produces a muon that survives all selection cuts. It is obtained exactly this way, as a frequency over millions of simulated collisions, with the $1/\sqrt{n_A}$ scatter quoted as the "Monte Carlo statistical uncertainty".

The frequency definition is not a recipe for every probability. Equation (1.11) needs repeatable, identical runs. It gives no meaning to P(the Hubble tension is new physics). Conversely, a degree of belief is not arbitrary: once it is written as a probability it must obey all the rules above, and it must be updated by the rule of Section 1.7.

1.3 Equally likely outcomes: probability as counting

A silicon pixel sensor has $N = 10^4$ pixels. In one readout frame, $k = 100$ charged particles cross it, each landing on a pixel at random. If two particles hit the same pixel, the pixel reports one hit, the two tracks share it, and the tracking code may lose one of them. How often does at least one such merger happen in a frame? The occupancy is only $100/10^4 = 1\%$, so one expects "rarely". The answer is about 39%.

When every outcome is equally likely, a probability is a ratio of two counts, so all the work is in counting. Four basic counting formulas cover the cases: ordered or not, with or without replacement. Physicists know them as Maxwell–Boltzmann, Bose–Einstein and Fermi–Dirac counting. They settle the pixel problem, which is the birthday problem in disguise, and we solve it three ways: exactly, approximately and by simulation.

The microcanonical ensemble. In statistical mechanics an isolated system is equally likely to be in any of its $\Omega(E)$ microstates of energy E. The probability of a macrostate is then the number of microstates that realise it divided by $\Omega(E)$, and entropy $S = k_B \ln \Omega$ is a logarithm of a count. Probability on a finite sample space with equally likely outcomes is exactly this: fix the list of equally likely microstates, and every probability is a ratio of counts.

Solving a counting problem. Every counting problem here is solved by the same four steps.

1. Find a sample space in which every outcome is equally likely. (This is a physical statement about the mechanism, not a counting fact.)
2. Count all outcomes, $|\Omega|$.
3. Count the outcomes in the event, $|A|$. Break the count into a sequence of simple choices.
4. Divide. If counting A is hard, count A^c and use $1 - |A^c|/|\Omega|$.

Working definition 1.6 (uniform distribution on a finite set)

If $\Omega = \{\omega_1, \dots, \omega_N\}$ is finite and every outcome has the same probability, $\mathrm{P}(\{\omega_i\}) = 1/N$, then for every event

$$\mathrm{P}(A) = \frac{|A|}{|\Omega|}. \tag{1.12}$$

This is the **uniform probability distribution** on Ω [probability of A is the number of outcomes in A over the number of outcomes in Ω]. *Operational test:* can you argue, from

symmetry of the mechanism, that no outcome is preferred? If not, (1.12) does not apply.

The equal value $1/N$ is forced: the N single-outcome events are disjoint with union Ω, so $N \cdot c = 1$. Then

$$\mathrm{P}(A) \overset{(1)}{=} \sum_{\omega \in A} \mathrm{P}(\{\omega\}) \overset{(2)}{=} \sum_{\omega \in A} \frac{1}{N} \overset{(3)}{=} \frac{|A|}{N}.$$

(1) Finite additivity over the outcomes in A, (1.8).
(2) Every outcome has probability $1/N$.
(3) The sum has $|A|$ equal terms.

Example 1.10 (Two dice). Roll a die twice. $\Omega = \{(i, j) : i, j \in \{1, \dots, 6\}\}$ has 36 ordered pairs, all equally likely for fair dice rolled separately. The sum is 11 for $(5, 6)$ and $(6, 5)$, so $\mathrm{P}(\text{sum} = 11) = 2/36$. The sum is 7 for $(1, 6), (2, 5), \dots, (6, 1)$, so $\mathrm{P}(\text{sum} = 7) = 6/36$. Here the pitfall of Section 1.1 bites: on the coarser space of sums $\{2, \dots, 12\}$ the eleven values are *not* equally likely, and (1.12) would give the wrong answer $1/11$.

1.3.1 The multiplication principle

Fundamental theorem of counting. If a job consists of k separate choices, and the i-th choice can be made in n_i ways whatever the earlier choices were, the job can be done in $n_1 \times n_2 \times \cdots \times n_k$ ways.

For $k = 2$ choices, a table proves it. Write the possibilities with n_1 rows (first choice) and n_2 columns (second choice). Each cell is one way of doing the job, and there are $n_1 n_2$ cells. For k choices, treat the first $k - 1$ as one job and apply the $k = 2$ case again and again. The phrase "whatever the earlier choices were" matters. The *number* of options at each step must not depend on what was chosen before, although the options themselves may: when dealing two cards, the second card has 51 options whichever card came first, but which 51 depends on the first card.

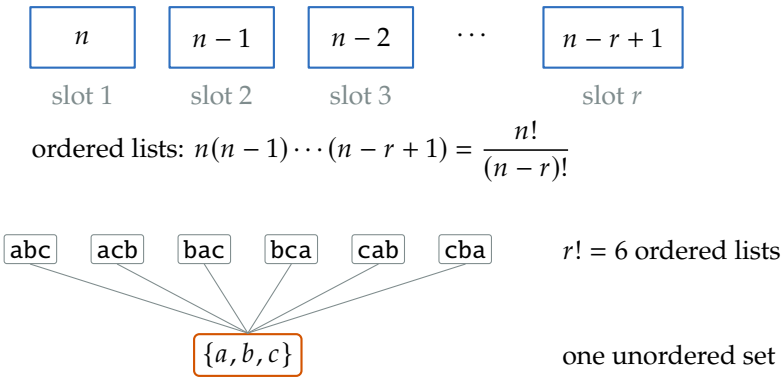


Figure 1.8: The upper panel fills r ordered slots from n distinct objects without replacement. Each choice leaves one fewer option for the next slot. The lower panel shows why this procedure overcounts unordered sets. All $3! = 6$ arrangements of a, b, c give the same set, and in general each set of r objects occurs $r!$ times in the ordered count.

Permutations. The number of ways to arrange n distinct objects in a row is $n! = n(n-1)\cdots 2 \cdot 1$. Call it P_n. There are n choices for the first position, and then the remaining $n - 1$ objects can be arranged in P_{n-1} ways, so $P_n = n\, P_{n-1}$ with $P_1 = 1$ [SeeingTheory p. 22]. Unwinding the recursion

gives $P_n = n!$. We set $0! = 1$: there is exactly one way to arrange nothing, and $P_1 = 1 \cdot P_0$ requires $P_0 = 1$.

Ordered selection without replacement. Filling r ordered slots from n objects (Figure 1.8, top) can be done in

$$n(n-1)\cdots(n-r+1) \overset{(1)}{=} \frac{n(n-1)\cdots(n-r+1)\,(n-r)!}{(n-r)!} \overset{(2)}{=} \frac{n!}{(n-r)!}$$

(1) Multiply and divide by $(n-r)! = (n-r)(n-r-1)\cdots 1$.
(2) The numerator is now the full product $n(n-1)\cdots 1$.

ways, by the multiplication principle with $n_i = n - i + 1$.

Unordered selection without replacement. Now we want the number of *sets* of r objects, ignoring order. Call it $\binom{n}{r}$. Every set of r objects can be put in order in $r!$ ways (Figure 1.8, bottom), and every ordered list arises from exactly one set. So we can count ordered lists in two ways:

$$\frac{n!}{(n-r)!} \overset{(1)}{=} \binom{n}{r} \times r! \quad \Longrightarrow \quad \binom{n}{r} = \frac{n!}{r!\,(n-r)!}. \tag{1.13}$$

(1) Choose the set first ($\binom{n}{r}$ ways), then its order ($r!$ ways), and apply the multiplication principle.

This is the **binomial coefficient**, read "n choose r" [Wasserman p. 8, (1.2)]. The method, counting something in two ways and equating, is the most useful trick in combinatorics.

Two properties follow. $\binom{n}{0} = \binom{n}{n} = 1$, since $0! = 1$: there is one way to choose nothing and one way to choose everything. And $\binom{n}{r} = \binom{n}{n-r}$, since choosing the r objects that are in is the same as choosing the $n - r$ that are out. Algebraically, swapping $r \leftrightarrow n - r$ in $n!/(r!(n-r)!)$ swaps the two factorials in the denominator.

Example 1.11 (Committees, and a poker hand). A class of 20 elects a committee of 3 [Wasserman p. 8]:

$$\binom{20}{3} \overset{(1)}{=} \frac{20!}{3!\,17!} \overset{(2)}{=} \frac{20 \times 19 \times 18}{3 \times 2 \times 1} = 1140.$$

(1) Definition (1.13).
(2) Cancel $17!$ from $20! = 20 \cdot 19 \cdot 18 \cdot 17!$.

A five-card poker hand is an unordered set of 5 of 52 cards, so there are $\binom{52}{5} = 2\,598\,960$ equally likely hands. "Four aces" fixes four cards and leaves 48 choices for the fifth, so $\mathrm{P}(\text{four aces}) = 48/2\,598\,960 \approx 1.8 \times 10^{-5}$. "Exactly one pair" is built by a sequence of choices [CasellaBerger (1.2.11)]: the pair's value (13 ways), its two suits ($\binom{4}{2} = 6$), the three other values, all different from each other and from the pair ($\binom{12}{3} = 220$), and a suit for each ($4^3 = 64$). So

$$\mathrm{P}(\text{exactly one pair}) = \frac{13 \cdot 6 \cdot 220 \cdot 64}{2\,598\,960} = \frac{1\,098\,240}{2\,598\,960} \approx 0.423.$$

Selection with replacement. If an object is put back after being chosen, every slot again has n options, and there are n^r ordered lists.

Unordered selection with replacement. This is the hardest case. An unordered choice of r objects from n types, repeats allowed, is fixed by how many of each type we took. Draw the n types as bins separated by $n-1$ walls, and each chosen object as a star in its bin (Figure 1.9). A choice is then a row of r stars and $n-1$ walls, and every such row is one choice. Counting rows means choosing which r of the $r+n-1$ positions hold stars:

$$\#\{\text{unordered selections with replacement}\} = \binom{n+r-1}{r}. \tag{1.14}$$



Figure 1.9: Stars and bars. Four objects of four possible types are recorded as a row of four stars and three walls. The stars between consecutive walls are the objects of one type, so this row means one of type 1, two of type 2, none of type 3 and one of type 4. The two outer walls (grey) never move and are not counted.

Key idea: The four ways to choose r objects from n

	without replacement	with replacement
ordered	$\dfrac{n!}{(n-r)!}$	n^r
unordered	$\binom{n}{r}$	$\binom{n+r-1}{r}$

In the New York lottery (6 numbers from 44) these are 5 082 517 440, 7 256 313 856, 7 059 052 and 13 983 816 tickets [CasellaBerger Table 1.2.1].

Counting in statistical mechanics. Put r particles into n single-particle states.

particles	counting	configurations
distinguishable (Maxwell–Boltzmann)	ordered, with replacement	n^r
identical bosons (Bose–Einstein)	unordered, with replacement	$\binom{n+r-1}{r}$
identical fermions (Fermi–Dirac)	unordered, without replacement	$\binom{n}{r}$

"With replacement" means a state can hold more than one particle. "Unordered" means we only record occupation numbers, not which particle is where.

Watch out: The count of outcomes is not a statement about their probabilities

Take $r = 2$ items from $n = 3$ with replacement, by drawing, noting and putting back. The nine ordered pairs are equally likely. The six unordered outcomes $\{1,1\},\{2,2\},\{3,3\},\{1,2\},\{1,3\},\{2,3\}$ are *not*: $\{1,2\}$ comes from $(1,2)$ and $(2,1)$ and has probability $2/9$, while $\{1,1\}$ has $1/9$ [CasellaBerger Ex. 1.2.19]. Formula (1.14) counts the unordered outcomes, but only the mechanism decides which list is equally likely. Physics gives a striking instance. For two particles in two states, the chance that both sit in the same state is $1/2$ if they are distinguishable (states 11, 22 out of 11, 12, 21, 22), $2/3$ for identical bosons (the three occupations $(2,0),(1,1),(0,2)$ are the equally likely quantum states), and

0 for fermions. That bosons "bunch" and fermions "anti-bunch" is measured directly in Hanbury Brown–Twiss correlations. It is a fact about nature, not about counting.

Try it: Toss until two heads. A fair coin is tossed until the second head appears [Wasserman p. 14, Ex. 5]. Show that exactly k tosses are needed with probability $(k-1)/2^k$ for $k \geq 2$. *Answer:* every particular sequence of k tosses has probability 2^{-k}. The event requires toss k to be heads and exactly one head among the first $k-1$ tosses, which can be placed in $\binom{k-1}{1} = k-1$ ways. As a check, $\sum_{k\geq 2}(k-1)2^{-k} = 1$: differentiate $\sum_{k\geq 0} x^k = 1/(1-x)$ to get $\sum_{k\geq 1} kx^{k-1} = 1/(1-x)^2$, and set $x = \frac{1}{2}$ after multiplying by x^2.

1.3.2 The birthday problem, and merged hits

Now back to the pixel sensor from the start of this section: $N = 10^4$ pixels, $k = 100$ particles per frame. Each of the k particles lands on one of the N pixels, independently and uniformly. (Independence is made precise in Section 1.6. Here it means that all N^k ordered lists of pixels, one entry per particle, are equally likely.) Let D be the event "at least two particles share a pixel". Its complement D^c, "all k pixels are different", is easier to count, because it is an ordered selection without replacement.

$$\mathrm{P}(D) \overset{(1)}{=} 1 - \mathrm{P}(D^c) \overset{(2)}{=} 1 - \frac{N(N-1)\cdots(N-k+1)}{N^k} \overset{(3)}{=} 1 - \prod_{j=0}^{k-1}\Big(1 - \frac{j}{N}\Big). \tag{1.15}$$

(1) Complement rule (1.5).
(2) Uniform distribution on the N^k ordered lists (ordered, with replacement). D^c contains the lists with all entries different (ordered, without replacement).
(3) Divide each of the k factors $N-j$ of the numerator by one factor N of the denominator.

For $N = 365$ days this is the birthday problem. To see why the answer grows so fast, take the logarithm and use $\ln(1-x) \approx -x$ for small x (the first term of the Taylor series $\ln(1-x) = -x - x^2/2 - \cdots$):

$$\ln \mathrm{P}(D^c) \overset{(1)}{=} \sum_{j=0}^{k-1} \ln\Big(1 - \frac{j}{N}\Big) \overset{(2)}{\approx} -\sum_{j=0}^{k-1} \frac{j}{N} \overset{(3)}{=} -\frac{k(k-1)}{2N}. \tag{1.16}$$

(1) The logarithm of a product is the sum of the logarithms.
(2) $\ln(1-x) \approx -x$, valid when $j/N \ll 1$, i.e. $k \ll N$.
(3) Arithmetic series, $0 + 1 + \cdots + (k-1) = k(k-1)/2$.

So $\mathrm{P}(D) \approx 1 - e^{-k(k-1)/2N}$. The exponent has a meaning: $k(k-1)/2 = \binom{k}{2}$ is the number of *pairs* of particles, and each pair shares a pixel with probability $1/N$. The number of pairs grows like k^2, which is why coincidences come early. The probability reaches one half when $k(k-1)/2N \approx \ln 2$, i.e. $k \approx \sqrt{2N \ln 2} \approx 1.18\sqrt{N}$.

Example 1.12 (Birthdays and pixels). For $N = 365$ and $k = 23$, $\binom{23}{2} = 253$ pairs, and (1.15) gives $\mathrm{P}(D) = 0.5073$, while the approximation gives $1 - e^{-253/365} = 0.5000$. The smallest group with better than even odds has $k = 23$ people, close to $1.18\sqrt{365} \approx 22.5$. With 50 people the probability is 0.970. For the pixel sensor, $N = 10^4$ and $k = 100$ give $\binom{100}{2} = 4950$ pairs, and P(some merged hit) = 0.3914 exactly (approximation 0.3904), at a mean occupancy of only 1%.

Listing 1.3: Exact, approximate and simulated birthday probabilities (`code/ch01/03_birthday.py`)

```
20 def p_shared_exact(k, N):
21     """1 - prod_{j=0}^{k-1} (N - j)/N, computed as a product of factors < 1."""
22     j = np.arange(k)
23     return 1.0 - np.prod((N - j) / N)
24
25
26 def p_shared_approx(k, N):
27     return 1.0 - np.exp(-k * (k - 1) / (2.0 * N))
28
29
30 def p_shared_mc(k, N, trials, rng):
31     """Draw k labels uniformly from N, many times; count draws with a repeat."""
32     draws = rng.integers(0, N, size=(trials, k))
33     draws.sort(axis=1)
34     has_repeat = np.any(draws[:, 1:] == draws[:, :-1], axis=1)
35     return has_repeat.mean()
```

The exact function (lines 20–23) evaluates the product in (1.15) as a product of numbers below one, which avoids the enormous factorials. The simulation (lines 30–35) draws a (trials $\times$ k) array of random labels in $\{0, \dots, N-1\}$, one row per group. Sorting each row puts equal labels next to each other, so a row contains a repeat exactly when two neighbours are equal. The mean of that boolean column is the frequency of D. With 200 000 simulated groups of 23 the frequency is 0.5070 ± 0.0011, and for 10^5 simulated pixel frames it is 0.3951 ± 0.0015. The error bars are the scatter $\sqrt{q(1-q)/N}$ of a frequency. The two simulated values lie 0.3 and 2.3 error bars from the exact ones.

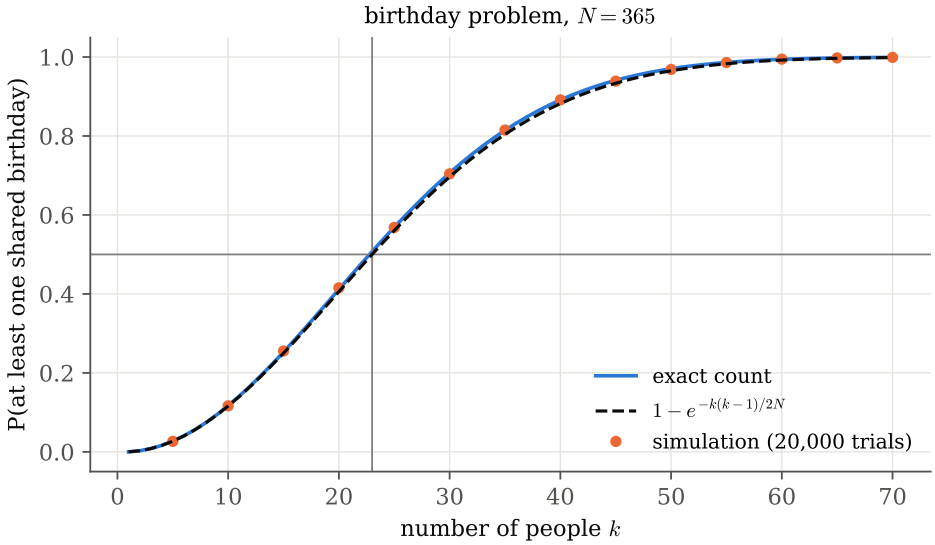


Figure 1.10: Probability that at least two of k people share a birthday. The exact count (solid) rises steeply and crosses one half between 22 and 23 people. The exponential approximation (dashed) lies just below it, and the simulated frequencies (dots) fall on the exact curve.

What this bought us. Coincidences are governed by the number of *pairs*, $\binom{k}{2} \approx k^2/2$, not by k. A collision among N equally likely labels becomes likely at $k \sim \sqrt{N}$ draws, far below N.

In research: where the $\sqrt{N}$ rule bites. Pixel and strip detectors are designed so that occupancy stays low enough for merged clusters to be rare, and the merging rate is estimated exactly as in

Example 1.12, then refined with a full simulation that includes the real, non-uniform hit density near the beam. The same rule applies to random seeds. If each of many parallel Monte Carlo jobs picks a 32-bit seed at random, two jobs share a seed (and produce identical "independent" samples) with probability one half after only $1.18\sqrt{2^{32}} \approx 77\,000$ jobs. This is why `common.py` derives seeds deterministically through `SeedSequence` (Chapter 6).

1.4 Conditional probability: shrinking the sample space

At a collider, the trigger decides in microseconds which collisions are kept. Everything an analyst later measures, such as the fraction of events with two muons, the fraction whose muons have opposite charge or the fraction that fall in a mass window, is a fraction *among triggered events*. The events that were not kept are gone. The analyst's world has shrunk from "all collisions" to "collisions that fired the trigger".

That is the whole idea: *information shrinks the sample space*. The question is this. Once we learn that an event B happened, what is the new probability of an event A? Conditional probability is the rule that answers it, by computing probabilities inside the shrunken space. It is how a probability changes when we learn something. Trees, the law of total probability, independence and Bayes' theorem are all built from it. In a simulation it is literally a selection: keep the rows in which B holds, and count among them.

In the area picture, conditioning changes only the unit of measurement. Before we learn anything, $\mathrm{P}(A)$ is the fraction of the whole box Ω covered by A. If we learn that the outcome lies in B, every point outside B is ruled out, and B becomes the new box. The probability of A is now the fraction of B covered by A: the area of the overlap AB, measured in units of the area of B (Figure 1.11). Nothing inside B is re-weighted relative to anything else inside B.

In Figure 1.11 the two regions are a hypothesis H and an observed event E. Their overlap HE (green) can be measured against either one. Measured against H, it is the share of the hypothesis region in which E also happens. Measured against E, it is the share of the observed region in which H was also true. One patch, two yardsticks. Bayes' theorem (Section 1.7) is the relation between those two readings of the same patch, so the picture already contains it.

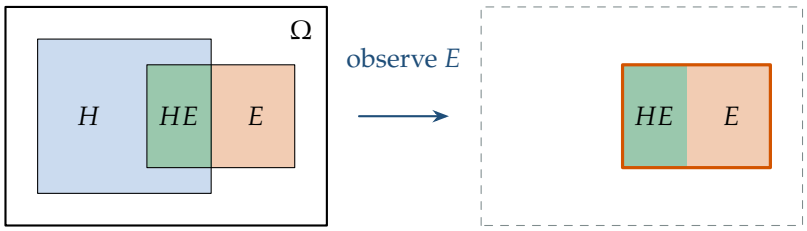


Figure 1.11: Conditioning as a change of reference space. Left: the hypothesis region H (blue) and the observed region E (orange) overlap in the green rectangle. Right: after we observe E, everything outside it is discarded and the green area is measured as a fraction of the orange one. Measured against H instead, the same green area gives $\mathrm{P}(E \mid H)$.

In words, computing a conditional probability takes four steps:

1. What do we know has happened? That event, B, is the new sample space.
2. Discard every outcome outside B.
3. Among the outcomes that remain, how much probability lies in A as well?

4. Rescale, so that the surviving outcomes carry total probability one.

These four steps determine the formula completely. Write $\mathrm{P}_B(A)$ for the new probability of A once B is known. We ask for three things: (i) P_B is a probability, (ii) $\mathrm{P}_B(B) = 1$, since B is now certain, and (iii) inside B nothing changes except the unit, $\mathrm{P}_B(C) = c\ \mathrm{P}(C)$ for every $C \subset B$, with one constant c. Then for any event A,

$$\mathrm{P}_B(A) \overset{(1)}{=} \mathrm{P}_B(AB) + \mathrm{P}_B(AB^c) \overset{(2)}{=} \mathrm{P}_B(AB) \overset{(3)}{=} c\ \mathrm{P}(AB), \qquad \text{and} \qquad 1 \overset{(4)}{=} \mathrm{P}_B(B) \overset{(5)}{=} c\ \mathrm{P}(B).$$

(1) A is the disjoint union of its parts inside and outside B, and P_B is additive by (i).
(2) $AB^c \subset B^c$, and $\mathrm{P}_B(B^c) = 1 - \mathrm{P}_B(B) = 0$ by (ii), so by monotonicity $\mathrm{P}_B(AB^c) = 0$.
(3) $AB \subset B$, so (iii) applies.
(4) Requirement (ii).
(5) Requirement (iii) with $C = B$.

The last equation fixes $c = 1/\mathrm{P}(B)$, and the first gives $\mathrm{P}_B(A) = \mathrm{P}(AB)/\mathrm{P}(B)$. There is no other choice.

Working definition 1.7 (conditional probability)

If $\mathrm{P}(B) > 0$, the **conditional probability of A given B** is

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(AB)}{\mathrm{P}(B)}. \tag{1.17}$$

Read the bar as "given" [restricted to the outcomes in which B is true]. *Frequency reading:* among the runs in which B happened, the fraction in which A also happened, $\mathrm{P}(A \mid B) \approx n_{AB}/n_B$ [n_{AB} counts runs with both, n_B runs with B, whatever A did]. *Operational test in code:* `A[B].mean()`.

Three remarks. First, $\mathrm{P}(A) = \mathrm{P}(A \mid \Omega)$: every probability is conditional on the sample space we chose [Cowan p. 2]. Second, the bar carries no arrow of time.

$$\mathrm{P}(\text{rain last night} \mid \text{wet ground this morning})$$

is a perfectly good conditional probability: among mornings with wet ground, the fraction that followed a rainy night [Kruschke §4.4.1]. Third, if A and B are disjoint then $\mathrm{P}(A \mid B) = 0$: once B has happened, A cannot have.

The shrinking sample space, drawn to scale. Very often a problem does not hand us $\mathrm{P}(AB)$ directly. It tells us how probable a situation H is, and how often an observation E occurs *inside* each situation. Take $\mathrm{P}(H) = 0.25$, and suppose E happens in 60% of the H-world and in 10% of the H^c-world. This is the picture of Figure 1.11 drawn to scale (Figure 1.12). Draw Ω as a rectangle of area one and H as a vertical strip of width 0.25. Inside each strip, mark the part where E happens as a piece running the full height of the strip: it covers 0.6 of the H strip and 0.1 of the H^c strip. Set the two pieces against the line between the strips, so that together they make one box, the event E. The share of a strip that its piece covers is itself a conditional probability: it is E measured against the strip as its reference space, so the two shares are $\mathrm{P}(E \mid H)$ and $\mathrm{P}(E \mid H^c)$.

Before we observe anything, $\mathrm{P}(H)$ is the width 0.25 of the H strip, a fraction of the whole rectangle. Now learn that E happened. Every point outside the box is ruled out. The surviving

region is the new sample space, and probabilities are re-measured against its area:

$$\mathrm{P}(H \mid E) \overset{(1)}{=} \frac{\text{area}(HE)}{\text{area}(E)} \overset{(2)}{=} \frac{\text{area}(HE)}{\text{area}(HE) + \text{area}(H^c E)} \overset{(3)}{=} \frac{0.25 \times 0.6}{0.25 \times 0.6 + 0.75 \times 0.1} = \frac{0.15}{0.225} = \frac{2}{3}.$$

(1) Definition of conditional probability as re-measured area. The areas are probabilities because Ω has area one.
(2) E is the union of its two pieces, which do not overlap because H and H^c are disjoint.
(3) Each piece is the strip's width times the share of the strip it covers, the width being the strip's probability and the share the conditional probability of E inside the strip.

Learning E moved the probability of H from $\frac{1}{4}$ to $\frac{2}{3}$. Why? The figure shows it. The H strip was only a third as wide as the H^c strip (0.25 against 0.75). But E is six times as common inside H as outside it (0.6 against 0.1). So, of the area that survives, the H piece (0.15) is twice the H^c piece (0.075), and H holds two thirds of it.

Step (3) used the multiplication rule $\mathrm{P}(HE) = \mathrm{P}(E \mid H)\,\mathrm{P}(H)$, read off the picture here and derived as (1.18) below. With every term named, and the denominator written as a sum over all hypotheses, the same calculation is Bayes' theorem (Section 1.7 and Chapter 5). Here it is only the definition of conditional probability, drawn to scale.

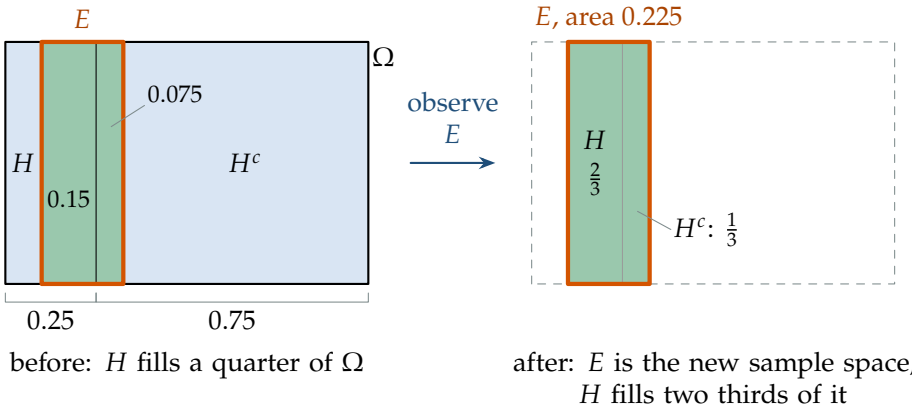


Figure 1.12: Conditioning drawn to scale, in the picture of Figure 1.11. Left: the rectangle Ω has area 1, split into the strip H of width 0.25 and the strip H^c of width 0.75. The event E (orange outline) covers 0.6 of the H strip and 0.1 of the H^c strip, so its two green pieces have areas $0.25 \times 0.6 = 0.15$ and $0.75 \times 0.1 = 0.075$. Right: once E is observed, everything outside the orange box is discarded. H keeps the share of the box that its piece fills, $0.15/0.225 = \frac{2}{3}$, although its strip was only a quarter of Ω.

Example 1.13 (Two dice, by counting in the shrunken space). Roll two fair dice. Let A be "the sum is 8" and B "at least one die shows 6". By counting, B has 11 of the 36 outcomes (six with a first 6, six with a second 6, minus the double-counted $(6,6)$), $A = \{(2,6),(3,5),(4,4),(5,3),(6,2)\}$ has 5, and $AB = \{(2,6),(6,2)\}$ has 2. Then

$$\mathrm{P}(A \mid B) \overset{(1)}{=} \frac{\mathrm{P}(AB)}{\mathrm{P}(B)} \overset{(2)}{=} \frac{2/36}{11/36} = \frac{2}{11} \approx 0.182, \qquad \mathrm{P}(B \mid A) = \frac{2/36}{5/36} = \frac{2}{5} = 0.4.$$

(1) Definition (1.17).
(2) Uniform distribution on 36 outcomes. The 1/36 cancels, so the answer is a ratio of counts *inside* B: $|AB|/|B|$.

Learning B raises the probability of A from $5/36 \approx 0.139$ to 0.182. The two conditional probabilities use the same overlap and differ only in what it is divided by.

Example 1.14 (Hair and eye colour: reallocating credibility). A sample of people has the joint proportions of eye and hair colour below [Kruschke Table 4.1] (rows do not add exactly because of rounding in the original data):

eye \ hair	black	brunette	red	blond	marginal
brown	0.11	0.20	0.04	0.01	0.37
blue	0.03	0.14	0.03	0.16	0.36
hazel	0.03	0.09	0.02	0.02	0.16
green	0.01	0.05	0.02	0.03	0.11
marginal	0.18	0.48	0.12	0.21	1.0

Learning "blue eyes" keeps only the second row and divides it by its total:

$$\mathrm{P}(\text{black} \mid \text{blue}) = \tfrac{0.03}{0.36} = 0.08, \qquad \mathrm{P}(\text{brunette} \mid \text{blue}) = \tfrac{0.14}{0.36} = 0.39,$$
$$\mathrm{P}(\text{red} \mid \text{blue}) = \tfrac{0.03}{0.36} = 0.08, \qquad \mathrm{P}(\text{blond} \mid \text{blue}) = \tfrac{0.16}{0.36} = 0.45.$$

The probability of blond hair rises from 0.21 to 0.45. The four numbers still add up to one, because the row was divided by its own total. Learning "blue eyes" has moved probability from some hair colours to others. Kruschke calls this a reallocation of credibility. It is already Bayesian inference in miniature.

1.4.1 $\mathrm{P}(\cdot \mid B)$ is a probability, $\mathrm{P}(A \mid \cdot)$ is not

For fixed B with $\mathrm{P}(B) > 0$, the function $A \mapsto \mathrm{P}(A \mid B)$ satisfies the three axioms [Wasserman p. 14, Ex. 9]. Axiom 1 holds because a ratio of a nonnegative number and a positive number is nonnegative. Axiom 2 holds because $\mathrm{P}(\Omega \mid B) = \mathrm{P}(\Omega B)/\mathrm{P}(B) = \mathrm{P}(B)/\mathrm{P}(B) = 1$. For Axiom 3, let $A_1, A_2, \ldots$ be disjoint:

$$\mathrm{P}\Big(\bigcup_i A_i \,\Big|\, B\Big) \overset{(1)}{=} \frac{\mathrm{P}\big((\bigcup_i A_i) \cap B\big)}{\mathrm{P}(B)} \overset{(2)}{=} \frac{\mathrm{P}\big(\bigcup_i (A_i B)\big)}{\mathrm{P}(B)} \overset{(3)}{=} \frac{\sum_i \mathrm{P}(A_i B)}{\mathrm{P}(B)} \overset{(4)}{=} \sum_i \mathrm{P}(A_i \mid B).$$

(1) Definition (1.17).
(2) Distributive law (1.2), for a countable family.
(3) The sets $A_i B$ are disjoint, because $A_i B \subset A_i$ and the A_i are. Then Axiom 3 for P.
(4) Split the sum and use the definition for each term.

So every rule of Section 1.2 holds with "$\mid B$" appended everywhere: $\mathrm{P}(A^c \mid B) = 1 - \mathrm{P}(A \mid B)$, inclusion–exclusion, Boole. This matters later, when all the machinery of probability is applied to a posterior, which is a probability conditional on the data.

The same is *not* true in the other slot. For fixed A, the map $B \mapsto \mathrm{P}(A \mid B)$ is not a probability. In general $\mathrm{P}(A \mid B \cup C) \neq \mathrm{P}(A \mid B) + \mathrm{P}(A \mid C)$, even for disjoint B and C. The quickest check: take $A = \Omega$, and any disjoint B and C of positive probability. Every conditional probability of Ω is one, so the left side is 1 and the right side is $1 + 1 = 2$. The rules of probability act only on events to the left of the bar [Wasserman p. 10].

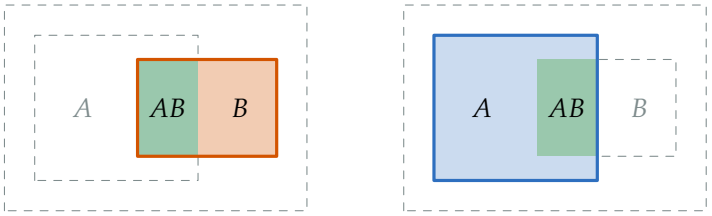
1.4.2 $\mathrm{P}(A \mid B)$ is not $\mathrm{P}(B \mid A)$

The probability of spots given measles is essentially one. The probability of measles given spots is small, since many things cause spots. The two conditional probabilities share the numerator

$\mathrm{P}(AB)$ and differ in the denominator. Figure 1.13 draws it with the rectangles of Figure 1.11, now labelled A and B. Conditioning on B makes B the whole world, and we ask what fraction of it the overlap fills. Swapping the roles makes A the whole world, and the same overlap is now measured against a bigger region. In the picture the overlap has area 1.6, B has area 3.68 and A has area 6.48, so

$$\mathrm{P}(A \mid B) = \frac{\text{area}(AB)}{\text{area}(B)} = \frac{1.6}{3.68} = 0.43, \qquad \mathrm{P}(B \mid A) = \frac{\text{area}(AB)}{\text{area}(A)} = \frac{1.6}{6.48} = 0.25.$$

Same patch, different yardstick, different answer. The two are equal only when A and B have the same probability.



condition on B: $\mathrm{P}(A \mid B) = \text{area}(AB)/\text{area}(B)$

condition on A: $\mathrm{P}(B \mid A) = \text{area}(AB)/\text{area}(A)$

Figure 1.13: Swapping what we condition on. The rectangles are those of Figure 1.11. Left: we learn that B happened, so B (orange outline) is the new reference space and the green overlap is read as a fraction of B. Right: we learn that A happened instead, so A (blue outline) is the reference space and the same green overlap is read as a fraction of the larger region A. Same numerator, different denominator: $\mathrm{P}(A \mid B) = 0.43$ but $\mathrm{P}(B \mid A) = 0.25$.

Mistaking one for the other is common enough in court to have a name, the **prosecutor's fallacy**: "the chance that an innocent person matches the DNA sample is one in a million, so the chance that the defendant is innocent is one in a million". In statistics the same mistake reads "$\mathrm{P}(\text{data this extreme} \mid H_0) = 0.01$, so $\mathrm{P}(H_0 \mid \text{data}) = 0.01$", which confuses a p-value with a posterior probability (Chapters 4 and 5).

Example 1.15 (A medical test). A test for a disease D has outcomes $+$ and $-$, and the joint probabilities are [Wasserman p. 10, Ex. 1.13]

	D	D^c
$+$	0.009	0.099
$-$	0.001	0.891

Condition on the column (the true state) to get the test's performance:

$$\mathrm{P}(+ \mid D) \overset{(1)}{=} \frac{\mathrm{P}(+ \cap D)}{\mathrm{P}(D)} \overset{(2)}{=} \frac{0.009}{0.009 + 0.001} = 0.9, \qquad \mathrm{P}(- \mid D^c) = \frac{0.891}{0.891 + 0.099} = 0.9.$$

(1) Definition (1.17).
(2) $\mathrm{P}(D)$ is the sum of its column, since $D = (+ \cap D) \cup (- \cap D)$ disjointly.

The test is right 90% of the time for the sick and for the healthy. Now condition on the row (the result) instead:

$$\mathrm{P}(D \mid +) \overset{(1)}{=} \frac{\mathrm{P}(+ \cap D)}{\mathrm{P}(+)} \overset{(2)}{=} \frac{0.009}{0.009 + 0.099} = \frac{9}{108} \approx 0.08.$$

(1) Definition (1.17), now with $+$ as the conditioning event.

(2) $\mathrm{P}(+)$ is the sum of the first row.

A positive result means the disease with probability 8%, not 90%. Among positives, the healthy outnumber the sick eleven to one, because there are ninety-nine times more healthy people to produce false positives (Figure 1.14).

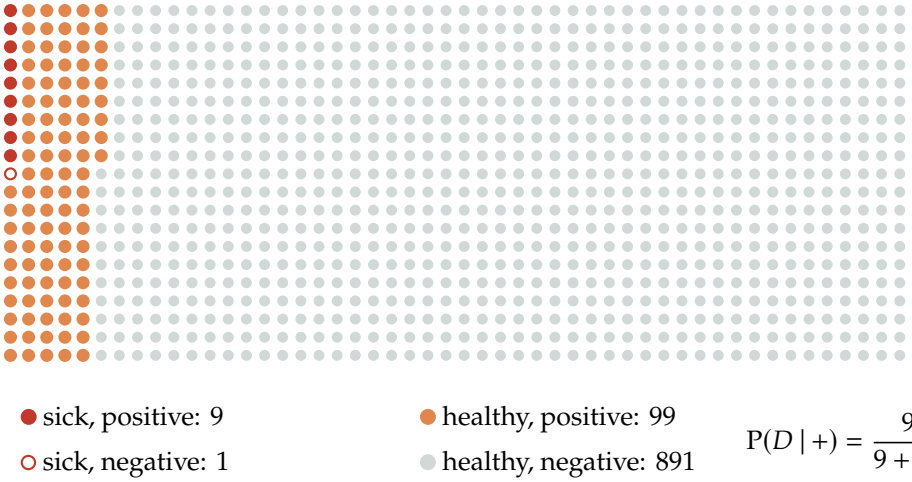


Figure 1.14: The medical test as a thousand people. The red dots at the top left are the ten who are sick, nine of whom test positive. The orange block is the ninety-nine healthy people who also test positive. Conditioning on a positive result keeps only the filled coloured dots, and the red ones are a small minority of them.

Example 1.16 (A radioactive nucleus does not age). A nucleus with decay constant λ survives to time t with probability $\mathrm{P}(S_t) = e^{-\lambda t}$, where S_t is "not yet decayed at time t". Given that it has survived to t, what is the probability that it survives a further time s? Surviving to $t+s$ implies surviving to t, so $S_{t+s} \subset S_t$ and $S_{t+s}S_t = S_{t+s}$. Then

$$\mathrm{P}(S_{t+s} \mid S_t) \overset{(1)}{=} \frac{\mathrm{P}(S_{t+s}S_t)}{\mathrm{P}(S_t)} \overset{(2)}{=} \frac{\mathrm{P}(S_{t+s})}{\mathrm{P}(S_t)} \overset{(3)}{=} \frac{e^{-\lambda(t+s)}}{e^{-\lambda t}} \overset{(4)}{=} e^{-\lambda s} = \mathrm{P}(S_s).$$

(1) Definition (1.17).
(2) $S_{t+s} \subset S_t$.
(3) The survival law.
(4) $e^{a+b} = e^a e^b$.

A nucleus that has already lived a thousand years is exactly as likely to survive the next second as a freshly made one. The exponential law is the only survival law with this **memoryless** property, and it is why the decay constant is meaningful at all. Taking $s = \mathrm{d}t$ small, $\mathrm{P}(\text{decay in } [t, t+\mathrm{d}t] \mid S_t) = 1 - e^{-\lambda\,\mathrm{d}t} \approx \lambda\,\mathrm{d}t$: the decay constant *is* a conditional probability per unit time, the same at every age. A human, or a photomultiplier tube, does age, and its survival law is not exponential.

1.4.3 The multiplication rule

Multiplying (1.17) by $\mathrm{P}(B)$, and doing the same with the roles of A and B exchanged, gives the **multiplication rule** [Wasserman p. 11, Lemma 1.14]:

$$\mathrm{P}(AB) = \mathrm{P}(A \mid B)\ \mathrm{P}(B) = \mathrm{P}(B \mid A)\ \mathrm{P}(A). \tag{1.18}$$

Read it as a two-step story: for A and B both to happen, first A must happen, and then, in the world where A has happened, B must happen [Lambert §3.6]. Applying it repeatedly gives the

chain rule [Wasserman p. 15, Ex. 17]:

$$\mathrm{P}(ABC) \overset{(1)}{=} \mathrm{P}\big(A \mid BC\big)\ \mathrm{P}(BC) \overset{(2)}{=} \mathrm{P}(A \mid BC)\ \mathrm{P}(B \mid C)\ \mathrm{P}(C), \tag{1.19}$$

(1) Multiplication rule (1.18) with the two events A and BC.
(2) Multiplication rule again, for $\mathrm{P}(BC)$.

and in general $\mathrm{P}(A_1A_2\cdots A_n) = \mathrm{P}(A_1)\,\mathrm{P}(A_2 \mid A_1)\,\mathrm{P}(A_3 \mid A_1A_2)\cdots\mathrm{P}(A_n \mid A_1\cdots A_{n-1})$. The chain rule is the arithmetic of a path through a tree, which is the subject of the next section.

Example 1.17 (Cards drawn without replacement). Draw two cards from a deck without putting the first back. Let A be "first is the ace of clubs" and B "second is the queen of diamonds". Then $\mathrm{P}(A) = 1/52$, and once A has happened the second card is uniform over the 51 cards that remain, one of which is the queen of diamonds. So $\mathrm{P}(AB) = \mathrm{P}(A)\,\mathrm{P}(B \mid A) = \frac{1}{52}\cdot\frac{1}{51}$ [Wasserman p. 11, Ex. 1.15].

Four aces in a row: with E_j "the j-th card is an ace",

$$\mathrm{P}(E_1E_2E_3E_4) \overset{(1)}{=} \mathrm{P}(E_1)\,\mathrm{P}(E_2 \mid E_1)\,\mathrm{P}(E_3 \mid E_1E_2)\,\mathrm{P}(E_4 \mid E_1E_2E_3) \overset{(2)}{=} \frac{4}{52}\cdot\frac{3}{51}\cdot\frac{2}{50}\cdot\frac{1}{49} = \frac{1}{270\,725}.$$

(1) Chain rule.
(2) At each step the shrunken deck is uniform: after j aces, $52-j$ cards remain, $4-j$ of them aces.

This equals $1/\binom{52}{4}$, as it must. Now condition on having started well [CasellaBerger Ex. 1.3.3]. Given that the first i cards are aces, the deck has $52-i$ cards with $4-i$ aces, and the probability that the next $4-i$ cards are exactly those aces is $1/\binom{52-i}{4-i}$. For $i = 1, 2, 3$ this is 0.00005, 0.00082 and 0.0204. Each piece of good news makes the event four aces far more probable, though still unlikely.

1.4.4 Conditioning is selection: a simulation

In code, conditioning on B means keeping only the rows where B is true. We repeat Example 1.13 with 200000 simulated rolls.

Listing 1.4: Conditioning by filtering (`code/ch01/04_conditioning_by_filtering.py`)

```
21 rng = rng_for("ch01", "1a_conditioning")
22 N = 200_000
23 d1 = rng.integers(1, 7, size=N)
24 d2 = rng.integers(1, 7, size=N)
25 A = (d1 + d2 == 8)
26 B = (d1 == 6) | (d2 == 6)
27
28 # --- conditioning = filtering: keep only the rows in B, then count A among them
29 A_given_B = A[B].mean()
30 B_given_A = B[A].mean()
31 # the same thing written as the ratio of frequencies, P(AB)/P(B)
32 ratio = (A & B).mean() / B.mean()
```

Lines 25–26 build the two masks. Line 29 is the whole of conditional probability: `A[B]` is the sub-array of A over the rows in which B happened, and its mean is the fraction of those rows in which A also happened, n_{AB}/n_B. Line 32 computes the same number as a ratio of two unconditional frequencies, $\hat{\mathrm{P}}(AB)/\hat{\mathrm{P}}(B)$, and the two agree exactly (0.1817 and 0.1817), because $n_{AB}/n_B = (n_{AB}/N)/(n_B/N)$.

Of the 200000 rolls, 61090 had at least one six, 27774 had sum eight, and 11100 had both. The estimates are $\hat{\mathrm{P}}(A \mid B) = 0.1817$ (exact $2/11 = 0.1818$) and $\hat{\mathrm{P}}(B \mid A) = 0.3997$ (exact $2/5 = 0.4000$).

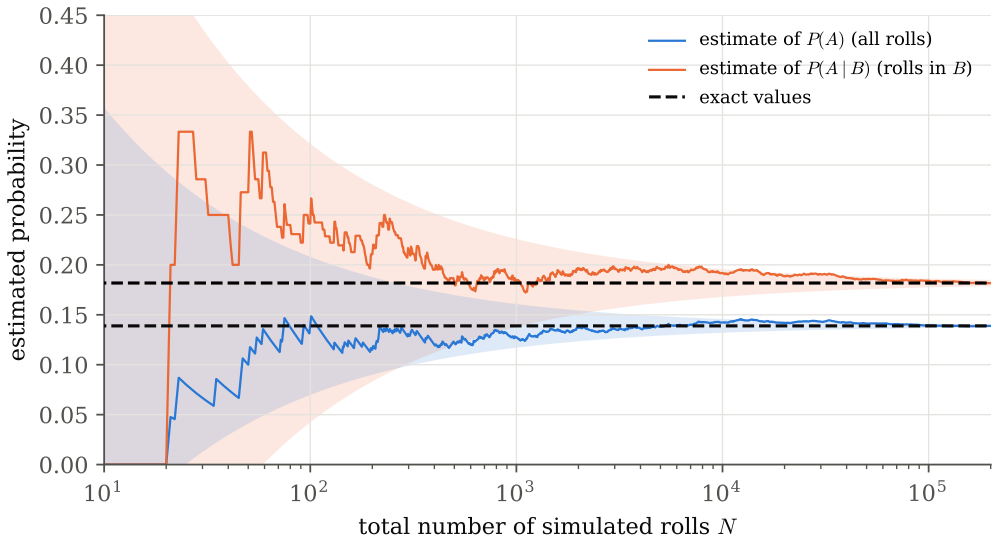


Figure 1.15: Running estimates of $\mathrm{P}(A)$ from all rolls (blue) and of $\mathrm{P}(A \mid B)$ from the rolls in B only (orange), against the total number of simulated rolls, with the exact values dashed. Both converge, but the conditional estimate wanders more and its band is wider, because only about 30% of the rolls survive the selection.

The figure shows the price of conditioning. The estimate of $\mathrm{P}(A \mid B)$ uses only $n_B \approx N\,\mathrm{P}(B)$ rolls, so its scatter is about $\sqrt{\mathrm{P}(A \mid B)(1-\mathrm{P}(A \mid B))/(N\,\mathrm{P}(B))}$, the running-frequency rule of Section 1.2 with n replaced by the number of survivors. Conditioning on a rare event leaves few survivors, and then the estimate is noisy however large N was.

So conditioning is selection followed by renormalisation. In a simulation it is one line, `A[B].mean()`, and its statistical error is set by how many events survive the selection, not by how many were generated.

Two traps are worth naming. First, $\mathrm{P}(A \mid B)$ and $\mathrm{P}(B \mid A)$ are different numbers, and they can be very different (Example 1.15), so always say in words which event is known. Second, (1.17) needs $\mathrm{P}(B) > 0$. Conditioning on an event of probability zero, such as "the measured energy is exactly 91.1876 GeV", is still meaningful but needs densities, which come in the next part.

How this is used in real research: Tag-and-probe

The efficiency of a muon identification algorithm is a conditional probability, P(identified | real muon). In data we do not know which tracks are real muons, so we make a subset in which we do. Select events with two tracks whose invariant mass is at the Z peak and in which one track (the "tag") passes a very tight identification. The other track (the "probe") is then almost certainly a muon, and the fraction of probes that pass the algorithm is the efficiency. The whole technique is the definition (1.17), implemented as a selection on real events. Here it is the research instrument: no formula gives this efficiency.

Everything in this section comes back to one picture: the conditioning event B becomes the new, shrunk sample space [Wasserman pp. 10–12, §1.6]. It is certain inside itself, $\mathrm{P}(B \mid B) = \mathrm{P}(B)/\mathrm{P}(B) = 1$ [CasellaBerger §1.3], and every other probability is re-measured as a share of it (Seeing Theory draws the same idea as "shelves", [SeeingTheory p. 27]). The medical test shows why we should not trust intuition here: the shrunk space can hold a very different mix of cases from the original one.

1.5 Probability trees and the law of total probability

One collision in a hundred contains a real high-momentum muon (S, for signal). The other 99% do not (B, background). The hardware level-1 trigger (L1) fires for 95% of the signal collisions and, because of fake muons from hadrons punching through the calorimeter, for 2% of the background. The collisions that pass L1 are examined by the software high-level trigger (HLT), which, *among those that passed L1*, keeps 90% of the signal and 10% of the background. What fraction of all collisions is written to disk, and how much of what is written is signal?

A situation built from successive stages is drawn as a tree. Two rules then answer both questions. The probability of each complete path is the product of the conditional probabilities along it. The probability of an observation is the sum over all the paths that lead to it. That sum is the law of total probability. Read the right way, it is also marginalisation: averaging out the stage we do not observe.

Water flowing through pipes. Pour one litre of water in at the root. At every junction the water divides among the outgoing pipes in fixed proportions (the conditional probabilities), and no water is lost, so the proportions at each junction add up to one. The water reaching the end of a pipe is the product of the proportions along the way. The total water collected in a bucket that several pipes drain into is the sum of what each pipe delivers.

Building and reading a tree. The recipe, in words, is as follows.

1. What are the stages of the situation, in the order they are easiest to describe? (Nature picks S or B, then L1 decides, then HLT decides.)
2. At each node, what are the mutually exclusive possibilities for the next stage, and what is the probability of each *given the path so far*?
3. The probability of a complete path is the product along it.
4. The probability of an event is the sum over the leaves where the event is true.

Working definition 1.8 (probability tree)

A **probability tree** is a rooted tree whose branches leaving each node are the possible results of the next stage [mutually exclusive and exhaustive, given everything above the node]. Each branch carries the conditional probability of its result given the path from the root to its node, so the labels leaving a node sum to one. A **path** from the root to a leaf is one complete history, and the leaves partition Ω [every outcome follows exactly one path]. *Rules:* multiply along a path, add across paths.

1.5.1 Multiply along a path

A path is an intersection of events, one per stage, so its probability is the chain rule (1.19). For the top path of Figure 1.16,

$$\mathrm{P}(S \cap \mathrm{L1} \cap \mathrm{HLT}) \overset{(1)}{=} \mathrm{P}(S)\ \mathrm{P}(\mathrm{L1} \mid S)\ \mathrm{P}(\mathrm{HLT} \mid S \cap \mathrm{L1}) \overset{(2)}{=} 0.01 \times 0.95 \times 0.90 = 0.00855. \tag{1.20}$$

(1) Chain rule, with the events taken in the order of the stages.
(2) The branch labels. The last one is the HLT efficiency *for signal that passed L1*, which is exactly how the HLT number was specified.

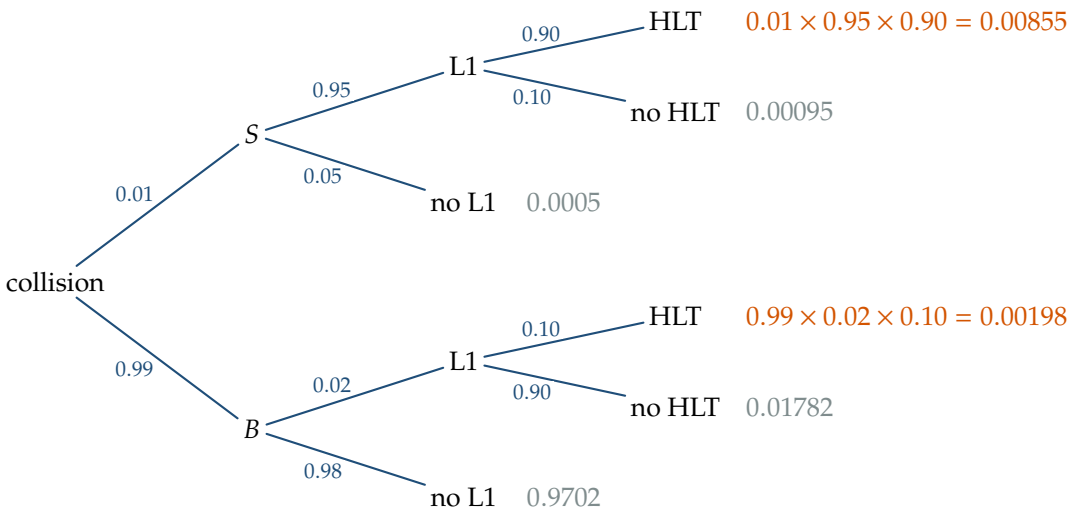


Figure 1.16: The two-level trigger as a tree. The first split is nature's choice between signal and background. Each later split is a trigger decision, and its numbers are conditional on everything to the left. The numbers at the right are the path probabilities, which add up to one. The two orange paths end in "written to disk".

The path probabilities at the right of Figure 1.16 sum to one. The reason is that the labels leaving each node sum to one. So the paths through the branches leaving a node add up to the probability of the path that reaches that node. Repeating this from the leaves back to the root leaves the root's probability, one. In the water picture, no water is lost at a junction.

1.5.2 Add across paths: the law of total probability

Let "accepted" be Acc = L1 ∩ HLT. Two paths end in it, one through S and one through B. They are disjoint (a collision cannot be both signal and background), so their probabilities add:

$$\mathrm{P}(\text{Acc}) = 0.00855 + 0.99 \times 0.02 \times 0.10 = 0.00855 + 0.00198 = 0.01053.$$

About one collision in a hundred is kept. Nothing here was special to the trigger: whenever the first stage splits Ω into disjoint pieces, the probability of an event is the sum over the paths that end in it.

Key idea: Law of total probability

Let $A_1, \ldots, A_k$ be a partition of Ω. Then for any event B,

$$\mathrm{P}(B) = \sum_{i=1}^{k} \mathrm{P}(B \mid A_i)\ \mathrm{P}(A_i). \tag{1.21}$$

The first stage of the tree is the partition, the second stage is "does B happen", and (1.21) adds the paths that end in B.

To prove it, cut B along the partition as in Figure 1.3: $C_i = BA_i$. Every outcome of B lies in exactly one A_i, so the C_i are disjoint and $B = \bigcup_i C_i$ [Wasserman p. 12, Thm 1.16]. Then

$$\mathrm{P}(B) \overset{(1)}{=} \mathrm{P}\Big(\bigcup_i C_i\Big) \overset{(2)}{=} \sum_i \mathrm{P}(BA_i) \overset{(3)}{=} \sum_i \mathrm{P}(B \mid A_i)\ \mathrm{P}(A_i).$$

(1) B is the union of its pieces C_i.

(2) Finite additivity for the disjoint pieces $C_i = BA_i$.
(3) Multiplication rule (1.18) for each piece. If some $\mathrm{P}(A_i) = 0$, the term $\mathrm{P}(BA_i) = 0$ is simply dropped.

The law holds just as well for a countable partition $A_1, A_2, \ldots$, by Axiom 3 in place of finite additivity.

Total probability is an average. The weights $\mathrm{P}(A_i)$ in (1.21) are nonnegative and sum to one. So $\mathrm{P}(B)$ is a *weighted average* of the conditional probabilities $\mathrm{P}(B \mid A_i)$, and it lies between the smallest and the largest of them. In the trigger, the acceptance 0.01053 lies between the background acceptance 0.002 ($= 0.02 \times 0.10$) and the signal acceptance 0.855 ($= 0.95 \times 0.90$). It sits very close to the background value, because background carries 99% of the weight.

This averaging is what it means to **marginalise**. We average out the stage we do not observe, here the true nature of the collision. That stage plays the role of a *nuisance* variable: it matters for the answer, but we do not ask about it. In Chapter 5 the same sum becomes an integral over parameters $\boldsymbol{\theta}$ given data $\boldsymbol{d}$, $p(\boldsymbol{d}) = \int p(\boldsymbol{d} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}$, the evidence.

Example 1.18 (The 662 keV line of caesium-137). A ^{137}Cs nucleus β-decays either to the excited isomer ^{137m}Ba (probability about 0.944) or directly to the ground state of ^{137}Ba (about 0.056). The isomer de-excites by emitting a 662 keV photon with probability about 0.902. Otherwise the energy is carried off by an internal-conversion electron. The direct branch emits no 662 keV photon. The probability that one decay produces a 662 keV photon is

$$\begin{aligned} \mathrm{P}(\gamma_{662}) &\overset{(1)}{=} \mathrm{P}(\gamma_{662} \mid {}^{137m}\mathrm{Ba})\ \mathrm{P}({}^{137m}\mathrm{Ba}) + \mathrm{P}(\gamma_{662} \mid \text{ground})\ \mathrm{P}(\text{ground}) \\ &\overset{(2)}{=} 0.902 \times 0.944 + 0 \times 0.056 \approx 0.851, \end{aligned}$$

(1) Law of total probability over the two decay branches.
(2) Branch probabilities from the nuclear data tables.

which is the "85.1% gamma intensity" printed in the tables. A third stage, "the photon deposits its full energy in the detector" with probability ε, multiplies the useful path by ε. A source of activity $\mathcal{A}$ (decays per second) therefore gives a photopeak rate $0.851\, \varepsilon \mathcal{A}$. Every calibration with a radioactive source uses this tree.

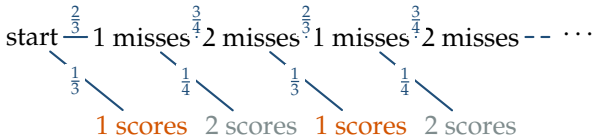
Example 1.19 (A coin from a bag). A bag holds a fair coin and a coin with $\mathrm{P}(\text{heads}) = 0.95$. We pick one blindly and toss it three times [SeeingTheory p. 27]. Let F be "fair coin picked" and H_3 "three heads". Given the coin, the tosses are independent (Section 1.6), so $\mathrm{P}(H_3 \mid F^c) = 0.95^3 = 0.857$ and $\mathrm{P}(H_3 \mid F) = 0.5^3 = 0.125$. Then

$$\mathrm{P}(H_3) \overset{(1)}{=} \mathrm{P}(H_3 \mid F^c)\, \mathrm{P}(F^c) + \mathrm{P}(H_3 \mid F)\, \mathrm{P}(F) = 0.857 \times 0.5 + 0.125 \times 0.5 \approx 0.491.$$

(1) Law of total probability over the partition $\{F, F^c\}$.

We will reverse this tree in Section 1.7 to ask which coin we hold.

Example 1.20 (Taking turns: a tree with infinitely many paths). Two players take turns shooting at a basket, player 1 first. Player 1 scores with probability 1/3, player 2 with probability 1/4, independently each time. What is the probability that player 1 scores first [Wasserman p. 9, Ex. 1.11]?



Let A_j be "the first basket is by player 1, on her j-th attempt". Each A_j is one path: $j-1$ rounds of "both miss", then player 1 scores. The A_j are disjoint and their union is the event E we want. Each round of two misses has probability $\frac{2}{3}\cdot\frac{3}{4}=\frac{1}{2}$, so $\mathrm{P}(A_j)=(\frac{1}{2})^{j-1}\frac{1}{3}$, and

$$\mathrm{P}(E)\overset{(1)}{=}\sum_{j=1}^{\infty}\mathrm{P}(A_j)\overset{(2)}{=}\frac{1}{3}\sum_{j=1}^{\infty}\left(\frac{1}{2}\right)^{j-1}\overset{(3)}{=}\frac{1}{3}\cdot\frac{1}{1-\frac{1}{2}}=\frac{2}{3}.$$

(1) Axiom 3, since the paths A_j are disjoint.
(2) Multiply along each path.
(3) Geometric series, $\sum_{m=0}^{\infty} r^m = 1/(1-r)$ for $0 \le r < 1$. It follows from $(1-r)(1+r+\cdots+r^M) = 1-r^{M+1}$ and $r^{M+1}\to 0$. More generally $\sum_{j=k}^{\infty} r^j = r^k/(1-r)$.

A shortcut avoids the series. Look only at the first round. With probability $\frac{1}{3}$ player 1 scores at once. With probability $\frac{1}{2}$ both miss, and then the game starts afresh, so player 1 scores first with probability $\mathrm{P}(E)$ again. In the remaining case player 2 scored. So $\mathrm{P}(E)=\frac{1}{3}+\frac{1}{2}\mathrm{P}(E)$, which again gives $\mathrm{P}(E)=\frac{2}{3}$.

Try it: Until both faces have appeared. Toss a fair coin until both a head and a tail have appeared [Wasserman p. 15, Ex. 13]. The sample space is $\{HT, TH, HHT, TTH, HHHT, TTTH, \dots\}$. Show that exactly three tosses are needed with probability $1/4$. *Answer:* two paths, HHT and TTH, each of probability $(1/2)^3$.

1.5.3 Following each collision down the tree

The trigger tree can be simulated literally: generate the nature of each collision, then each trigger decision with the probability appropriate to that branch.

Listing 1.5: The trigger tree, exactly and by simulation (`code/ch01/05_trigger_chain.py`)

```
24 # --- exact: multiply along each path, add the paths that end in "accepted"
25 path_S = s * e1S * e2S
26 path_B = (1 - s) * e1B * e2B
27 p_acc = path_S + path_B
28 purity = path_S / p_acc      # used again in the Bayes section
29
30 # --- simulation: follow every collision down the tree
31 rng = rng_for("ch01", "1a_trigger_chain")
32 N = 2_000_000
33 is_S = rng.random(N) < s
34 l1 = np.where(is_S, rng.random(N) < e1S, rng.random(N) < e1B)
35 hlt = l1 & np.where(is_S, rng.random(N) < e2S, rng.random(N) < e2B)
36 mc_acc = hlt.mean()
37 mc_path_S = (hlt & is_S).mean()
38 mc_path_B = (hlt & ~is_S).mean()
39 mc_purity = is_S[hlt].mean()        # conditioning = filtering on "accepted"
```

Lines 25–28 are the tree read as arithmetic: multiply along the two accepted paths and add them. Line 33 is nature's choice at the root. Line 34 is the L1 split: `np.where` uses the signal efficiency on signal rows and the background efficiency on background rows, which is precisely what "conditional on the branch" means. Line 35 applies the HLT only to rows that passed L1 (the `l1 &`), again with a branch-dependent efficiency. Lines 37–38 count the accepted collisions separately on each path. Line 39 is a preview of Section 1.7: among accepted collisions, the fraction that is signal.

With 2×10^6 simulated collisions, 21248 are accepted, a frequency of 0.01062 against the exact 0.01053. The signal path contributes 0.00865 (exact 0.00855) and the background path 0.00198 (exact 0.00198).

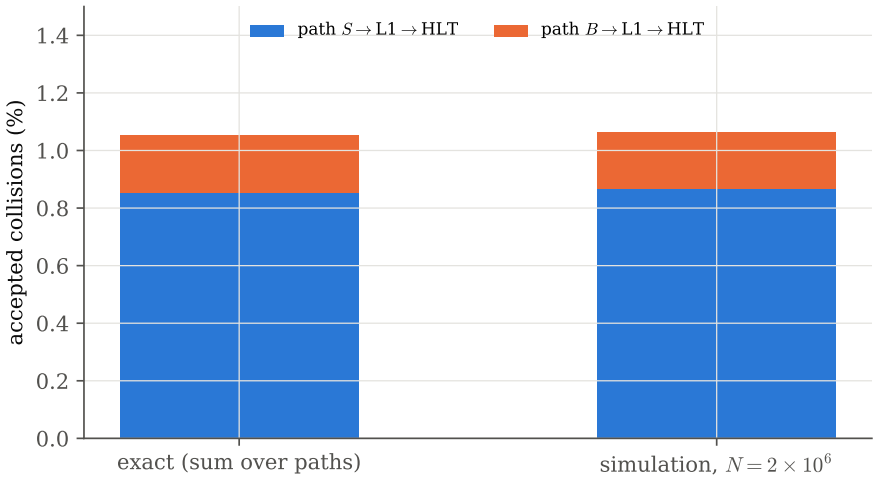


Figure 1.17: The fraction of collisions written to disk, split into the two paths of the tree that lead there. The left bar is the exact sum over paths, the right bar the simulation. Signal (blue) supplies most of the accepted collisions although only one collision in a hundred is signal.

What this bought us. A tree turns a many-stage problem into bookkeeping: conditional probabilities on the branches, products along paths, sums across paths. The law of total probability is the sum across the paths that end in the observed event, and it is a weighted average of the branch-wise probabilities.

In research: trigger rates and emulation. In a real experiment each trigger level has a rate budget (for example 100 kHz out of L1 and a few kHz out of the HLT), and the total rate is computed exactly as a sum over the physics processes that can fire it, each weighted by its cross section: the law of total probability with the processes as the partition. Here the simulation only verified the tree. In practice the branch efficiencies depend on the momentum, angle and pile-up of each collision, and no closed formula exists. One then runs the trigger emulation on millions of simulated events and counts, so the simulation is the instrument and the tree is the check that one understands the result.

Watch out: Branch labels are conditional on the path

Two traps. First, the HLT keeps 90% of signal *that passed L1*. That is P(HLT | $S \cap$ L1), not P(HLT | S) and certainly not P(HLT). Multiplying unconditional probabilities along a path is correct only when the stages are independent, which is the subject of the next section. Second, the paths that are added must be disjoint. The leaves of a tree always are, which is why drawing the tree is safer than adding probabilities by eye.

1.6 Independence: when learning B tells us nothing about A

In Example 1.6 we used a table in which the pattern 111 had probability $0.729 = 0.9^3$ and each pattern with one silent paddle had $0.081 = 0.9^2 \times 0.1$. Those numbers come from one assumption: each paddle fires with probability 0.9 *whatever the other paddles do*. The paddles are

separate pieces of plastic with separate photomultipliers, so this is plausible. It fails if the three tubes share a high-voltage supply that sometimes sags, because then one silent paddle makes it more likely that the others are silent too.

Independence is the precise version of "whatever the others do". It is the one situation in which the probability of "A and B" is the product of the two probabilities. That product rule lets us write the probability of a data set as a product over data points, and every likelihood in this book is built that way. There are two things that independence is not. First, independence is not the same as disjointness. Second, when there are three or more events, independence does not mean merely that every pair of events is independent.

Crossed strips. Draw Ω as the unit square. Let A be a vertical strip of width $\mathrm{P}(A)$ and B a horizontal strip of height $\mathrm{P}(B)$. Their overlap is a rectangle of area $\mathrm{P}(A)\,\mathrm{P}(B)$. Inside A, the strip B covers the same fraction $\mathrm{P}(B)$ as it covers of the whole square, so learning "A happened" does not change the probability of B. Independence is this product structure. Two dice, two separate detectors or two separate nuclei behave like the two axes of the square.

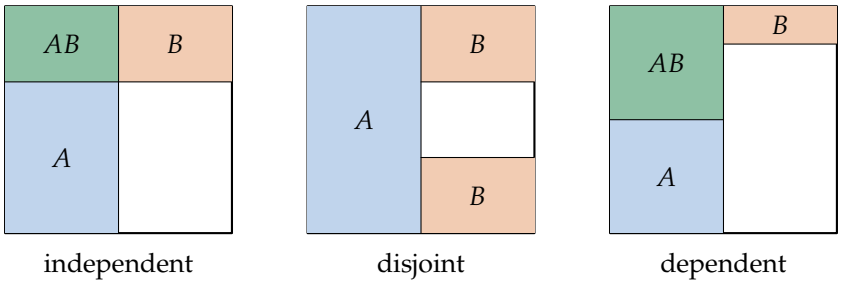


Figure 1.18: Three ways for two events to sit in the unit square, all drawn to scale and all with $\mathrm{P}(A) = \frac{1}{2}$ (the left half) and $\mathrm{P}(B) = \frac{1}{3}$. Left: B is the top third, so it fills a third of A and a third of the square, and $\mathrm{P}(AB) = \frac{1}{6}$, exactly the product: independent. Middle: A and B do not meet, so $\mathrm{P}(AB) = 0$ and knowing A rules B out. Right: B still has area $\frac{1}{3}$ but leans into A, so $\mathrm{P}(AB) = \frac{1}{4} \neq \frac{1}{6}$. The two overlaps have the same width, so compare their heights: B covers half of A on the right but only a third of A on the left. Covering more of A than its $\frac{1}{3}$ share of the square, B becomes more likely once we learn A.

Checking independence. Whenever a calculation seems to need independence, we ask four questions.

1. In words: does knowing that B happened change how likely A is?
2. In numbers: is $\mathrm{P}(AB)$ equal to $\mathrm{P}(A)\,\mathrm{P}(B)$?
3. Is independence being *assumed* (no physical mechanism links the two) or *derived* (computed from a known distribution)?
4. For more than two events, check every sub-collection, not just the pairs.

Working definition 1.9 (independence)

Two events A and B are **independent**, written $A \perp\!\!\!\perp B$, if

$$\mathrm{P}(AB) = \mathrm{P}(A)\ \mathrm{P}(B). \tag{1.22}$$

A collection $\{A_i : i \in I\}$ is **independent** (or **mutually independent**) if $\mathrm{P}\big(\bigcap_{i\in J} A_i\big) = \prod_{i\in J} \mathrm{P}(A_i)$ for *every* finite subset $J \subseteq I$ [for three events that means the three pairs and the triple, four equations in all]. Otherwise the events are dependent. *Operational test:* with

$\mathrm{P}(B) > 0$, independence is $\mathrm{P}(A \mid B) = \mathrm{P}(A)$.

The conditional form is the meaning, and the product form is the convenient definition, because it is symmetric and also makes sense when $\mathrm{P}(B) = 0$. They agree whenever $\mathrm{P}(B) > 0$ [Wasserman p. 11, Lemma 1.14]:

$$\mathrm{P}(A \mid B) \overset{(1)}{=} \frac{\mathrm{P}(AB)}{\mathrm{P}(B)} \overset{(2)}{=} \frac{\mathrm{P}(A)\,\mathrm{P}(B)}{\mathrm{P}(B)} = \mathrm{P}(A),$$

(1) Definition of conditional probability.
(2) Independence (1.22).

and conversely, if $\mathrm{P}(A \mid B) = \mathrm{P}(A)$ then multiplying by $\mathrm{P}(B)$ gives (1.22). By the symmetry of (1.22), $\mathrm{P}(B \mid A) = \mathrm{P}(B)$ too when $\mathrm{P}(A) > 0$: if B tells us nothing about A, then A tells us nothing about B.

Assumed or derived. Often independence is an *assumption* that expresses physics: a coin has no memory of its last toss, two nuclei do not communicate, the noise in two separate amplifiers has separate sources. Sometimes it is a *result* [Wasserman p. 8]. Roll one fair die and let $A = \{2, 4, 6\}$ and $B = \{1, 2, 3, 4\}$. Then $AB = \{2, 4\}$ and

$$\mathrm{P}(AB) = \tfrac{2}{6} = \tfrac{1}{3} = \tfrac{1}{2} \times \tfrac{2}{3} = \mathrm{P}(A)\,\mathrm{P}(B),$$

so A and B are independent, although both are about the same single roll. Nobody assumed it. With $C = \{4, 5, 6\}$ instead, $AC = \{4, 6\}$ and $\mathrm{P}(AC) = \frac{1}{3} \neq \frac{1}{2} \times \frac{1}{2}$: knowing the roll is high makes it more likely to be even.

1.6.1 Consequences

Complements of independent events are independent. If $A \perp\!\!\!\perp B$,

$$\mathrm{P}(AB^c) \overset{(1)}{=} \mathrm{P}(A) - \mathrm{P}(AB) \overset{(2)}{=} \mathrm{P}(A) - \mathrm{P}(A)\,\mathrm{P}(B) \overset{(3)}{=} \mathrm{P}(A)\big(1 - \mathrm{P}(B)\big) \overset{(4)}{=} \mathrm{P}(A)\,\mathrm{P}(B^c).$$

(1) A is the disjoint union of AB and AB^c, so $\mathrm{P}(A) = \mathrm{P}(AB) + \mathrm{P}(AB^c)$.
(2) Independence of A and B.
(3) Factor out $\mathrm{P}(A)$.
(4) Complement rule.

So $A \perp\!\!\!\perp B^c$ [Wasserman p. 15, Ex. 11], [CasellaBerger Thm 1.3.9]. Applying the same argument with the roles of the events exchanged gives $A^c \perp\!\!\!\perp B^c$. If a paddle's firing tells us nothing about another's, neither does its silence.

Sure and impossible events are independent of everything [Wasserman p. 15, Ex. 14]. If $\mathrm{P}(A) = 0$, then $\mathrm{P}(AB) \leq \mathrm{P}(A) = 0$ by monotonicity, so $\mathrm{P}(AB) = 0 = \mathrm{P}(A)\,\mathrm{P}(B)$. If $\mathrm{P}(A) = 1$, then $\mathrm{P}(A^c) = 0$, so A^c is independent of every B, and by the complement result so is A. Finally, an event independent of itself satisfies $\mathrm{P}(A) = \mathrm{P}(AA) = \mathrm{P}(A)^2$, so $\mathrm{P}(A) \in \{0, 1\}$.

Key idea: Disjoint is not independent

If A and B are disjoint and both have positive probability, they are *dependent*: $\mathrm{P}(A)\,\mathrm{P}(B) > 0$, but $\mathrm{P}(AB) = \mathrm{P}(\varnothing) = 0$ [Wasserman p. 9]. Disjoint events are as dependent as events can be: learning that one happened tells us for certain that the other did not.

Apart from this case, a Venn diagram cannot show independence. Whether the overlap equals the product depends on the areas, not on the shapes. In Figure 1.18 the right panel looks as "overlapping" as the left one, yet only the left is independent.

Two more warnings [Lambert §3.4]. First, take a playing card. Its colour and its suit are dependent: "red" forces hearts or diamonds. Its suit and its value are independent: $\mathrm{P}(\heartsuit \mid \text{queen}) = \frac{1}{4} = \mathrm{P}(\heartsuit)$. Second, dependence is not causation. The colour does not *cause* the suit. Dependence only means that knowing one tells us something about the other.

1.6.2 Examples

Example 1.21 (At least one head in ten tosses). Let A be "at least one head" in ten tosses of a fair coin and T_j "tails on toss j" [Wasserman p. 9, Ex. 1.10]. Then

$$\mathrm{P}(A) \overset{(1)}{=} 1 - \mathrm{P}(A^c) \overset{(2)}{=} 1 - \mathrm{P}(T_1 T_2 \cdots T_{10}) \overset{(3)}{=} 1 - \mathrm{P}(T_1)\,\mathrm{P}(T_2) \cdots \mathrm{P}(T_{10}) \overset{(4)}{=} 1 - \left(\frac{1}{2}\right)^{10} \approx 0.999.$$

(1) Complement rule.
(2) "No head" means "tails on every toss".
(3) The tosses are independent (assumed: the coin has no memory).
(4) Each factor is $\frac{1}{2}$.

The Chevalier de Méré's bet "at least one six in four rolls of a die" is the same computation, $1 - (5/6)^4 \approx 0.518$ [CasellaBerger Ex. 1.3.8], a slightly favourable bet. His other bet, "at least one double six in 24 rolls of two dice", gives $1 - (35/36)^{24} \approx 0.491$, slightly unfavourable, which puzzled him enough to write to Pascal.

Example 1.22 (Telescope efficiency, and the first decay in a sample). *(a)* Each paddle fires with probability ε, independently. The trigger needs at least two paddles. The pattern 111 has probability ε^3, and each of the three patterns with exactly one silent paddle has probability $\varepsilon^2(1-\varepsilon)$, by the product rule for the three independent events "paddle j fired or not" (complements are independent too). So

$$\mathrm{P}(T) \overset{(1)}{=} 3\varepsilon^2(1-\varepsilon) + \varepsilon^3 = 3\varepsilon^2 - 2\varepsilon^3,$$

(1) Finite additivity over the four disjoint patterns in T.

which is 0.972 at $\varepsilon = 0.9$, reproducing the table of Example 1.6. A two-out-of-three majority turns 90% paddles into a 97% trigger.

(b) A sample holds N nuclei, each surviving to time t with probability $e^{-\lambda t}$ (Example 1.16), independently of the others. The probability that *none* has decayed by t is

$$\mathrm{P}\Big(\bigcap_{i=1}^{N} S_t^{(i)}\Big) \overset{(1)}{=} \prod_{i=1}^{N} \mathrm{P}\big(S_t^{(i)}\big) = \big(e^{-\lambda t}\big)^N = e^{-N\lambda t}.$$

(1) Mutual independence of the N survival events.

"None has decayed by t" is the same event as "the first decay comes after t". So the waiting time until the first click of a Geiger counter has the survival law $e^{-N\lambda t}$ of a single nucleus with rate $N\lambda$: it is itself exponential, with rate $N\lambda$, the activity of the sample. This is the seed of the Poisson process of Chapter 2.

Example 1.23 (Blue eyes). Each of three children has blue eyes with probability $\frac{1}{4}$, independently [Wasserman p. 15, Ex. 15]. Let K be the number of blue-eyed children. *(a)* Given that at

least one has blue eyes, what is the probability that at least two do? The event $\{K \geq 2\}$ lies inside $\{K \geq 1\}$, so

$$\mathrm{P}(K \geq 2 \mid K \geq 1) \overset{(1)}{=} \frac{\mathrm{P}(K \geq 2)}{\mathrm{P}(K \geq 1)} \overset{(2)}{=} \frac{3 \cdot \frac{1}{16} \cdot \frac{3}{4} + \frac{1}{64}}{1 - \left(\frac{3}{4}\right)^3} = \frac{10/64}{37/64} = \frac{10}{37} \approx 0.27.$$

(1) Definition of conditional probability, with $\{K \geq 2\} \cap \{K \geq 1\} = \{K \geq 2\}$.
(2) Numerator: exactly two blue (three choices of the brown-eyed child) or all three. Denominator: complement of "none blue". Independence gives each pattern's probability as a product.

(b) Given that the *youngest* has blue eyes, at least two are blue-eyed exactly when at least one of the other two is. By independence the youngest tells us nothing about them, so the answer is $1 - (\frac{3}{4})^2 = \frac{7}{16} \approx 0.44$. The two conditions sound alike but carry different information, and (b) is the larger because it pins down *which* child is blue-eyed.

1.6.3 Pairwise is not mutual

For three events, one might hope that the triple product $\mathrm{P}(ABC) = \mathrm{P}(A)\,\mathrm{P}(B)\,\mathrm{P}(C)$ is enough, or that it is enough for each pair to be independent. Neither is [CasellaBerger Exs. 1.3.10–1.3.11].

Example 1.24 (The triple product without the pairs). Roll two dice. Let A be "doubles", B "the sum is between 7 and 10" and C "the sum is 2, 7 or 8". Counting among the 36 outcomes, $\mathrm{P}(A) = \frac{6}{36} = \frac{1}{6}$, $\mathrm{P}(B) = \frac{6+5+4+3}{36} = \frac{1}{2}$ and $\mathrm{P}(C) = \frac{1+6+5}{36} = \frac{1}{3}$. The only double with sum in $\{7, 8\}$ is $(4, 4)$, so $\mathrm{P}(ABC) = \frac{1}{36} = \frac{1}{6} \cdot \frac{1}{2} \cdot \frac{1}{3}$. But $\mathrm{P}(BC) = \mathrm{P}(\text{sum 7 or 8}) = \frac{11}{36} \neq \frac{1}{6} = \mathrm{P}(B)\,\mathrm{P}(C)$, and $\mathrm{P}(AB) = \mathrm{P}(\{(4, 4), (5, 5)\}) = \frac{1}{18} \neq \frac{1}{12}$. The triple product holds by coincidence, while the pairs are dependent.

Example 1.25 (The pairs without the triple). Let Ω consist of the nine strings aaa, bbb, ccc, abc, bca, cab, acb, bac, cba, each with probability $\frac{1}{9}$, and let A_i be "the i-th letter is a". Each A_i contains three strings (for A_1: aaa, abc, acb), so $\mathrm{P}(A_i) = \frac{1}{3}$. Any two positions are both a only in aaa, so $\mathrm{P}(A_iA_j) = \frac{1}{9} = \mathrm{P}(A_i)\,\mathrm{P}(A_j)$: every pair is independent. But all three are a also only in aaa, so $\mathrm{P}(A_1A_2A_3) = \frac{1}{9} \neq \frac{1}{27}$. Knowing that the first two letters are a makes the third certain to be a.

Try it: Conditionally independent, marginally dependent. Suppose that on 95% of days the telescope's shared high-voltage supply is fine and each paddle fires with probability 0.92, independently, while on 5% of days the voltage sags and each fires with probability 0.5, independently. Show, with the law of total probability, that $\mathrm{P}(\text{top}) = \mathrm{P}(\text{middle}) \approx 0.899$ but $\mathrm{P}(\text{top and middle}) \approx 0.817 \neq 0.899^2 \approx 0.808$. The paddles are independent *given* the state of the supply, but not independent overall. The unobserved common cause creates the dependence. This is the structure of the hierarchical models of Chapter 5.

1.6.4 Testing independence on simulated data

Take the die events $A = \{2, 4, 6\}$, $B = \{1, 2, 3, 4\}$ (independent) and $C = \{4, 5, 6\}$ (dependent), simulate rolls, and compare $\hat{\mathrm{P}}(AB)$ with $\hat{\mathrm{P}}(A)\hat{\mathrm{P}}(B)$ [Wasserman p. 17, Ex. 23].

Listing 1.6: Independence checked by simulation (`code/ch01/06_independence_dice.py`)

```
rng = rng_for("ch01", "1a_independence")
N = 10_000
roll = rng.integers(1, 7, size=N)
```

```
A = np.isin(roll, [2, 4, 6])
B = np.isin(roll, [1, 2, 3, 4])
C = np.isin(roll, [4, 5, 6])

fA, fB, fC = A.mean(), B.mean(), C.mean()
fAB, fAC = (A & B).mean(), (A & C).mean()

# repeat the whole N-roll experiment many times to see the scatter of the test
reps = 4000
rolls = rng.integers(1, 7, size=(reps, N), dtype=np.int8)
Ar, Br, Cr = np.isin(rolls, [2, 4, 6]), np.isin(rolls, [1, 2, 3, 4]), np.isin(rolls, [4, 5,
    6])
diff_AB = (Ar & Br).mean(1) - Ar.mean(1) * Br.mean(1)
diff_AC = (Ar & Cr).mean(1) - Ar.mean(1) * Cr.mean(1)
```

Lines 21–24 roll one die N times and build the three masks with `np.isin`. Lines 26–27 are the frequencies. Lines 30–34 repeat the whole N-roll experiment 4000 times, as rows of a two-dimensional array, and compute for each repeat the difference $\hat{\mathrm{P}}(\text{both}) - \hat{\mathrm{P}}(\text{first})\hat{\mathrm{P}}(\text{second})$.

In a single run of $N = 10000$ rolls, $\hat{\mathrm{P}}(A) = 0.4964$, $\hat{\mathrm{P}}(B) = 0.6706$, $\hat{\mathrm{P}}(C) = 0.4927$. For the independent pair, $\hat{\mathrm{P}}(AB) = 0.3317$ against $\hat{\mathrm{P}}(A)\hat{\mathrm{P}}(B) = 0.3329$. For the dependent pair, $\hat{\mathrm{P}}(AC) = 0.3280$ against $\hat{\mathrm{P}}(A)\hat{\mathrm{P}}(C) = 0.2446$. Even the independent pair does not satisfy the product rule *exactly*, and no finite simulation would, because frequencies scatter. So the real question is whether the difference is larger than that random scatter. The repeats answer it, by showing how large the difference typically is.

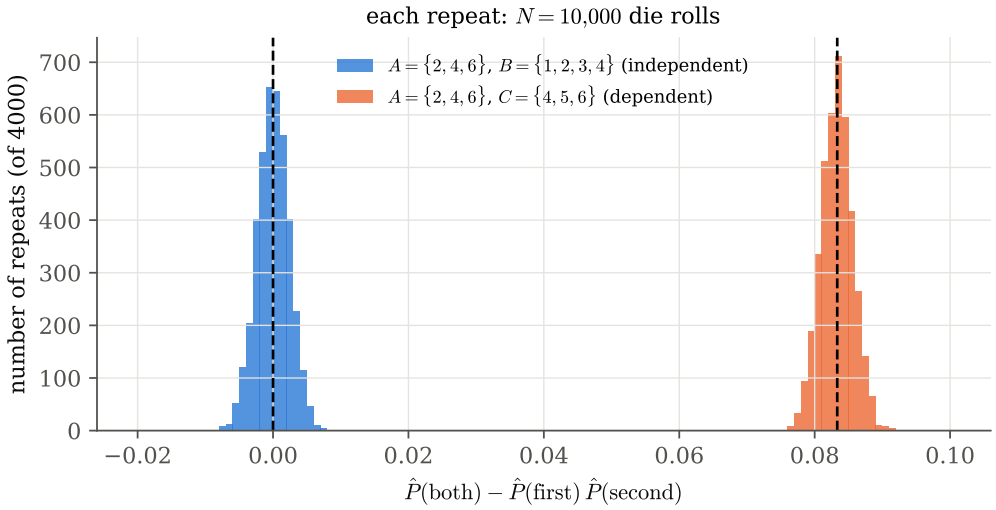


Figure 1.19: The difference between the frequency of "both" and the product of the two frequencies, over many repeats of a 10^4-roll experiment. For the independent pair (blue) the differences scatter symmetrically around zero. For the dependent pair (orange) they scatter around $\frac{1}{3} - \frac{1}{4} = \frac{1}{12}$, far from zero compared with the width of either histogram.

For the independent pair the differences have mean near zero and standard deviation 0.0024. For the dependent pair their mean is 0.0833, the exact $\frac{1}{12}$, some thirty-five standard deviations away. A difference of $0.3317 - 0.3329$ in the single run is typical of independence. A difference of 0.08 is not.

What this bought us. With data, "is $\mathrm{P}(AB) = \mathrm{P}(A)\,\mathrm{P}(B)$?" becomes "is the observed difference large compared with its scatter under independence?". That is a hypothesis test with independence as the null hypothesis, and Chapter 4 makes it precise.

In research: independence is what lets us multiply. Every likelihood in this book is a product, and every product is an assumption of independence. Poisson counts in different bins of a histogram are independent if the events arrive independently (colliders, photon counting). The spherical-harmonic coefficients $a_{\ell m}$ of a statistically isotropic Gaussian sky are independent for different (ℓ, m) (Chapter 7), which is why the full-sky CMB likelihood factorises over ℓ. For stationary detector noise, the Fourier amplitudes at different frequencies are independent, which is why the gravitational-wave log-likelihood is a sum over frequency bins weighted by the power spectral density (Chapter 10). A mask breaks the $a_{\ell m}$ independence and correlated noise breaks the frequency independence. Handling that loss is much of the work of Chapter 8.

Watch out: Three ways to get independence wrong

(1) Treating disjoint events as independent. (2) Checking only pairs, or only the full product, for three or more events. (3) Assuming independence where a hidden common cause exists (a shared power supply, a common calibration constant, a shared beam condition). Common-mode effects are the usual reason why a redundant system fails more often than the product of the individual failure rates predicts.

1.7 Bayes' theorem: reading the tree backwards

A tree can be read in two directions, and the analyst at a collider needs the direction it was not drawn in. Recall the two-level trigger of Section 1.5: a fast hardware trigger (L1) followed by a software high-level trigger (HLT), which together write about one collision in a hundred to disk. Its tree runs in the order in which things happen. First nature makes the collision signal (S) or background (B). Then L1 decides. Then the HLT decides. Read in that order, the tree answers *prediction* questions: if a collision is signal, how likely is it to be kept?

The analyst who opens the files on disk never sees nature's first choice. She sees only the collisions that were kept. So she asks the opposite question: of the collisions written to disk, what fraction is signal? That is an *inference* question. It starts from an observation and asks which branch of the tree produced it. These are the two directions of probability of Chapter 0. For a hypothesis H and an observed event E, $\mathrm{P}(E \mid H)$ runs forward, from H to what it produces. $\mathrm{P}(H \mid E)$ runs backward, from what we saw to what produced it.

Bayes' theorem turns the first kind of number into the second. We need it because the two kinds come from different places. The forward numbers are what physics and calibration give us: efficiencies, misidentification rates, branching fractions, the probability of the data under a model. The backward numbers are what we want: the purity of a sample, the probability that a track is a kaon, the credibility of a cosmological model. So the question is: given a tree built in the forward direction, how do we read it backwards?

The analyst's question can be said in words, in four steps, before any formula.

1. What did she see? A collision on disk, one that the trigger kept.
2. What could have put it there? It was either a signal collision or a background collision.
3. If it was signal, how probable was it to be kept? And if it was background?
4. Given those two answers, how must the shares of signal and background that nature started with be revised?

Step 3 is ordinary prediction, done once for each hypothesis. Only step 4 is new, and it needs nothing more than the definition of conditional probability, used once.

1.7.1 Two routes to the same leaf

The whole theorem rests on one fact: a leaf of a tree is the same event whichever order the tree was drawn in. Take a hypothesis H and an observed event E. The leaf "H and E", written HE for $H \cap E$, can be reached in two orders. In the forward tree we pass first through H and then through E. In a tree drawn the other way round we pass first through E and then through H [Lambert §3.6.1]. Both routes end at the same event HE, so they give it the same probability. The multiplication rule (1.18), $\mathrm{P}(HE) = \mathrm{P}(E \mid H)\,\mathrm{P}(H)$, computes it along the forward route. Dividing by $\mathrm{P}(E)$, which we assume is positive, gives the backward number:

$$\mathrm{P}(H \mid E) \overset{(1)}{=} \frac{\mathrm{P}(HE)}{\mathrm{P}(E)} \overset{(2)}{=} \frac{\mathrm{P}(E \mid H)\ \mathrm{P}(H)}{\mathrm{P}(E)}. \tag{1.23}$$

(1) Definition (1.17) of conditional probability, with E as the conditioning event: E is the new, shrunk sample space.
(2) Multiplication rule along the forward route, $\mathrm{P}(HE) = \mathrm{P}(E \mid H)\,\mathrm{P}(H)$.

The numerator is one leaf of the forward tree. The denominator $\mathrm{P}(E)$ is the total probability of the observation, and it too comes from the forward tree. Suppose the possible hypotheses $H_1, \ldots, H_k$ form a partition of the sample space Ω [exactly one of them is true]. Then the law of total probability (1.21) gives $\mathrm{P}(E)$ as the sum of all the forward leaves that end in E. Putting that sum in the denominator gives the general form [Wasserman p. 12, Thm 1.17].

Key idea: Bayes' theorem

Let $H_1, \ldots, H_k$ be a partition of Ω with every $\mathrm{P}(H_i) > 0$, and let $\mathrm{P}(E) > 0$. Then for each i,

$$\mathrm{P}(H_i \mid E) = \frac{\mathrm{P}(E \mid H_i)\ \mathrm{P}(H_i)}{\sum_{j=1}^{k} \mathrm{P}(E \mid H_j)\ \mathrm{P}(H_j)}. \tag{1.24}$$

In the tree: the posterior of H_i is the leaf "H_i then E" divided by the sum of all leaves that end in E.

Each ingredient has a name [Wasserman p. 12, Remark 1.18].

- $\mathrm{P}(H_i)$ is the **prior** probability of H_i: how probable H_i was before the observation.
- $\mathrm{P}(H_i \mid E)$ is its **posterior** probability: how probable it is after.
- $\mathrm{P}(E \mid H_i)$ is the **likelihood** of H_i: the probability of the observation we actually made, computed as if H_i were true.
- $\mathrm{P}(E)$ is the **evidence**. It is the same number for every H_i, and its only job is to make the posteriors add up to one.

Nothing here goes beyond the definition of conditional probability and the law of total probability. The only new thing is the direction in which we read the tree.

Inverting a tree means: compute the leaves forward, then regroup them by the observation instead of by the hypothesis. Figure 1.20 shows this for the trigger. The forward tree and the inverted tree have the same four leaves with the same probabilities, because the leaves are the same events. What changes is the order of the two splits, and with it the conditional probabilities written on the branches.

Example: inverting the trigger tree. *What she saw.* A collision on disk: it passed both L1 and the HLT. Call this one merged decision "accepted", Acc = L1 ∩ HLT.

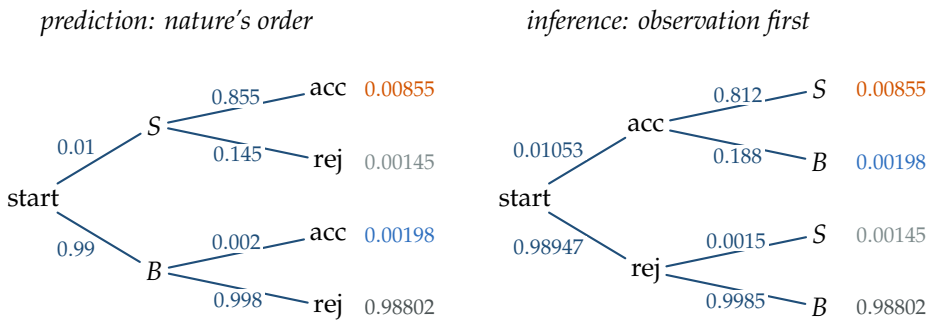


Figure 1.20: The trigger of Figure 1.16 with the two trigger levels merged into one decision, accepted (acc) or rejected (rej), drawn in both orders. Left: nature chooses signal or background, then the trigger decides. Right: the same four leaves, same colours, same probabilities, regrouped by the trigger decision first. The branches on the right are the inference numbers: among accepted collisions, 81% are signal.

What could have produced it? The collision was either signal, S, or background, B. These are the two hypotheses.

How probable was acceptance under each? Multiply along the paths of Figure 1.16. A signal collision passes L1 with probability 0.95 and then the HLT with probability 0.90, so $\mathrm{P}(\mathrm{Acc}\,|\,S) = 0.95 \times 0.90 = 0.855$. A background collision passes L1 with probability 0.02 and then the HLT with probability 0.10, so $\mathrm{P}(\mathrm{Acc}\,|\,B) = 0.02 \times 0.10 = 0.002$.

Then the priors, and normalise. Before the trigger, one collision in a hundred is signal: $\mathrm{P}(S) = 0.01$, $\mathrm{P}(B) = 0.99$. The signal fraction of the recorded sample, called its **purity**, is

$$\mathrm{P}(S\,|\,\mathrm{Acc}) \overset{(1)}{=} \frac{\mathrm{P}(\mathrm{Acc}\,|\,S)\,\mathrm{P}(S)}{\mathrm{P}(\mathrm{Acc}\,|\,S)\,\mathrm{P}(S) + \mathrm{P}(\mathrm{Acc}\,|\,B)\,\mathrm{P}(B)} \overset{(2)}{=} \frac{0.855 \times 0.01}{0.855 \times 0.01 + 0.002 \times 0.99}$$
$$= \frac{0.00855}{0.00855 + 0.00198} = \frac{0.00855}{0.01053} \approx 0.812.$$

(1) Bayes' theorem (1.24) with the partition $\{S, B\}$.
(2) The branch probabilities of the merged tree.

Only one collision in a hundred is signal, yet four out of five recorded collisions are. The trigger has raised the probability of "signal" from a prior of 0.01 to a posterior of 0.81. The reason is the ratio of the two likelihoods: a signal collision is $0.855/0.002 \approx 430$ times more likely to be accepted than a background one. The simulation of Section 1.5 (the last line of its listing) counted the signal fraction among the accepted collisions directly and found 0.8138, against the exact 0.8120.

Example: dots and dashes. Morse code sends dots and dashes in the proportion 3 : 4, so $\mathrm{P}(\text{dot sent}) = \frac{3}{7}$ [CasellaBerger Ex. 1.3.6]. Noise on the line turns a dot into a dash, or a dash into a dot, with probability $\frac{1}{8}$. We receive a dot. How sure can we be that a dot was sent?

What we saw is "dot received". Two things could have produced it: a dot was sent and arrived intact, or a dash was sent and noise flipped it. Under the first, a received dot has probability $\frac{7}{8}$; under the second, $\frac{1}{8}$. The priors are $\frac{3}{7}$ for a dot and $\frac{4}{7}$ for a dash. Then

$$\mathrm{P}(\text{dot sent}\,|\,\text{dot rec.}) \overset{(1)}{=} \frac{\frac{7}{8}\cdot\frac{3}{7}}{\frac{7}{8}\cdot\frac{3}{7} + \frac{1}{8}\cdot\frac{4}{7}} \overset{(2)}{=} \frac{21/56}{21/56 + 4/56} = \frac{21}{25} = 0.84.$$

(1) Bayes' theorem: the likelihood of "dot received" is $\frac{7}{8}$ if a dot was sent and $\frac{1}{8}$ if a dash was sent.

(2) Put both products over the common denominator 56.

So a received dot is wrong 4 times in 25, about one time in six, even though each symbol is flipped only one time in eight. The extra errors come from the dashes. Dashes are sent more often than dots, so the flipped dashes make up a larger share of the received dots than the flip rate alone suggests.

1.7.2 The recipe: prior times likelihood, then normalise

The evidence $\mathrm{P}(E)$ is the same for every hypothesis. So we can leave it out, compute prior times likelihood for each hypothesis, and divide by the total at the end:

$$\mathrm{P}(H_i \mid E) \;\propto\; \mathrm{P}(E \mid H_i)\,\mathrm{P}(H_i), \qquad \text{then divide by the sum over } i. \tag{1.25}$$

This is the most useful form in practice. It turns every discrete Bayes problem into a table with one row per hypothesis and four columns: the prior; the likelihood; their product, which is the forward leaf that ends in E; and the product divided by the sum of its column, which is the posterior. The sum of the product column is $\mathrm{P}(E)$.

Example: spam. An email is spam (A_1), low priority (A_2) or high priority (A_3), with prior probabilities 0.7, 0.2 and 0.1 [Wasserman pp. 12–13, Ex. 1.19]. The word "free" (E) appears in 90% of spam and in 1% of each of the other two kinds.[1] An email arrives containing "free". That is what we saw, and the three kinds of email are the hypotheses that could have produced it. The table is

hypothesis	prior $\mathrm{P}(A_i)$	likelihood $\mathrm{P}(E \mid A_i)$	product	posterior $\mathrm{P}(A_i \mid E)$
spam	0.7	0.90	0.630	0.9953
low priority	0.2	0.01	0.002	0.0032
high priority	0.1	0.01	0.001	0.0016
sum	1	0.92	0.633	1

so $\mathrm{P}(A_1 \mid E) = 0.630/0.633 \approx 0.995$.

The column sums carry a lesson: the likelihoods are not a probability distribution over the hypotheses. The priors add up to one, and so do the posteriors, because both are probabilities *over the hypotheses*. The likelihoods add up to 0.92, and nothing forces them to add up to one. Each likelihood is a probability over *outcomes* (does "free" appear or not?), computed in a different world: one where the email is spam, one where it is low priority, one where it is high priority. Read across those worlds, as a function of the hypothesis, the list of likelihoods is not a probability distribution. This is the statement "$\mathrm{P}(A \mid \cdot)$ is not a probability" of Section 1.4. It is also why, in Chapter 3, the likelihood of a parameter is called a function and not a density.

Example: which operating system? Among computer owners 30% use a Mac, 50% Windows and 20% Linux, and the fractions infected by a virus are 0.65, 0.82 and 0.50 [Wasserman p. 16, Ex. 19]. A randomly chosen owner turns out to be infected. The products are $0.3 \times 0.65 = 0.195$, $0.5 \times 0.82 = 0.410$ and $0.2 \times 0.50 = 0.100$, with sum $\mathrm{P}(\text{infected}) = 0.705$, so

$$\mathrm{P}(\text{Windows} \mid \text{infected}) = \frac{0.410}{0.705} \approx 0.582, \qquad \mathrm{P}(\text{Mac} \mid \cdot) \approx 0.277, \qquad \mathrm{P}(\text{Linux} \mid \cdot) \approx 0.142.$$

[1] Wasserman prints $\mathrm{P}(B \mid A_1) = .01$ for the third likelihood. The subscript should be A_3, as the computation that follows it shows.

Windows rises from 0.50 to 0.58. Mac falls from 0.30 to 0.28, and Linux from 0.20 to 0.14. The rule is visible in the numbers. The average infection rate is P(infected) = 0.705. Windows owners are infected more often than that (0.82), so Windows gains. Mac (0.65) and Linux (0.50) owners are infected less often, so they lose.

That rule is general: probability can only move between the hypotheses, never be created or destroyed [Wasserman p. 16, Ex. 18]. The reason is that the priors and the posteriors both add up to one. So if one hypothesis loses, another must gain. To see it in symbols, suppose A_1 loses, $\mathrm{P}(A_1 \mid E) < \mathrm{P}(A_1)$. Then

$$\sum_{i=2}^{k} \mathrm{P}(A_i \mid E) \overset{(1)}{=} 1 - \mathrm{P}(A_1 \mid E) \overset{(2)}{>} 1 - \mathrm{P}(A_1) \overset{(3)}{=} \sum_{i=2}^{k} \mathrm{P}(A_i),$$

(1) $\mathrm{P}(\cdot \mid E)$ is a probability (Section 1.4), and the A_i partition Ω, so the posteriors add up to one.
(2) The assumption $\mathrm{P}(A_1 \mid E) < \mathrm{P}(A_1)$.
(3) The priors add up to one.

A sum of $k-1$ terms can be larger than another sum of $k-1$ terms only if at least one of its terms is larger. So some other hypothesis gains: $\mathrm{P}(A_i \mid E) > \mathrm{P}(A_i)$ for at least one $i \geq 2$. For exactly this reason Kruschke calls Bayesian inference a *reallocation of credibility* across the possibilities [Kruschke §2.1].

The extreme case is a hypothesis whose likelihood is zero. The observation is impossible under it, so its posterior is zero. Its prior weight goes to the hypotheses that survive, shared in proportion to each one's prior times likelihood. Repeat this with more observations until only one hypothesis is left, and you get Sherlock Holmes's rule: once every other cause has been ruled out, the one that remains carries all the probability.

1.7.3 Base rates and the odds form

Bayes' theorem is least intuitive when the prior is small. A test for a rare disease detects it with probability $\mathrm{P}(+ \mid D) = 0.98$, and gives a false positive with probability $\mathrm{P}(+ \mid D^c) = 0.03$. The prevalence is $\mathrm{P}(D) = 0.001$ [Cowan p. 4]. You test positive.

What happened. The test came back positive, written +. *What could have produced it?* Two situations: you have the disease, D, or you do not, D^c. A positive result can come out of either. *How probable was a positive under each?* With the disease it is the detection probability, 0.98; without it, the false-positive rate, 0.03. *Then the base rates.* Before the test, D had probability 0.001 and D^c had 0.999. The two forward leaves that end in "+" give

$$\mathrm{P}(D \mid +) \overset{(1)}{=} \frac{0.98 \times 0.001}{0.98 \times 0.001 + 0.03 \times 0.999} \overset{(2)}{=} \frac{0.00098}{0.00098 + 0.02997} \approx 0.032.$$

(1) Bayes' theorem (1.24) with the partition $\{D, D^c\}$.
(2) The two forward leaves that end in "+".

The test is wrong only 2% or 3% of the time, yet after a positive result the probability of disease is only about 3%. The two leaves show why. The false-positive leaf, 0.02997, is thirty times larger than the true-positive leaf, 0.00098. Healthy people rarely test positive, but there are a thousand times more of them, so they produce most of the positives. The medical test of Example 1.15 ($\mathrm{P}(D \mid +) \approx 0.08$) and the screening example of [Lambert §3.6.2] (also 0.08) have the same structure. Ignoring the prior in such a problem is called the **base-rate fallacy**. It is the prosecutor's fallacy of Section 1.4 in another guise: it confuses $\mathrm{P}(+ \mid D)$ with $\mathrm{P}(D \mid +)$.

With two hypotheses, the effect of an observation is a single multiplication. To see it, write Bayes' theorem (1.23) for H_1 and for H_2, and divide one by the other. The evidence $\mathrm{P}(E)$ cancels:

$$\frac{\mathrm{P}(H_1 \mid E)}{\mathrm{P}(H_2 \mid E)} \overset{(1)}{=} \frac{\mathrm{P}(E \mid H_1)\,\mathrm{P}(H_1)/\mathrm{P}(E)}{\mathrm{P}(E \mid H_2)\,\mathrm{P}(H_2)/\mathrm{P}(E)} \overset{(2)}{=} \underbrace{\frac{\mathrm{P}(E \mid H_1)}{\mathrm{P}(E \mid H_2)}}_{\text{likelihood ratio}} \times \underbrace{\frac{\mathrm{P}(H_1)}{\mathrm{P}(H_2)}}_{\text{prior odds}} . \tag{1.26}$$

(1) Bayes' theorem (1.23) in the numerator and in the denominator.
(2) Cancel $\mathrm{P}(E)$ and regroup.

The ratio on the left is the **posterior odds**. The observation multiplies the prior odds by the **likelihood ratio**, and it does nothing else. For the rare disease, take $H_1 = D$ and $H_2 = D^c$. The likelihood ratio is $0.98/0.03 \approx 32.7$ and the prior odds are $0.001/0.999 \approx 0.0010$, so the posterior odds are 0.0327. Turning odds back into a probability, $\mathrm{P}(D \mid +) = 0.0327/(1 + 0.0327) \approx 0.032$, as before. The test is good evidence (it multiplies the odds by about thirty), but it starts from odds of one to a thousand.

The odds form makes repeated observations easy: each new observation multiplies the odds again. Suppose a second test is also positive, and the two results are independent *given the disease status* (Section 1.6). Then yesterday's posterior is today's prior, and the odds are multiplied by 32.7 again, to 1.07, so $\mathrm{P}(D \mid ++) \approx 0.52$. A third positive test gives odds 34.9 and $\mathrm{P}(D \mid +++) \approx 0.97$. In logarithms the multiplications become additions: each independent observation *adds* its log likelihood ratio to the log odds. This is why analyses add up log likelihoods rather than multiply likelihoods. The condition "independent given the disease status" matters. A test that fails for a reason specific to the patient will fail again, and then repeating it adds much less [Lambert Problem 3.8.5].

Watch out: Efficiency is not purity

A detector's efficiency $\mathrm{P}(\text{tag} \mid \text{true})$ is a forward number, measured once on a calibration sample, and it does not depend on what else is in the beam. The purity $\mathrm{P}(\text{true} \mid \text{tag})$ is an inverse number, and it depends strongly on the prior, the composition of the sample. A selection with 90% efficiency and a 5% fake rate can give a sample that is mostly fakes, if the true particles are rare. Quote both, and never call one by the other's name.

Example: particle identification with misidentification. A hadron beam contains pions, kaons and protons in the proportions 0.80 : 0.15 : 0.05. A Cherenkov detector tags a track as "kaon" with probability 0.90 for a real kaon, 0.05 for a pion and 0.10 for a proton. A track is tagged as a kaon. Which particle is it?

What we saw is the tag "kaon". Any of the three species could have produced it, so they are the three hypotheses. Their priors are the beam proportions, and their likelihoods are the three tagging probabilities. The table:

species	prior	$\mathrm{P}(\text{tag } K \mid \text{species})$	product	posterior
π	0.80	0.05	0.040	0.2222
K	0.15	0.90	0.135	0.7500
p	0.05	0.10	0.005	0.0278
sum	1		0.1800	1

One tagged track in four is not a kaon, and almost all of the impostors are pions. Pions are misidentified only 5% of the time, but there are more than five times as many pions as kaons in the beam. So the purity depends on the beam, not only on the detector. As the kaon fraction

falls, the purity collapses (Figure 1.21). At a kaon fraction of 1% it is only 0.148, with the same detector and the same 90% efficiency.

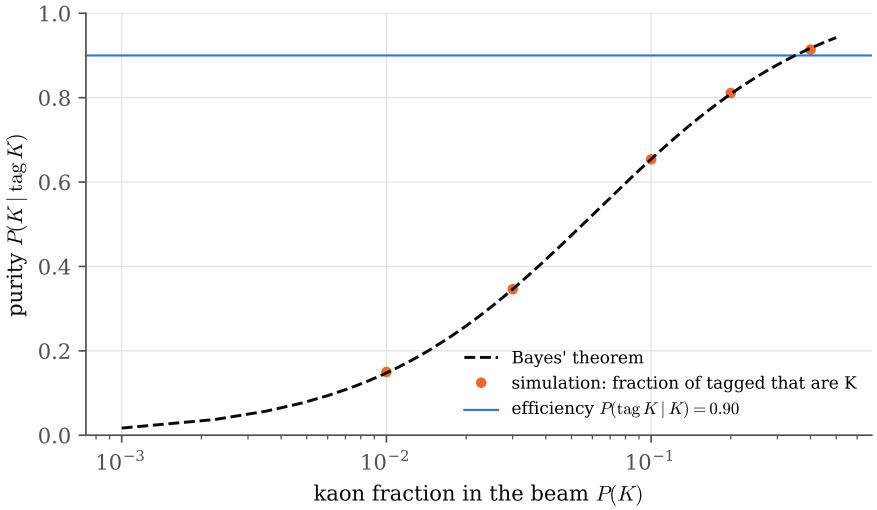


Figure 1.21: Purity of the kaon-tagged sample against the kaon fraction of the beam, on a logarithmic axis. The dashed curve is Bayes' theorem, the dots are simulated beams in which we tag tracks and count the true kaons among the tagged ones. The horizontal line is the efficiency, which stays at 0.90 whatever the beam. The two agree only when kaons dominate.

1.7.4 When the observation carries information in how it was produced

When the observation comes out of a person's choice, intuition tends to ignore how it was produced, and goes wrong. In the trigger, the detector and the medical test, the observation came from a fixed mechanism that was easy to see. In three famous puzzles, Monty Hall, the three prisoners and the three cards, the mechanism is hidden inside someone's decision. The four steps of the reasoning handle them without any new idea. Everything turns on the third step: under each hypothesis, how probable was *exactly the observation we made*?

Example: Monty Hall. A car is behind one of three doors, each with probability $\frac{1}{3}$, and goats are behind the other two. We pick door 1. The host, Monty, knows where the car is. He always opens one of the two doors we did not pick, never the car door, and if he has a choice he picks at random. He opens door 3 [Wasserman p. 14, Ex. 10]. Should we switch to door 2?

What we saw. Monty opened door 3. Call this event M_3.

What could have made him open it? Monty's choice depends on one thing we do not know, where the car is. So the candidate explanations are the three places the car can be: H_1, H_2 and H_3, the car behind door 1, 2 or 3. We want to know which of these places could have led Monty to open door 3, and how strongly.

How likely was door 3 under each location? Take the three in turn. Monty knows where the car is, he never reveals it, and he never opens our door.

- H_1, the car is behind our door 1: both other doors hide goats, so Monty can open door 2 or door 3 and chooses at random, $\mathrm{P}(M_3 \mid H_1) = \frac{1}{2}$.
- H_2, the car is behind door 2: he may not open our door 1 and may not reveal the car behind door 2, so he is forced to open door 3, $\mathrm{P}(M_3 \mid H_2) = 1$.

- H_3, the car is behind door 3: opening door 3 would reveal the car, so he cannot open it, $\mathrm{P}(M_3 \mid H_3) = 0$.

These are the second-stage branches of the tree in Figure 1.23. *Then the priors, and normalise.* Each location had prior $\mathrm{P}(H_d) = \frac{1}{3}$, so the three leaves that end in M_3 weigh $\frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6}$, $\frac{1}{3} \cdot 1 = \frac{1}{3}$ and $\frac{1}{3} \cdot 0 = 0$. Normalising by their sum,

$$\mathrm{P}(H_2 \mid M_3) \overset{(1)}{=} \frac{\mathrm{P}(M_3 \mid H_2)\,\mathrm{P}(H_2)}{\sum_{d=1}^{3} \mathrm{P}(M_3 \mid H_d)\,\mathrm{P}(H_d)} \overset{(2)}{=} \frac{1 \cdot \frac{1}{3}}{\frac{1}{2} \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3}} = \frac{1/3}{1/2} = \frac{2}{3},$$

(1) Bayes' theorem with the partition $\{H_1, H_2, H_3\}$.
(2) The three likelihoods above and the uniform prior. The evidence is $\mathrm{P}(M_3) = \frac{1}{6} + \frac{1}{3} = \frac{1}{2}$.

and in the same way $\mathrm{P}(H_1 \mid M_3) = \frac{1}{6} / \frac{1}{2} = \frac{1}{3}$ and $\mathrm{P}(H_3 \mid M_3) = 0$. Switching wins with probability $\frac{2}{3}$.

The whole calculation is one look at the picture of Figure 1.11, drawn to scale in Figure 1.22. The three places the car can be fill the sample space in three equal strips. The event M_3 is a box that takes half of the H_1 strip, all of the H_2 strip and none of the H_3 strip, exactly the three likelihoods above. Once we see M_3, that box is the whole world, and the H_2 piece of it is twice the H_1 piece.

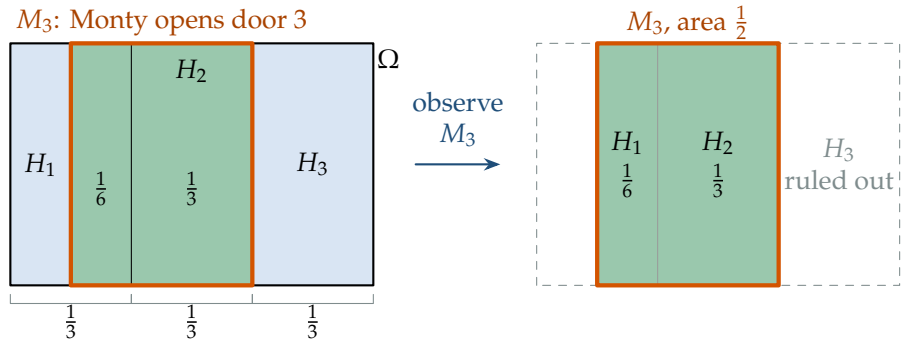


Figure 1.22: Monty Hall drawn to scale in the picture of Figure 1.11. Left: the rectangle Ω has area 1, and since the car is behind exactly one door, the three hypotheses H_1, H_2, H_3 fill it in three strips of area $\frac{1}{3}$. The event M_3, "Monty opens door 3" (orange outline), covers half of H_1 (with the car behind our door he picks door 2 or 3 at random), all of H_2 (he is forced to open door 3) and none of H_3 (he never reveals the car), so its area is $\frac{1}{6} + \frac{1}{3} = \frac{1}{2}$. Right: seeing M_3 discards everything outside the orange box, and each hypothesis keeps the share of the box that its piece fills: $\frac{1}{3}$ for H_1, $\frac{2}{3}$ for H_2.

Most people's first answer is that, with one door gone, the two doors left must be even. That answer throws away what Monty's choice told us. Door 3 was not equally likely to be opened under the two locations still possible: with the car behind door 2, Monty had no other move, while with the car behind our door 1 he would have opened door 3 only half the time. What we saw is therefore twice as probable under H_2 as under H_1, and the posteriors inherit that factor of two, $\frac{2}{3}$ against $\frac{1}{3}$.

The fifty–fifty answer gets one step right and skips another. It shrinks the sample space correctly: door 3 is out. But it forgets the likelihoods. A counterexample shows that they decide the answer. Suppose Monty did *not* know where the car is, opened door 2 or door 3 at random, and happened to show a goat behind door 3. Under H_1 he opens door 3 with probability $\frac{1}{2}$ and it hides a goat. Under H_2 he also opens door 3 with probability $\frac{1}{2}$, and it again hides a goat. So

$\mathrm{P}(M_3, \text{goat} \mid H_1) = \mathrm{P}(M_3, \text{goat} \mid H_2) = \frac{1}{2}$, and the answer really would be fifty–fifty. Same doors, same observation, different mechanism, different posterior.

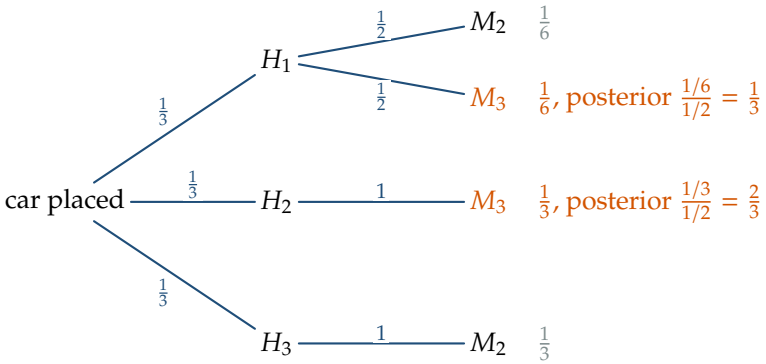


Figure 1.23: Monty Hall as a tree. We always pick door 1. The first split is where the car is. The second is which door Monty opens, which is random only when the car is behind our door. The two orange leaves are the ways of producing the observation "Monty opened door 3". They have probabilities $\frac{1}{6}$ and $\frac{1}{3}$, and dividing each by their total $\frac{1}{2}$ gives the posterior. The car cannot be behind door 3, so that hypothesis has no leaf in the observation.

Example: the three prisoners. The same tree appears as an older puzzle [CasellaBerger Ex. 1.3.4]. One of three prisoners A, B, C is to be pardoned, each with probability $\frac{1}{3}$. Prisoner A asks the warden to name one of the *other two* who will be executed.

What happened. The warden names B. Call this event W.

What could have made him name B? The one thing A does not know is who is pardoned, so the hypotheses are "A pardoned", "B pardoned" and "C pardoned", written simply A, B, C. Take them in turn.

- If A is pardoned, both B and C will be executed. The warden can name either and picks at random, so $\mathrm{P}(W \mid A) = \frac{1}{2}$.
- If B is pardoned, the warden cannot name B as one to be executed, so $\mathrm{P}(W \mid B) = 0$.
- If C is pardoned, then of B and C only B will be executed, so the warden must name B: $\mathrm{P}(W \mid C) = 1$.

Then the priors, and normalise. These are the Monty Hall likelihoods, with prisoner A in the place of door 1, and the priors are again $\frac{1}{3}$ each. So $\mathrm{P}(A \mid W) = \frac{1}{3}$ and $\mathrm{P}(C \mid W) = \frac{2}{3}$.

Prisoner A reasons "now it is A or C, so my chance has risen to $\frac{1}{2}$". He is wrong. The warden is right that he told A nothing about A's own fate. But he told A a great deal about C's.

Example: three cards. One card is green on both sides, one red on both sides, one green on one side and red on the other. We draw a card at random, look at one side at random, and see green [Wasserman p. 15, Ex. 12]. What is the probability that the other side is green too?

What we saw is a green face. It could have come from any card that has a green side, so the hypotheses are the three cards: GG, RR and GR, with prior $\frac{1}{3}$ each. Take them in turn. The GG card shows green whichever side we look at, so the likelihood is 1. The RR card never shows green: 0. The GR card shows green only if we happen to look at its green side: $\frac{1}{2}$. The posterior of GG is

$$\mathrm{P}(GG \mid \text{see } G) \overset{(1)}{=} \frac{1 \cdot \frac{1}{3}}{1 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + \frac{1}{2} \cdot \frac{1}{3}} \overset{(2)}{=} \frac{1/3}{1/2} = \frac{2}{3},$$

(1) Bayes' theorem with the partition $\{GG, RR, GR\}$.
(2) The evidence is $\mathrm{P}(\text{see } G) = \frac{1}{2}$, as it must be by symmetry between the colours.

and the hidden side is green exactly when the card is GG. The popular answer, $\frac{1}{2}$, counts cards: two cards have a green side, so it treats them as equally likely. But what we saw was a face, not a card. The correct answer counts green *faces*. There are three, and two of them are on the GG card.

1.7.5 A posterior over a parameter: coins in a box

The hypotheses can be the possible values of a number, not just named situations. Then the posterior is a probability distribution over that number. That is the step from the puzzles to parameter estimation.

Example: which coin? A bag holds a fair coin and a biased coin with $\mathrm{P}(\text{heads}) = 0.95$. We pick one at random and toss it three times, getting three heads, H_3 [SeeingTheory pp. 28–29]. Which coin did we pick? The hypotheses are "biased" and "fair", each with prior $\frac{1}{2}$. Given the coin, the tosses are independent, so the likelihoods of three heads are $0.95^3 = 0.857$ for the biased coin and $0.5^3 = 0.125$ for the fair one. In Example 1.19 we computed the two forward leaves, $\mathrm{P}(H_3 \mid \text{biased})\,\mathrm{P}(\text{biased}) = 0.857 \times 0.5$ and $\mathrm{P}(H_3 \mid \text{fair})\,\mathrm{P}(\text{fair}) = 0.125 \times 0.5$, and their sum $\mathrm{P}(H_3) \approx 0.491$. Reading that tree backwards,

$$\mathrm{P}(\text{biased} \mid H_3) = \frac{0.857 \times 0.5}{0.491} \approx 0.873.$$

Three heads move the biased coin from $\frac{1}{2}$ to 0.87. In the odds form (1.26): the prior odds were 1, the likelihood ratio is $0.857/0.125 \approx 6.9$, so the posterior odds are 6.9.

Example: five coins, and a prediction from a posterior. A box holds five coins whose probabilities of heads are $p_1 = 0$, $p_2 = \frac{1}{4}$, $p_3 = \frac{1}{2}$, $p_4 = \frac{3}{4}$ and $p_5 = 1$ [Wasserman p. 16, Ex. 20]. We pick coin C_i at random, prior $\frac{1}{5}$ each, toss it once and get heads (H_1). The likelihood of coin i is p_i, so

$$\mathrm{P}(C_i \mid H_1) \overset{(1)}{=} \frac{p_i \cdot \frac{1}{5}}{\sum_j p_j \cdot \frac{1}{5}} \overset{(2)}{=} \frac{p_i}{5/2} = \big(0,\ 0.1,\ 0.2,\ 0.3,\ 0.4\big) \quad \text{for } i = 1, \ldots, 5.$$

(1) Bayes' theorem. The uniform prior cancels between numerator and denominator.
(2) $\sum_j p_j = 0 + \frac{1}{4} + \frac{1}{2} + \frac{3}{4} + 1 = \frac{5}{2}$.

With a flat prior, the posterior is just the likelihood, rescaled to add up to one. The two-headed coin ($p_5 = 1$) is now the most probable, and the two-tailed coin ($p_1 = 0$) is ruled out.

Now toss the same coin again. What is the probability of another head, $\mathrm{P}(H_2 \mid H_1)$, where H_2 is "heads on the second toss"? We do not know which coin we hold, so we average over the coins. Given the coin C_i, the tosses are independent, so $\mathrm{P}(H_2 \mid C_i H_1) = p_i$. The conditional probability $\mathrm{P}(\cdot \mid H_1)$ is itself a probability (Section 1.4), so every rule holds for it, including the law of total probability. Hence

$$\mathrm{P}(H_2 \mid H_1) \overset{(1)}{=} \sum_{i=1}^{5} \mathrm{P}(H_2 \mid C_i H_1)\ \mathrm{P}(C_i \mid H_1) \overset{(2)}{=} \sum_{i=1}^{5} p_i\ \mathrm{P}(C_i \mid H_1)$$
$$\overset{(3)}{=} \frac{\sum_i p_i^2}{\sum_i p_i} = \frac{0 + \frac{1}{16} + \frac{1}{4} + \frac{9}{16} + 1}{5/2} = \frac{15/8}{5/2} = \frac{3}{4}.$$

(1) Law of total probability (1.21) for the probability $\mathrm{P}(\cdot \mid H_1)$, over the partition $\{C_i\}$. Each term is $\mathrm{P}(H_2 C_i \mid H_1) = \mathrm{P}(H_2 \mid C_i H_1)\,\mathrm{P}(C_i \mid H_1)$ by the multiplication rule applied inside $\mathrm{P}(\cdot \mid H_1)$.
(2) Given the coin, the second toss does not care about the first.
(3) The posterior from the previous chain.

The chance of heads has gone from $\frac{1}{2}$ to $\frac{3}{4}$. Before any toss, it was the prior average of p_i, which is $\frac{1}{2}$. After one head it is the *posterior* average of p_i, which is $\frac{3}{4}$. This is a **posterior predictive** probability: a prediction made by averaging over what we still do not know, weighted by what we now believe. So the two tosses are *not* independent: $\mathrm{P}(H_2 \mid H_1) = \frac{3}{4} \neq \mathrm{P}(H_2) = \frac{1}{2}$. They are independent only given the coin. The first head changes the forecast for the second because it tells us something about the coin.

Finally, suppose instead that we toss the chosen coin until the first head, and it arrives on toss 4 (B_4). The likelihood of coin i is three tails and then a head, $(1-p_i)^3 p_i$:

$$\big((1-p_i)^3 p_i\big)_{i=1}^{5} = \Big(0,\ \tfrac{27}{256},\ \tfrac{16}{256},\ \tfrac{3}{256},\ 0\Big),$$
$$\mathrm{P}(C_i \mid B_4) = \Big(0,\ \tfrac{27}{46},\ \tfrac{16}{46},\ \tfrac{3}{46},\ 0\Big) \approx (0,\ 0.59,\ 0.35,\ 0.07,\ 0),$$

where the second list is the first divided by its sum $\frac{46}{256}$ (the flat prior again cancels). A late first head favours a coin with a small p. Both extreme coins are eliminated: $p_1 = 0$ never gives a head, and $p_5 = 1$ gives one on the first toss.

The five values p_i are a coarse grid over a continuous parameter, the probability of heads $p \in [0,1]$. With a finer and finer grid, the posterior becomes a density $p(p \mid \boldsymbol{d})$, where $\boldsymbol{d}$ stands for the observed tosses, and the sum in the evidence becomes an integral. That is the subject of Chapter 5.

1.7.6 Bayes' theorem in a simulation: selecting on the observation

Bayes' theorem can be simulated in one line, because conditioning is selection (Section 1.4). The recipe has three steps.

1. Generate many trials in the forward order, exactly as the tree says: first the hypothesis, then the observation.
2. Keep only the trials in which the observation came out as it actually did.
3. Among those, count how often each hypothesis was the true one.

That fraction estimates the posterior, and no formula is involved. Here it is for the particle-identification beam of the kaon-tagging example.

Listing 1.7: Bayes' theorem by selecting tagged tracks (`code/ch01/07_particle_id_bayes.py`)

```
30 # --- simulation: conditioning = selecting the tagged tracks
31 rng = rng_for("ch01", "1a_particle_id")
32 N = 1_000_000
33 true = rng.choice(3, size=N, p=prior)
34 tagged = rng.random(N) < p_tag[true]
35 mc_post = np.bincount(true[tagged], minlength=3) / tagged.sum()
```

Line 33 is the first split of the tree, the true species of each of a million tracks, drawn from the beam composition (the prior). Line 34 is the second split, the detector's decision. `p_tag[true]` looks up, for each track, the tagging probability of its own species, so this line uses exactly the

likelihoods of the table. Line 35 is the inversion: `true[tagged]` keeps the true species of the tagged tracks only, and `np.bincount` counts how many are pions, kaons and protons. Of the 180515 tagged tracks, the fractions are 0.2228, 0.7492 and 0.0280, against the exact posteriors 0.2222, 0.7500 and 0.0278. The same selection, repeated for five beam compositions, gives the dots in Figure 1.21.

For Monty Hall the forward simulation must also contain Monty's rule, because his rule is the likelihood.

Listing 1.8: Monty Hall by selecting the games in which door 3 was opened (`code/ch01/08_monty_hall.py`)

```
26 prize = rng.integers(1, 4, size=N)            # door hiding the prize, uniform on {1,2,3}
27 coin = rng.random(N) < 0.5                     # Monty's tie-break when prize is behind
       door 1
28 # we always pick door 1. Monty opens a door that is neither ours nor the prize door.
29 opened = np.where(prize == 1, np.where(coin, 2, 3),   # free choice between 2 and 3
30                   np.where(prize == 2, 3, 2))         # forced choice
31 sel = opened == 3                               # the observation "Monty opened door 3"
32 post = np.array([np.mean(prize[sel] == d) for d in (1, 2, 3)])
```

Line 26 places the prize (the prior). Lines 29–30 are Monty's behaviour, which is the likelihood: with the prize behind door 1 he opens door 2 or 3 according to a fair coin, otherwise he opens the one empty door we did not pick. Line 31 selects the observation, "Monty opened door 3", which happened in 150332 of the 300000 games, a fraction 0.5011 (exact $\frac{1}{2}$, the evidence). Line 32 counts where the prize was in those games: door 1 in a fraction 0.3326, door 2 in 0.6674, door 3 in 0.0000. Over all games, switching won 0.6652 of the time and staying 0.3348 (Figure 1.24). The three-card problem is simulated the same way in the same script: of the 150117 draws that showed a green face, the hidden face was green in a fraction 0.6647 (exact $\frac{2}{3}$).

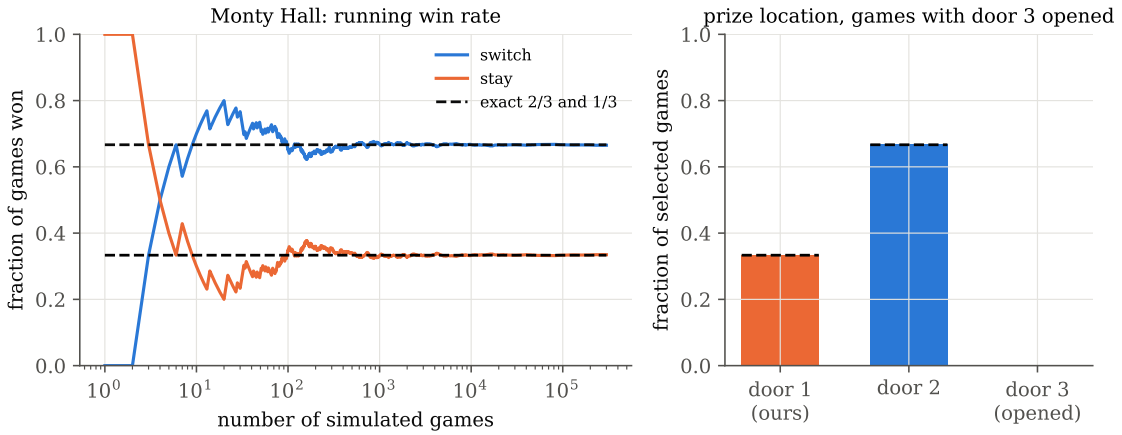


Figure 1.24: Monty Hall by simulation. Left: the running fraction of games won by always switching and by always staying, settling on $\frac{2}{3}$ and $\frac{1}{3}$. Right: among the games in which Monty opened door 3, where the prize was. The bars are the posterior, estimated by counting, and the dashed ticks are the exact values $\frac{1}{3}$ and $\frac{2}{3}$.

How this is used in real research: Selecting on the observation is how analyses estimate purity

Here the simulation only verified a posterior that Bayes' theorem gives in one line. In a real analysis the "hypotheses" are the physics processes (signal and every background), the "observation" is a long list of selection cuts, and the likelihood of passing them depends on the kinematics of each event, for which no formula exists. The expected composition of the selected sample is then computed exactly as above: simulate each process with its cross

section as the prior weight, apply the selection, and count what survives. The numbers that come out are the prior-times-likelihood column of the table, event by event, and the fractions among the survivors are the posterior. Bayes' theorem in its tabular form is also how particle-ID systems combine the responses of several detectors: the likelihood ratios of independent subdetectors multiply, as in (1.26), and the prior is the expected particle mix.

1.7.7 Where this goes

Every later chapter uses one of the two directions, prediction or inference, and many use both. Side by side:

prediction	inference
start from a hypothesis H	start from an observed event E
ask what outcomes it can produce	ask which hypotheses could have produced it
uses $\mathrm{P}(E \mid H)$, the forward tree	wants $\mathrm{P}(H \mid E)$, the tree read backwards
efficiency, sampling distribution, p-value	purity, posterior, credible interval

They are not rival interpretations of probability. They are two questions about one joint probability structure, the leaves of a single tree.

From the trigger to the five coins, every example was settled by two sentences said before any symbol. First: which hypotheses could have given what we saw? They are the rows of the table, and the posterior is spread over them. Second: how probable was the observation under each one? That is the likelihood column. Chapter 5 turns this into a standing habit.

The arithmetic is always the same, but the meaning of the posterior depends on the hypotheses. When they are outcomes of a repeatable process (signal or background, which coin, which door), a posterior is a long-run frequency among the selected trials, as the simulations showed. When the hypothesis is a value of a physical constant, which nature chose once, there are no repeated trials to count. Then the posterior can only be read as a degree of belief, the second interpretation of Section 1.2.

In the pipeline that runs from a physical model, through the data and a summary of them, to physics inferred, the forward tree is the sampling distribution $P(S \mid \theta)$ of a summary S computed from the data, given the model parameters θ. Reading that tree backwards is the last arrow, inference. Chapter 5 replaces the list of hypotheses by a continuous parameter $\boldsymbol{\theta}$, the likelihood by $\mathcal{L}(\boldsymbol{\theta}) = p(\boldsymbol{d} \mid \boldsymbol{\theta})$ (the probability density of the data $\boldsymbol{d}$ given $\boldsymbol{\theta}$), and the sum in the evidence by the integral $Z = p(\boldsymbol{d}) = \int p(\boldsymbol{d} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}$, which runs over every possible parameter value and is usually the hard part. Chapter 9 avoids computing it at all: in the proportional form (1.25) only ratios of posteriors appear, and Z cancels in every ratio, just as $\mathrm{P}(E)$ cancelled in the odds (1.26). In Chapter 4 we meet the forward direction on its own, in the p-value, which is a statement about $\mathrm{P}(\text{data} \mid H_0)$, the probability of the data if the null hypothesis H_0 ("nothing but chance") were true. It must never be read as $\mathrm{P}(H_0 \mid \text{data})$.

Reading the literature

Events. [Wasserman pp. 3–5, §1.2] Wasserman defines Ω, outcomes and events in three lines and gives the three examples of Example 1.2, the table of terminology (our dictionary, where $\varnothing$ is the "null event" that is always false and Ω the "true event"), disjointness, partitions, indicators and the limits of monotone sequences with the example $[0, 1/i)$. He writes $A \cap B$ as AB or (A, B), the convention used from here on. The proofs of De Morgan's laws and the distributive law are exercises there ([Wasserman p. 14, Ex. 4]) and worked

in [SeeingTheory pp. 20–21]. Kruschke adds the point that the sample space is set by the measurement operation [Kruschke §4.1].

Axioms. [Wasserman pp. 5–7, 13, §1.3, §1.9] Wasserman states the axioms (Definition 1.5, footnote on σ-fields), the two interpretations in one paragraph, the properties (1.1) without proof, inclusion–exclusion (Lemma 1.6) with the proof we expanded, the two-toss example (Example 1.7) and continuity (Theorem 1.8, increasing case only, the decreasing case being his Exercise 1). The appendix §1.9 defines σ-fields, probability spaces and the Borel σ-field. Cowan [Cowan §1.1] lists finite additivity instead of countable additivity as the third axiom, noting that this is a simplification. Countable additivity is what makes continuity and Example 1.9 work. The Bonferroni inequality and the disjointification in Boole's inequality are from [CasellaBerger §1.2]; the electron-mass argument is [Cowan §1.2.2]; calibration by gambles is [Kruschke §4.2.2].

Counting. [Wasserman pp. 7–8, §1.4] Wasserman treats finite sample spaces in a page: the 36 outcomes of two dice, $\mathrm{P}(\text{sum} = 11) = 2/36$, the uniform distribution $\mathrm{P}(A) = |A|/|\Omega|$, $n!$ with $0! = 1$, the binomial coefficient (1.2), the committee of 3 from 20 (1140), and $\binom{n}{0} = \binom{n}{n} = 1$, $\binom{n}{k} = \binom{n}{n-k}$. He writes that "we needn't delve into these in any great detail", which is why we supplied the derivations. The two-by-two table of counting rules, stars and bars, the lottery and poker counts, and the warning that unordered outcomes with replacement are not equally likely come from [CasellaBerger §1.2.3–1.2.4]. The recursion $P_n = nP_{n-1}$ and counting by overcounting are from [SeeingTheory pp. 22–24]. The birthday problem is not treated in these sources.

Trees. [Wasserman p. 12, Thm 1.16] Wasserman states the law of total probability for a finite partition and proves it in one line, as in our chain. He draws no trees. The tree picture, with the first stage as the hypotheses and the second as the outcomes they produce, organises the librarian–farmer, Monty Hall and medical-test problems above, and Lambert's "two routes to the joint probability" [Lambert §3.6.1] is a two-stage tree read in both orders. Cowan derives the same law as (1.7) [Cowan p. 3], and Seeing Theory's coins in a bag is Example 1.19 [SeeingTheory pp. 28–29].

Independence. [Wasserman pp. 8–10, §1.5] Wasserman defines independence by the product rule (Definition 1.9, with $A \amalg B$ for independence), distinguishes assumed from derived independence with the die events $\{2, 4, 6\}$ and $\{1, 2, 3, 4\}$, shows that disjoint events with positive probability are dependent and that a Venn diagram cannot display independence, and gives Examples 1.10 and 1.11 and a summary box. His Exercises 11, 14, 15 and 23 are solved above. The pairwise and mutual examples and de Méré's bet are from [CasellaBerger §1.3], the card-colour and causation remarks from [Lambert §3.4], and suit versus value from [Kruschke §4.4.2]. Seeing Theory lists $B = \{\text{second flip heads}\}$ as $\{HT, TT\}$ [SeeingTheory p. 12]. It should be $\{HH, TH\}$.

Conditioning and Bayes. Conditional probability and Bayes' theorem are [Wasserman pp. 11–13, §1.6–1.7], [Cowan §1.2], [CasellaBerger §1.3], [Kruschke §4.4] and [SeeingTheory pp. 26–30]. Wasserman gives the law of total probability and Bayes' theorem as Theorems 1.16 and 1.17 [Wasserman pp. 12–17], the names prior and posterior in Remark 1.18, the spam example as Example 1.19, and Monty Hall, the three cards, the rule that a losing hypothesis forces another to gain, the operating systems and the five coins as Exercises 10, 12, 18, 19 and 20. The two routes to the joint probability are [Lambert §3.6]. The disease example and the combined form of the theorem are [Cowan pp. 3–4, (1.6)–(1.8)]. The Morse-code and prisoners examples are [CasellaBerger §1.3], and the coins in a bag are [SeeingTheory pp. 28–29].

Recap: Events, conditional probability and trees

- An experiment has a sample space Ω of outcomes. Events are subsets, combined by $\cup$, $\cap$ and complement. In code an event is a Boolean mask over simulated outcomes.
- Probability obeys three axioms. Complement rule, monotonicity, inclusion–exclusion, Boole's inequality and continuity all follow. It can be read as a long-run frequency or as a degree of belief.
- With equally likely outcomes, probability is counting: $\mathrm{P}(A) = |A|/|\Omega|$, with $n!$ and $\binom{n}{k}$ as the tools (birthday problem).
- Information shrinks the sample space: $\mathrm{P}(A \mid B) = \mathrm{P}(AB)/\mathrm{P}(B)$, the overlap re-measured against the surviving region B. In code, `A[B].mean()`.
- Trees: multiply along a path (chain rule), add across paths (law of total probability, a weighted average over the unobserved stage).
- Independence, $\mathrm{P}(AB) = \mathrm{P}(A)\,\mathrm{P}(B)$, is not disjointness. Mutual independence needs every sub-collection.
- Bayes' theorem reads the tree backwards: $\mathrm{P}(H_i \mid E) \propto \mathrm{P}(E \mid H_i)\,\mathrm{P}(H_i)$, normalised by $\mathrm{P}(E) = \sum_j \mathrm{P}(E \mid H_j)\,\mathrm{P}(H_j)$. Posterior odds are the likelihood ratio times the prior odds.
- Efficiency is a forward number, purity an inverse one. Small priors make even good evidence weak (base rates).

1.8 Random variables: turning outcomes into numbers

A charged track crosses ten layers of a silicon tracker. Each layer either records a hit or misses it. The raw outcome of one track is therefore a 10-letter pattern such as `HHTHHTHHTT` (H for hit, T for miss). Nobody analyses patterns directly. The track-quality cut asks for a *number*: "how many hits?" A pattern-recognition algorithm asks for a different number: "what is the longest run of consecutive hits?" Both numbers are computed from the same pattern, and both are uncertain before the track arrives.

A **random variable** is exactly this: a rule that reads a number off each outcome. We want numbers because a pattern of letters cannot be put in a histogram, averaged or added, and a number can. Every data point, every estimator and every summary statistic in the rest of the book is a random variable.

Recall the box Ω of all possible outcomes from the first part of Chapter 1, where the probability of an event was its share of the area of the box. A random variable is a measuring device clipped onto that box. Nature picks the point ω, and the device prints the number $X(\omega)$. Nothing random happens inside the device. It is an ordinary function, and all the randomness is in which ω nature picked. Several devices can be clipped onto the same box, and each reads its own number off the *same* outcome. This is why two random variables can be related to each other (correlated, dependent): both are functions of one common ω.

Building a random variable takes four questions, asked in this order.

1. What is one outcome ω? (One full hit pattern, one collision event record, one sky map.)
2. What number do we want to read off it? Write it as a rule $\omega \mapsto X(\omega)$.
3. For each set A of numbers we might ask about, which outcomes give a value in A? That set of outcomes is an event, and its probability is the probability that X lands in A.

4. Collect those probabilities for all A. That list is the *distribution* of X, and from then on we can mostly forget Ω.

1.8.1 The definition, and the event hiding behind every statement about X

The four questions give one definition. They also give a rule we will use constantly: every statement about the number X is a statement about a set of outcomes.

Working definition 1.10 (random variable, distribution)

A **random variable** is a map

$$X : \Omega \to \mathbb{R}$$

that assigns a real number $X(\omega)$ to each outcome ω [a fixed rule, such as "count the H's", applied to whichever outcome occurs]. For a set $A \subseteq \mathbb{R}$ the **inverse image** is $X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\}$ [all outcomes whose reading lands in A. It is a set of outcomes, not an inverse function], and

$$\mathrm{P}(X \in A) := \mathrm{P}\big(X^{-1}(A)\big), \qquad \mathrm{P}(X = x) := \mathrm{P}\big(\{\omega : X(\omega) = x\}\big).$$

The collection of all the numbers $\mathrm{P}(X \in A)$ is the **distribution** of X [how the probability of Ω is spread along the real line by the map X]. *Operational test:* you have a random variable when, given any outcome, you can compute a single real number from it without further information.

Capital X is the random variable (the rule). Lower-case x is a particular value it can take. "$X = x$" is therefore an event: the set of outcomes whose reading is x. (Strictly, X must be **measurable**: every set $\{\omega : X(\omega) \le x\}$ must be an event we are allowed to assign a probability to. Every function met in practice is, and we will not mention it again [Wasserman p. 43].)

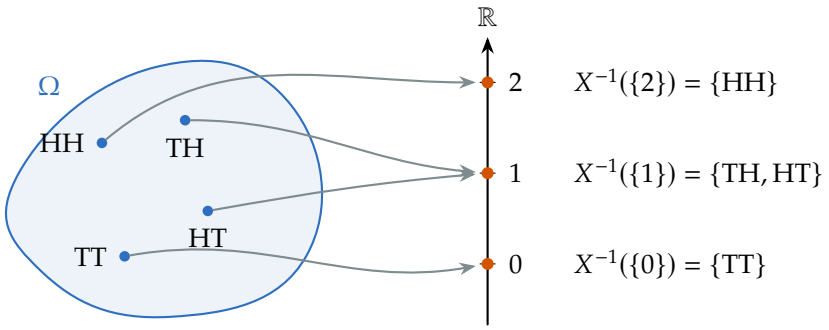


Figure 1.25: A random variable is a function from outcomes to numbers. For two coin flips, X = number of heads sends the four outcomes to three numbers. Two outcomes land on the value 1, so the probability that $X = 1$ collects the probability of both. The inverse images on the right are the events that statements about X secretly refer to.

Example 1.26 (Two flips: from the outcome table to the distribution). Flip a fair coin twice, $\Omega = \{\mathrm{TT}, \mathrm{TH}, \mathrm{HT}, \mathrm{HH}\}$, each outcome with probability $1/4$, and let X be the number of heads. Then

$$\begin{aligned} \mathrm{P}(X = 1) &\overset{(1)}{=} \mathrm{P}\big(\{\omega : X(\omega) = 1\}\big) \\ &\overset{(2)}{=} \mathrm{P}\big(\{\mathrm{TH}, \mathrm{HT}\}\big) \\ &\overset{(3)}{=} \mathrm{P}(\{\mathrm{TH}\}) + \mathrm{P}(\{\mathrm{HT}\}) = \tfrac{1}{4} + \tfrac{1}{4} = \tfrac{1}{2}. \end{aligned}$$

(1) Definition of $\mathrm{P}(X = x)$ as the probability of the inverse image.
(2) Read off which outcomes contain exactly one H.
(3) The two outcomes are disjoint events, so the axiom of additivity (first part of Chapter 1) adds their probabilities.

In the same way $\mathrm{P}(X = 0) = \mathrm{P}(\{\mathrm{TT}\}) = 1/4$ and $\mathrm{P}(X = 2) = \mathrm{P}(\{\mathrm{HH}\}) = 1/4$. Two tables summarise the situation:

ω	$\mathrm{P}(\{\omega\})$	$X(\omega)$
TT	1/4	0
TH	1/4	1
HT	1/4	1
HH	1/4	2

x	$\mathrm{P}(X = x)$
0	1/4
1	1/2
2	1/4

The left table describes the outcomes themselves. The right table is the distribution of X, and it answers every question about X alone. What the right table has lost, *which* flip was the head, is exactly what X does not measure.

Example 1.27 (Several random variables on one disk). Pick a point uniformly at random in the unit disk $\Omega = \{(x, y) : x^2 + y^2 \le 1\}$, so that the probability of a region is its area divided by π. The outcome is the point $\omega = (x, y)$. Four random variables on the same Ω are $X(\omega) = x$, $Y(\omega) = y$, $Z(\omega) = x + y$ and $W(\omega) = \sqrt{x^2 + y^2}$ [Wasserman p. 20]. Let us find the distribution of the distance W from the centre. For $0 \le r \le 1$,

$$\mathrm{P}(W \le r) \overset{(1)}{=} \mathrm{P}\big(\{(x, y) : x^2 + y^2 \le r^2\}\big) \overset{(2)}{=} \frac{\pi r^2}{\pi} = r^2.$$

(1) $W \le r$ happens exactly for the points inside the circle of radius r.
(2) Uniform on the disk means probability = area/π, and the small disk has area πr^2.

So $\mathrm{P}(W \le \frac{1}{2}) = \frac{1}{4}$: a "uniformly random point" is *not* uniformly distributed in its distance from the centre. Three quarters of the area lies in the outer ring $\frac{1}{2} < W \le 1$. A detector that is uniformly illuminated sees most hits near its rim, simply because the rim has more area.

1.8.2 Bernoulli and binomial: one trial and many trials

The simplest random variables take only two values. Many of the most useful ones in the book are sums of them.

Working definition 1.11 (Bernoulli and binomial)

X is **Bernoulli**(p) if $\mathrm{P}(X = 1) = p$ and $\mathrm{P}(X = 0) = 1 - p$ [a single yes-or-no trial, 1 for success]. Compactly $\mathrm{P}(X = x) = p^x(1 - p)^{1-x}$ for $x \in \{0, 1\}$. X is **binomial**(n, p) if it counts the successes in n successive, independent Bernoulli(p) trials [the same trial repeated n times, with outcomes that do not influence each other]. *Operational test:* a single detector layer that fires or not is Bernoulli. The number of the ten layers that fired is binomial, provided the layers have the same efficiency and fire independently.

Bernoulli is for a single trial; binomial is for multiple successive trials. That is the whole difference between the two. A binomial variable is a sum of Bernoulli variables, $X = X_1 + \cdots + X_n$ with X_i the indicator of success on trial i.

Example 1.28 (The binomial distribution, by counting paths). Generalise Example 1.26 to n trials with success probability p [Wasserman p. 20 ("try generalizing this to n flips")]. An outcome is a

sequence of n letters. Consider one particular sequence with exactly k successes, say $\mathrm{S}^k\mathrm{F}^{n-k}$ (all successes first). Then

$$\mathrm{P}(\mathrm{S}\cdots\mathrm{S}\,\mathrm{F}\cdots\mathrm{F}) \overset{(1)}{=} \underbrace{p\cdots p}_{k}\ \underbrace{(1-p)\cdots(1-p)}_{n-k} = p^k(1-p)^{n-k},$$

$$\mathrm{P}(X=k) \overset{(2)}{=} \sum_{\text{sequences with } k \text{ S}} p^k(1-p)^{n-k} \overset{(3)}{=} \binom{n}{k} p^k(1-p)^{n-k}, \qquad k=0,1,\ldots,n.$$

(1) The trials are independent, so the probability of a sequence is the product of the probabilities of its letters.
(2) $\{X=k\}$ is the disjoint union of all sequences with exactly k successes, and every one of them has the same probability (only the number of S's matters, not their positions).
(3) The number of such sequences is the number of ways to choose which k of the n positions carry an S, $\binom{n}{k} = n!/(k!(n-k)!)$ (the counting rule of the first part of Chapter 1).

For $n=10$, $p=\frac{1}{2}$: $\mathrm{P}(X=5) = \binom{10}{5}/2^{10} = 252/1024 = 0.2461$. Simulated coin tosses give 0.2467, against the exact 0.2461. The binomial and its relatives (geometric, Poisson, multinomial) are derived from first principles in Chapter 2.

Example 1.29 (The sample space that "disappears"). The named distributions are usually stated with no Ω in sight. Is there still a set of outcomes behind them? Yes: one can always be built [Wasserman p. 27]. Take $\Omega=[0,1]$ with $\mathrm{P}([a,b]) = b-a$ (uniform), fix p, and define

$$X(\omega) = \begin{cases} 1, & \omega \le p, \\ 0, & \omega > p. \end{cases}$$

Then

$$\mathrm{P}(X=1) \overset{(1)}{=} \mathrm{P}(\{\omega:\ \omega \le p\}) \overset{(2)}{=} \mathrm{P}([0,p]) \overset{(3)}{=} p,$$

(1) Inverse image of the value 1 under the rule above.
(2) The outcomes with $\omega \le p$ form the interval $[0,p]$.
(3) Uniform probability of an interval is its length.

and $\mathrm{P}(X=0) = 1-p$, so $X \sim \text{Bernoulli}(p)$. This is not a curiosity: it is exactly how a computer makes a Bernoulli variable, `x = (u <= p)` with `u` a uniform random number (Chapter 6 makes this the general recipe). In practice we think of a random variable as "a random number". Formally it is still a function on some Ω.

Watch out: Three notational traps

- $X \sim F$ means "X *has* distribution F", not "X is approximately F" [Wasserman p. 25].
- In Binomial(n,p), X is random and x is a value it can take, but n and p are **parameters**: fixed numbers of the model. The parameter is usually unknown and has to be learned from data. That is what statistical inference is about. Do not confuse the random variable with the parameter [Wasserman p. 26].
- X and Y are **equal in distribution**, $X \overset{d}{=} Y$, if $\mathrm{P}(X \le x) = \mathrm{P}(Y \le x)$ for every x. This does not mean $X = Y$. If $\mathrm{P}(X=1) = \mathrm{P}(X=-1) = \frac{1}{2}$ and $Y = -X$, then X and Y have the same distribution but $\mathrm{P}(X=Y) = 0$ [Wasserman p. 25].

1.8.3 Seeing it: two random variables on the same outcomes

A simulation shows two random variables read off the same outcomes. The script first builds a Bernoulli variable from a uniform ω, as in Example 1.29. It then generates 2×10^5 outcomes of ten fair flips (each row of an array is one ω) and applies two different maps to every row: X = number of heads and Y = longest run of heads.

Listing 1.9: Two random variables read off the same simulated outcomes (code/ch01/10_rv_as_map.py)

```
25 # --- 1. Bernoulli from a uniform outcome omega in [0,1] ------------------------------
26 p = 0.3
27 omega = rng.uniform(0.0, 1.0, size=100_000)       # the outcomes
28 X_bern = (omega <= p).astype(int)                  # the map omega -> {0,1}
29 bern_frac = X_bern.mean()
30
31 # --- 2. Ten fair coin flips: one outcome = one sequence of 10 H/T -------------------
32 n_exp, n_flip = 200_000, 10
33 flips = rng.integers(0, 2, size=(n_exp, n_flip))  # 1 = heads; each row is an omega
34
35
36 def longest_run(row):
37     best = cur = 0
38     for b in row:
39         cur = cur + 1 if b else 0
40         best = max(best, cur)
41     return best
42
43
44 X = flips.sum(axis=1)                              # map 1: number of heads
45 Y = np.array([longest_run(r) for r in flips])      # map 2: longest run of heads
46
47 k = np.arange(n_flip + 1)
48 pmf_X = np.array([(X == j).mean() for j in k])
49 pmf_Y = np.array([(Y == j).mean() for j in k])
50 pmf_binom = stats.binom.pmf(k, n_flip, 0.5)
```

Lines 26–29 are Example 1.29 verbatim: `omega` is an array of 10^5 uniform outcomes and `X_bern` is the boolean test $\omega \le p$ turned into 0/1. Its average is the fraction of ones, 0.3016 for $p = 0.3$. Line 33 draws the outcomes of the second experiment: an array with one row per outcome and one column per flip. The two maps on lines 44–45 are ordinary Python functions of a row, which is the definition of a random variable made executable. Lines 47–50 estimate each distribution by the fraction of outcomes that land on each value.

The numbers the script produced: the simulated $P(X = 5)$ is 0.2467 (exact 0.2461), the average number of heads is 4.997 (we will show in Section 1.12 that the exact value is $np = 5$), the average longest run is 2.796, and the most likely longest run is 2, with $P(Y = 3) = 0.2606$. No simple formula gives the distribution of Y. The simulation gets it with the same effort as the binomial one, because once we can generate outcomes, *any* random variable is just a function applied to them.

What this bought us. A random variable is a function, so a histogram of a random variable is a histogram of a function evaluated on simulated outcomes. That single sentence is the whole of Monte Carlo simulation (Chapter 6).

How research uses it. A collider Monte Carlo generator produces outcomes ω: complete event records with every particle's four-momentum. An analysis then defines random variables on them: the missing transverse momentum, the invariant mass of the two leading jets, the

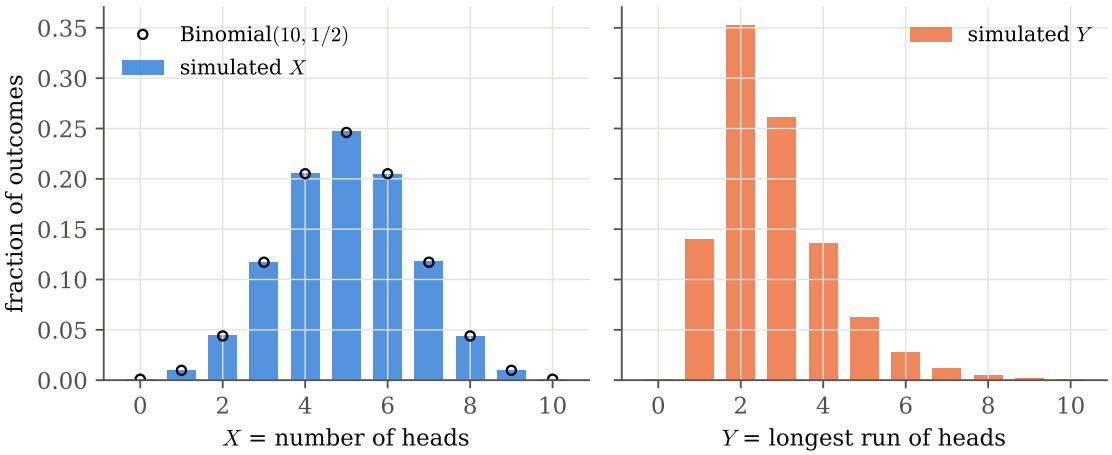


Figure 1.26: The same 2×10^5 simulated ten-flip outcomes, read by two different random variables. Left: the number of heads, whose fractions sit on the binomial probabilities (open circles). Right: the longest run of heads, which has a completely different, skewed distribution although it is computed from exactly the same outcomes.

number of b-tagged jets. The same ω's feed every observable, which is why one simulated sample serves a whole collaboration, and why two observables computed from the same events are in general correlated (Section 1.13). In cosmology the outcome is a whole simulated sky map, and the random variables are its summaries: a power spectrum $\widehat{C}_\ell$ at each ℓ, the number of hot spots above a threshold, the Euler characteristic of an excursion set (Chapters 7 and 12).

Later chapters use many named random variables. The table lists the named distributions of [Wasserman pp. 26–30], each with the mechanism that produces it. Chapter 2 derives every one of them from its mechanism.

Taxonomy: Named random variables and the story that produces each

distribution	generating story
point mass δ_a	no randomness at all, $\mathrm{P}(X = a) = 1$
discrete uniform	k equally likely labels (a fair die)
Bernoulli(p)	one yes/no trial
Binomial(n, p)	successes in n independent identical trials
Geometric(p)	number of trials up to and including the first success, $p(1-p)^{k-1}$
Poisson(λ)	counts of rare events in a fixed window (decays, events in a bin), $e^{-\lambda}\lambda^k/k!$
Uniform(a, b)	"a point at random" in an interval, density $1/(b-a)$
Normal $\mathcal{N}(\mu, \sigma^2)$	sums of many small independent effects (noise)
Exponential(β)	waiting time to the next rare event, density $e^{-x/\beta}/\beta$
Gamma, χ^2_p, t_ν, Beta, Cauchy	sums of waiting times, sums of squared Gaussians, ratios, probabilities of probabilities, resonance line shapes

1.9 Mass, density and the cumulative distribution

A test beam fires electrons of exactly 50 GeV into an electromagnetic calorimeter. Although the beam energy is fixed, fluctuations in the shower and noise in the readout make the measured energy E scatter around 50 with a spread of about 0.25 GeV. We want to answer questions such

as "what fraction of electrons are measured below 49.5?", "which energy is exceeded by only 5% of them?" and "how does the histogram of measured energies look if we take more data?"

Three objects describe how a random variable is spread out: the probability mass function for counts, the density for continuous readings, and the cumulative distribution for both. We need more than one because counts and continuous readings differ in kind. For a count (hits, heads) we can list the probability of each value. For a continuous reading like E we cannot, because the probability of measuring *exactly* $50.000\ldots$ is zero. Each of the three objects can be read off a histogram. The idea people get wrong most often is that a density is probability *per unit* x, so it carries units.

Mass and density are familiar from mechanics. A string of beads has point masses. A metal rod has a mass density $\lambda(x)$ in kg/m, and the mass of a piece is $\int \lambda \, dx$. Nobody asks "what is the mass at the point $x = 0.3\,\text{m}$?". The answer is zero, and yet $\lambda(0.3)$ is a perfectly good number. Probability works the same way. Discrete random variables are beads: the probability sits on isolated values. Continuous random variables are rods: the probability of a single point is zero, and $f(x)$ is probability per unit length. The cumulative distribution $F(x)$ is "the mass to the left of x", which makes sense for beads, rods and anything in between.

To describe how a random number is spread out, we ask four questions in turn.

1. Does X take isolated values (counts) or a continuum (energies, angles, positions)?
2. Isolated values: list $\mathrm{P}(X = x)$ for each value.
3. Continuum: ask instead for the probability of a small interval, divided by its width. That ratio is the density.
4. Either way: the probability of landing at or below x is a single function $F(x)$ that answers every interval question by subtraction, and whose inverse gives the quantiles.

1.9.1 The cumulative distribution function

Beads and rods need different bookkeeping, but one function works for both: the total probability collected up to a point.

Working definition 1.12 (cumulative distribution function)

The **cumulative distribution function** (CDF) of X is

$$F_X(x) = \mathrm{P}(X \le x), \qquad x \in \mathbb{R},$$

[the probability that the reading is at most x, defined for every real x, including values X never takes]. We write F when the variable is clear. *Operational test:* sort a sample of N readings. The fraction of readings $\le x$ is the **empirical CDF**, and it tends to $F(x)$ as $N \to \infty$.

Example 1.30 (The CDF of two flips). With X = number of heads in two fair flips (Example 1.26),

$$F_X(x) = \begin{cases} 0, & x < 0, \\ 1/4, & 0 \le x < 1, \\ 3/4, & 1 \le x < 2, \\ 1, & x \ge 2. \end{cases}$$

For instance

$$F_X(1.4) \overset{(1)}{=} \mathrm{P}(X \le 1.4) \overset{(2)}{=} \mathrm{P}(X = 0) + \mathrm{P}(X = 1) = \tfrac{1}{4} + \tfrac{1}{2} = \tfrac{3}{4}.$$

(1) Definition of the CDF at the point $x = 1.4$.
(2) X only takes the values 0, 1, 2, and the ones ≤ 1.4 are 0 and 1. The events $\{X = 0\}$ and $\{X = 1\}$ are disjoint, so their probabilities add.

The CDF is a staircase. It is flat between the values, jumps by $\mathrm{P}(X = x)$ at each value, and at a jump takes the *upper* value (the filled dots in Figure 1.27): $F(1) = 3/4$, not $1/4$, because $X \le 1$ includes $X = 1$.

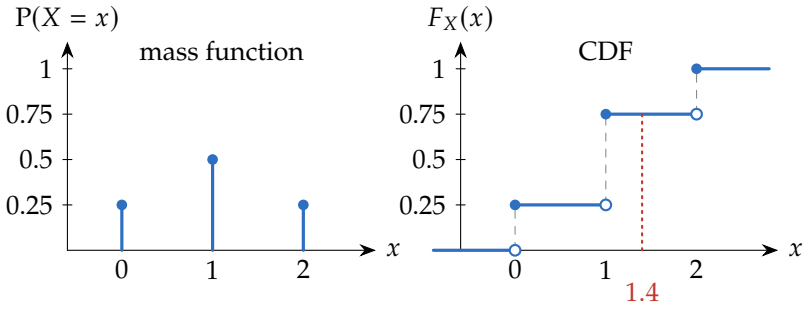


Figure 1.27: Two fair flips. Left: the mass function puts probability $1/4, 1/2, 1/4$ on the values 0, 1, 2. Right: the CDF accumulates it into a staircase that jumps by those amounts. Filled dots mark the value the CDF takes at each jump, open dots the value it approaches from the left. The dotted line reads off $F(1.4) = 3/4$.

Every CDF has three properties, and conversely any function with these properties is the CDF of some random variable [Wasserman Thm 2.8, p. 21]:

(i) *non-decreasing*: $x_1 < x_2 \Rightarrow F(x_1) \le F(x_2)$, because $\{X \le x_1\} \subseteq \{X \le x_2\}$;
(ii) *normalised*: $F(x) \to 0$ as $x \to -\infty$ and $F(x) \to 1$ as $x \to +\infty$;
(iii) *right-continuous*: $F(x) = \lim_{y \downarrow x} F(y)$.

Property (iii) is why the CDF takes the upper value at each jump (the filled dots). Why does it hold? The tool is the *continuity of probability* from the first part of Chapter 1: if $A_1 \supseteq A_2 \supseteq \cdots$ shrink to $A = \bigcap_i A_i$, then $\mathrm{P}(A) = \lim_i \mathrm{P}(A_i)$. Take any sequence $y_1 > y_2 > \cdots$ decreasing to x, and $A_i = \{X \le y_i\}$. Then

$$F(x) \overset{(1)}{=} \mathrm{P}(X \le x) \overset{(2)}{=} \mathrm{P}\Big(\bigcap_i \{X \le y_i\}\Big) \overset{(3)}{=} \lim_i \mathrm{P}(X \le y_i) = \lim_i F(y_i).$$

(1) Definition of F.
(2) $X \le x$ holds exactly when $X \le y_i$ for every i, because the y_i come arbitrarily close to x from above.
(3) The events $\{X \le y_i\}$ shrink, so continuity of probability applies.

From the left the argument fails. The events $\{X \le y_i\}$ with $y_i \uparrow x$ grow to $\{X < x\}$, and that event misses the point $X = x$. This asymmetry gives the handy rules of [Wasserman Lemma 2.15, pp. 24–25]. Write $F(x^-) = \lim_{y \uparrow x} F(y)$. Then

$$\mathrm{P}(X = x) = F(x) - F(x^-), \qquad \mathrm{P}(x < X \le y) = F(y) - F(x), \qquad \mathrm{P}(X > x) = 1 - F(x). \tag{1.27}$$

The first says that the size of the jump at x is the probability of the value x. The second is additivity: $\{X \le y\}$ is the disjoint union of $\{X \le x\}$ and $\{x < X \le y\}$. The third is the complement rule.

1.9.2 Discrete variables: the probability mass function

For counts the bookkeeping is the simplest one, a list of probabilities, one per allowed value.

Working definition 1.13 (probability mass function)

X is **discrete** if it takes countably many values $x_1, x_2, \ldots$ [values that can be listed one after another, like $0, 1, 2, \ldots$]. Its **probability mass function** (pmf) is $f_X(x) = \mathrm{P}(X = x)$. It satisfies $f_X(x) \geq 0$, $\sum_i f_X(x_i) = 1$, and $F_X(x) = \sum_{x_i \leq x} f_X(x_i)$. *Operational test:* from N simulated or measured values, the fraction equal to x_i estimates $f_X(x_i)$. The fractions are pure numbers and none can exceed 1.

A set is countable if its elements can be listed so that each one is reached after finitely many steps. The integers are countable: list them as $0, 1, -1, 2, -2, \ldots$. The real numbers in $[0, 1]$ are not. Cantor's diagonal argument builds, from any proposed list, a number that differs from the n-th entry in its n-th decimal, so it is missing from the list [SeeingTheory pp. 33–34]. A Poisson count can take infinitely many values and is still discrete.

1.9.3 Continuous variables: density is probability per unit x

Return to the calorimeter of the opening, which measures 50 GeV electrons with a spread of about 0.25 GeV. Divide the energy axis into bins of width Δ, record N electrons and let n_i be the number in bin i. The fraction n_i/N estimates the probability of the bin. It depends on the bin width: halve Δ and every fraction roughly halves. Dividing by the width removes this dependence. The ratio

$$\frac{n_i}{N\,\Delta} \approx \frac{\mathrm{P}(\text{bin } i)}{\Delta}$$

does *not* depend on the bin width once the bins are narrow, and as $N \to \infty$ and $\Delta \to 0$ it becomes a smooth function of E, the probability density [Cowan §1.3, Fig. 1.2] [Kruschke §4.3.2].

Working definition 1.14 (probability density function)

X is **continuous** if there is a function $f_X \geq 0$ with $\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = 1$ such that for every $a \leq b$

$$\mathrm{P}(a < X < b) = \int_a^b f_X(x)\,\mathrm{d}x.$$

f_X is the **probability density function** (pdf) [probability per unit of x, so $f_X(x)\,\mathrm{d}x$ is the probability of the small interval $[x, x + \mathrm{d}x]$]. Then $F_X(x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t$, and $f_X = F_X'$ wherever F_X is differentiable. *Operational test:* histogram N values in bins of width Δ and divide each count by $N\Delta$. The result has units of $1/[x]$ and converges to f_X.

Example 1.31 (A density has units, and its peak can exceed 1). Let $E \sim \mathcal{N}(\mu, \sigma^2)$ with $\mu = 50\,\mathrm{GeV}$ and $\sigma = 0.25\,\mathrm{GeV}$. The Gaussian density is

$$f(E) = \frac{1}{\sigma\sqrt{2\pi}} \exp\!\left[-\frac{(E-\mu)^2}{2\sigma^2}\right],$$

so its peak value is

$$f(\mu) \overset{(1)}{=} \frac{1}{\sigma\sqrt{2\pi}} = \frac{1}{0.25\,\mathrm{GeV} \times 2.5066} \overset{(2)}{=} 1.596\,\mathrm{GeV}^{-1} \overset{(3)}{=} 1.596 \times 10^{-3}\,\mathrm{MeV}^{-1}.$$

(1) At $E = \mu$ the exponent vanishes and $e^0 = 1$.
(2) $\sqrt{2\pi} = 2.5066$. The unit is inverse GeV because σ has units of GeV.
(3) $1\,\text{GeV} = 10^3\,\text{MeV}$, so $1\,\text{GeV}^{-1} = 10^{-3}\,\text{MeV}^{-1}$.

A density of 1.6 is not a probability of 1.6. The probability of a 10 MeV window at the peak is about $1.596\,\text{GeV}^{-1} \times 0.010\,\text{GeV} = 0.016$. The same physical statement written in MeV gives $1.596 \times 10^{-3}\,\text{MeV}^{-1} \times 10\,\text{MeV} = 0.016$. The density changed by a factor 1000 when we changed units, and the probability did not change at all. This is the first instance of the Jacobian rule of Section 1.11: densities transform, probabilities do not.

Example 1.32 (Valid and invalid densities). Checking a candidate density means checking $f \geq 0$ and $\int f = 1$ [Wasserman pp. 23–24].

- $f(x) = 1$ on $[0, 1]$ (Uniform$(0, 1)$). The CDF is $F(x) = x$ on $[0, 1]$, 0 to the left, 1 to the right.
- $f(x) = 5$ on $[0, \frac{1}{5}]$. Here $\int f = 5 \times \frac{1}{5} = 1$: a valid density whose value is 5 everywhere on its support.
- $f(x) = \frac{2}{3}x^{-1/3}$ on $(0, 1)$. It is unbounded at 0, yet

$$\int_0^1 \tfrac{2}{3}x^{-1/3}\mathrm{d}x \overset{(1)}{=} \tfrac{2}{3}\left[\tfrac{3}{2}x^{2/3}\right]_0^1 = 1.$$

  (1) The antiderivative of $x^{-1/3}$ is $x^{2/3}/(2/3)$, and it vanishes at $x = 0$, so the singularity is integrable.

- $f(x) = 1/(1+x)^2$ for $x \geq 0$: $\int_0^\infty (1+x)^{-2}\mathrm{d}x = \left[-(1+x)^{-1}\right]_0^\infty = 1$. Valid.
- $f(x) = 1/(1+x)$ for $x \geq 0$ is *not* a density: $\int_0^\infty \mathrm{d}x/(1+x) = \int_1^\infty \mathrm{d}u/u = \ln\infty = \infty$. It falls off too slowly to be normalised.

Watch out: A density is not a probability

If X is continuous then $\mathrm{P}(X = x) = 0$ for every x. Do not read $f(x)$ as $\mathrm{P}(X = x)$. Probabilities come from integrating f over an interval. A pdf can exceed 1 and can even be infinite at a point. A pmf can do neither. For a continuous X it does not matter whether interval ends are included: $\mathrm{P}(a < X < b) = \mathrm{P}(a \leq X < b) = \mathrm{P}(a < X \leq b) = \mathrm{P}(a \leq X \leq b) = F(b) - F(a)$.

Neither mass nor density. Some measured quantities are mixed. A zero-suppressed calorimeter cell reads exactly 0 with a finite probability (no signal above threshold) and a continuous energy otherwise. Such a variable has neither a pmf nor a pdf. Its CDF has a jump at 0 and rises smoothly afterwards. The CDF still exists and still determines everything: if two variables have the same CDF then $\mathrm{P}(X \in A) = \mathrm{P}(Y \in A)$ for every A [Wasserman Thm 2.7, p. 21]. That is why the CDF, not the density, is the fundamental object.[2]

1.9.4 Quantiles: reading the CDF backwards

The CDF answers "what fraction lies below x?". The reverse question, "below which x lies a fraction q?", is answered by the quantile function.

[2] Loosely, one hears that a variable is continuous "if it takes uncountably many values" [SeeingTheory p. 35]. The mixed variable above takes uncountably many values and has no density, so the definition through $\int f$ is the correct one.

Working definition 1.15 (quantile function, median, quartiles)

The **quantile function** or inverse CDF is

$$F^{-1}(q) = \inf\{x : F(x) \geq q\}, \qquad 0 < q < 1,$$

[the smallest x at which the CDF has reached height q]. If F is continuous and strictly increasing, $F^{-1}(q)$ is simply the unique x with $F(x) = q$. $F^{-1}(\frac{1}{2})$ is the **median**, $F^{-1}(\frac{1}{4})$ and $F^{-1}(\frac{3}{4})$ are the first and third **quartiles**. The **mode** is the value where the density (or pmf) is largest. *Operational test:* sort N values. The value at position $\lceil qN \rceil$ is the sample q-quantile.

The definition above uses $\geq$, and other books use a strict inequality.[3] Nothing below depends on the choice.

Example 1.33 (Gaussian probabilities and quantiles by standardising). Let $X \sim \mathcal{N}(3, 5)$, so $\mu = 3$ and $\sigma^2 = 5$ (the second argument is the variance). Write $Z = (X - \mu)/\sigma$, which is standard normal $\mathcal{N}(0, 1)$ (we prove this in Section 1.11), with CDF Φ. Then [Wasserman Ex. 2.17, p. 29]

$$\mathrm{P}(X > 1) \overset{(1)}{=} 1 - \mathrm{P}(X \leq 1) \overset{(2)}{=} 1 - \mathrm{P}\Big(Z \leq \frac{1-3}{\sqrt{5}}\Big) \overset{(3)}{=} 1 - \Phi(-0.8944) \overset{(4)}{=} 1 - 0.1855 = 0.81.$$

(1) Complement rule (1.27).
(2) $X \leq 1$ is the same event as $(X - 3)/\sqrt{5} \leq (1 - 3)/\sqrt{5}$, because subtracting and dividing by a positive number preserve the inequality.
(3) Definition of Φ, with $-2/2.2361 = -0.8944$.
(4) A table or `scipy.stats.norm.cdf(-0.8944)`.

For the 20% quantile q we need $\mathrm{P}(X < q) = 0.2$:

$$0.2 \overset{(1)}{=} \Phi\Big(\frac{q-3}{\sqrt{5}}\Big) \quad \Longrightarrow \quad \frac{q-3}{\sqrt{5}} \overset{(2)}{=} \Phi^{-1}(0.2) = -0.8416 \quad \Longrightarrow \quad q = 3 - 0.8416\sqrt{5} = 1.1181.$$

(1) Same standardisation as above.
(2) Apply the inverse of the strictly increasing function Φ to both sides. $\Phi^{-1}(0.2) = -0.8416$ from a table or `norm.ppf(0.2)`.

Example 1.34 (Exponential lifetimes: quantiles and a random-number recipe). A detector component has lifetime $X \sim \mathrm{Exp}(\beta)$, density $f(x) = e^{-x/\beta}/\beta$ for $x > 0$. Its CDF and quantile function are

$$F(x) \overset{(1)}{=} \int_0^x \frac{e^{-t/\beta}}{\beta}\mathrm{d}t = \big[-e^{-t/\beta}\big]_0^x = 1 - e^{-x/\beta},$$
$$F^{-1}(q) \overset{(2)}{=} -\beta \ln(1 - q).$$

(1) $f = 0$ for $t < 0$, so the integral starts at 0. The antiderivative of $e^{-t/\beta}/\beta$ is $-e^{-t/\beta}$.
(2) Solve $q = 1 - e^{-x/\beta}$ for x: $e^{-x/\beta} = 1 - q$, take logs.

The median lifetime is $\beta \ln 2 = 0.693\beta$, shorter than the mean β (Section 1.12): a few very long-lived components pull the mean up.

The quantile function is also a random-number generator. If $U \sim \mathrm{Uniform}(0, 1)$ and F is continuous and strictly increasing, then $X = F^{-1}(U)$ has CDF F:

$$\mathrm{P}\big(F^{-1}(U) \leq x\big) \overset{(1)}{=} \mathrm{P}\big(U \leq F(x)\big) \overset{(2)}{=} F(x).$$

[3] [Wasserman Def. 2.16] writes $\inf\{x : F(x) > q\}$. We use $\geq$, as in [CasellaBerger §2.1]. The two agree whenever F is strictly increasing and differ only on flat stretches of F. For two flips and $q = \frac{1}{4}$ the strict version gives 1, the $\geq$ version gives 0.

(1) F is strictly increasing, so applying it to both sides of $F^{-1}(U) \le x$ gives the equivalent inequality $U \le F(x)$.
(2) For a uniform U, $\mathrm{P}(U \le u) = u$ for $u \in [0,1]$.

So `x = -beta*np.log(1-u)` turns uniform numbers into exponential lifetimes. Running the argument backwards, $Y = F(X)$ is Uniform$(0,1)$ whenever X has the continuous CDF F (the **probability integral transform**) [Wasserman Exercise 2.15, p. 45] [CasellaBerger Thm 2.1.10]. Both facts become the inverse-transform method of Chapter 6 and the idea behind calibrated p-values in Chapter 4.

Example 1.35 (Which interval holds 95%?). A q-quantile pins one point. To summarise a spread we need an interval that holds, say, 95% of the probability. Many intervals do, so we must say which one we mean. Two choices are common. The **equal-tailed interval** $[F^{-1}(0.025), F^{-1}(0.975)]$ leaves 2.5% in each tail. The **highest density interval** (HDI) contains the values whose density exceeds a level W, with W chosen so that the enclosed probability is 95%. Every point inside it is more probable (per unit x) than every point outside [Kruschke §4.3.4]. For a Gaussian both are $\mu \pm 1.96\sigma$. For the exponential lifetime they differ:

$$\text{equal-tailed: } \big[-\beta \ln 0.975,\ -\beta \ln 0.025\big] = [0.025\beta,\ 3.689\beta], \quad \text{width } 3.66\beta,$$
$$\text{HDI: } \big[0,\ -\beta \ln 0.05\big] = [0,\ 2.996\beta], \quad \text{width } 3.00\beta.$$

The density is largest at $x = 0$ and decreases, so the highest-density region must start at 0 and extend until $F(x) = 0.95$. It is the narrowest interval with 95% probability. Both appear later: equal-tailed intervals in frequentist confidence constructions (Chapter 3), HDIs for Bayesian posteriors (Chapter 5).

1.9.5 Seeing it: from counts to density, and the empirical CDF

Listing 1.10: Histogram counts, divided by $N\Delta$, become a density (`code/ch01/11_density_histogram.py`)

```
mu, sigma = 50.0, 0.25                      # GeV
E_all = rng.normal(mu, sigma, size=100_000)
grid = np.linspace(mu - 4 * sigma, mu + 4 * sigma, 400)
pdf = stats.norm.pdf(grid, mu, sigma)       # per GeV

# --- figure 1: counts (left) and counts/(N*width) (right) ---------------------------
fig, axes = plt.subplots(1, 2)
cases = [(100, 0.2), (1_000, 0.1), (100_000, 0.025)]
for (N, w), c in zip(cases, SERIES):
    E = E_all[:N]
    bins = np.arange(mu - 4 * sigma, mu + 4 * sigma + w / 2, w)
    counts, edges = np.histogram(E, bins=bins)
    centres = 0.5 * (edges[1:] + edges[:-1])
    axes[0].step(centres, counts, where="mid", color=c, label=f"$N={N}$, $\\Delta={w}$")
    dens = counts / (N * w)
    axes[1].step(centres, dens, where="mid", color=c, label=f"$N={N}$")
theory_line(axes[1], grid, pdf, label="pdf")
axes[0].set_yscale("log")
axes[0].set_xlabel("$E$ [GeV]")
axes[0].set_ylabel("counts per bin")
axes[0].set_ylim(0.7, 2e5)
axes[0].legend(fontsize=7, loc="upper left")
axes[1].set_xlabel("$E$ [GeV]")
axes[1].set_ylabel("counts / ($N\\Delta$)  [GeV$^{-1}$]")
axes[1].legend(fontsize=7, loc="upper right")
fig.tight_layout()
```

The script draws 10^5 Gaussian energies with $\mu = 50$, $\sigma = 0.25$ (line 26). The loop over `cases` makes three histograms with different numbers of events and bin widths. The left panel plots raw counts per bin (line 38): they differ by orders of magnitude between the three cases, because counts scale with both N and Δ. Line 39 divides by $N\Delta$, and the right panel shows that all three collapse onto the same curve, the pdf drawn by `theory_line`.

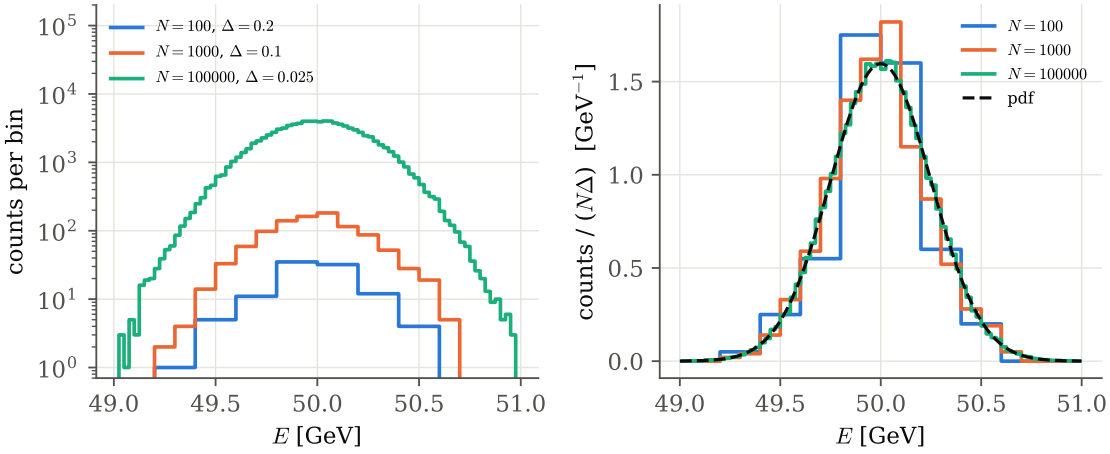


Figure 1.28: Left: counts per bin for three data sets ($N = 100$, 10^3, 10^5 with bins of 0.2, 0.1, 0.025 GeV). The heights are not comparable. Right: the same histograms divided by $N\Delta$. They now fluctuate around one curve, the Gaussian density, whose peak is above 1 per GeV.

With the finest bins the estimated peak density is $1.612\,\text{GeV}^{-1}$ against the exact $1.596\,\text{GeV}^{-1}$, and the fraction of electrons within $\pm\sigma$ of 50 GeV is 0.683 (the Gaussian value is 0.6827).

Listing 1.11: The empirical CDF is a sorted sample (`code/ch01/11_density_histogram.py`)

```
58
59 # --- figure 2: empirical CDF and quantiles ----------------------------------------
60 N = 1_000
61 E = np.sort(E_all[:N])
62 ecdf = np.arange(1, N + 1) / N
63 fig, axes = plt.subplots(1, 2)
64 ax = axes[0]
65 ax.step(E, ecdf, where="post", color=SERIES[0], label=f"empirical CDF, $N={N}$")
66 theory_line(ax, grid, stats.norm.cdf(grid, mu, sigma), label="$F(E)$")
67 for q, c in zip([0.25, 0.5, 0.75], SERIES[1:]):
68     xq = stats.norm.ppf(q, mu, sigma)
69     ax.plot([grid[0], xq], [q, q], color=c, lw=0.9)
```

Lines 60–61 are the whole empirical CDF: sort the sample, and assign to the i-th smallest value the height i/N. Lines 66–69 draw the quartiles as the construction "go across at height q, then down to the axis". With $N = 1000$ the sample quartiles are 49.826, 49.996, 50.146 GeV, against the exact 49.831, 50.000, 50.169 GeV.

What this bought us. Counts depend on how you bin. Densities do not, but they carry units of $1/[x]$. The CDF needs no binning at all, which is why tests that compare a sample to a model, such as Kolmogorov–Smirnov in Chapter 4, are built on the empirical CDF.

How research uses it. Here the histogram *verifies* a density we already know. In research the direction is often reversed: nobody knows the density of, say, the estimated tensor-to-scalar ratio after a realistic foreground cleaning pipeline, so we simulate the whole pipeline many

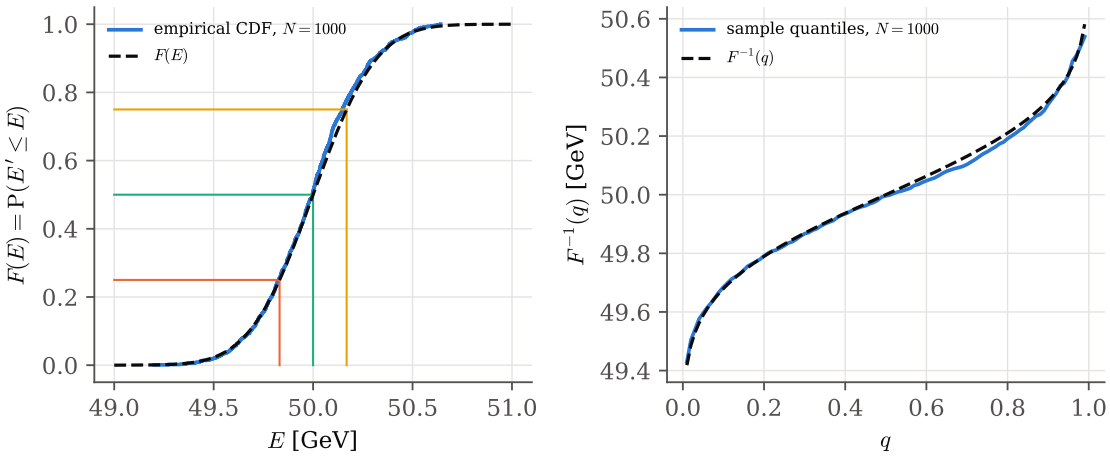


Figure 1.29: Left: the empirical CDF of 1000 simulated energies (a staircase with 1000 tiny steps) on top of the exact CDF. The coloured lines read off the quartiles: across at $q = 0.25, 0.5, 0.75$ and down to the energy axis. Right: the same information turned sideways, sample quantiles against q, compared with the exact quantile function.

times and the normalised histogram (or the empirical CDF) of the output *is* our best knowledge of its sampling distribution. Its quantiles become the error bars. Chapters 6 and 8 do exactly this.

1.10 Several variables at once: joint, marginal and conditional distributions

A calorimeter measures a photon. Two numbers describe the event: the true energy E that nature gave the photon, and the measured energy $E_{\text{meas}} = E +$ noise that the readout reports. The physics (a falling spectrum from a decay or a background process) tells us how E is distributed. The detector response tells us how E_{meas} is spread for a given E. But the histogram we actually fill contains only E_{meas}. The true energy is a **nuisance variable**: it is essential to the physics of the measurement, and we never observe it.

The same situation appears in every cosmological analysis. A fit returns a joint constraint on (Ω_m, σ_8), and a talk quotes "$\Omega_m = 0.31 \pm 0.01$" with σ_8 "marginalised over". In both cases we hold the joint distribution of several random variables and need one of two operations. *Marginalising* isolates one variable by averaging out the others. *Conditioning* fixes one variable and asks how the rest are distributed. Neither is new. Both are rules from the first part of Chapter 1 (the law of total probability and conditional probability), applied to densities.

Two pictures make both operations concrete. Draw a scatter plot of many simulated pairs (x, y). The density of points in the plane is the joint density. Squash all the points onto the x axis, ignoring y, and histogram them: that is the marginal of x. Instead, keep only the points in a thin vertical band at $x = x_0$, histogram their y values and normalise: that is the conditional density of y given x_0 [Cowan Figs. 1.4–1.6]. Projecting adds up all the ways a value of x could have arisen. Slicing throws away everything that disagrees with what we know. This is the "shrinking of the sample space" of the first part of this chapter.

So with two random variables there are four questions to ask.

1. How are the two variables distributed *together*? (Joint.)
2. How is one of them distributed if we do not care about, or cannot see, the other? (Marginal:

average out the nuisance.)

3. How is one distributed if we *know* the value of the other? (Conditional: slice and renormalise.)
4. Does knowing one change what we expect for the other? (If not, they are independent, and the joint is the product of the marginals.)

1.10.1 Joint distributions

The other three questions are all answered from the distribution of the pair taken together, so we define it first.

Working definition 1.16 (joint mass function, joint density)

For discrete X, Y the **joint mass function** is $f(x,y) = \mathrm{P}(X = x, Y = y)$ [the probability that both statements hold at once, $\mathrm{P}(X = x \text{ and } Y = y)$]. For continuous variables a **joint density** is a function $f(x,y) \geq 0$ with $\iint f \,\mathrm{d}x\,\mathrm{d}y = 1$ such that for every region A of the plane

$$\mathrm{P}\big((X,Y) \in A\big) = \iint_A f(x,y)\,\mathrm{d}x\,\mathrm{d}y,$$

[probability per unit area, so $f(x,y)\,\mathrm{d}x\,\mathrm{d}y$ is the probability of the small rectangle at (x,y)]. The joint CDF is $F(x,y) = \mathrm{P}(X \leq x, Y \leq y)$. *Operational test:* a 2-D histogram of N simulated pairs with cells $\Delta x \times \Delta y$, divided by $N\Delta x\Delta y$, estimates $f(x,y)$. Its units are $1/([x][y])$.

Example 1.36 (A joint table and a uniform square). Two binary variables with the joint table [Wasserman Ex. 2.18, p. 31]

	$Y=0$	$Y=1$	
$X=0$	$1/9$	$2/9$	$1/3$
$X=1$	$2/9$	$4/9$	$2/3$
	$1/3$	$2/3$	1

have $f(1,1) = \mathrm{P}(X=1, Y=1) = 4/9$. The row and column sums written in the margins of the table are the marginals of the next subsection. For a continuous example, let (X,Y) be uniform on the unit square, $f = 1$ there. Then $\mathrm{P}(X < \frac{1}{2}, Y < \frac{1}{2})$ is the integral of 1 over the lower-left quarter, which is its area, $\frac{1}{4}$.

Example 1.37 (A density on a curved region). Let $f(x,y) = c\,x^2 y$ on the region $x^2 \leq y \leq 1$ (and 0 elsewhere) [Wasserman Ex. 2.22, p. 32]. The region forces $-1 \leq x \leq 1$. Fix c by normalisation, integrating y first at fixed x:

$$1 \overset{(1)}{=} c\int_{-1}^{1}\!\!\int_{x^2}^{1} x^2 y\,\mathrm{d}y\,\mathrm{d}x \overset{(2)}{=} c\int_{-1}^{1} x^2\,\frac{1-x^4}{2}\,\mathrm{d}x \overset{(3)}{=} \frac{c}{2}\Big(\frac{2}{3}-\frac{2}{7}\Big) = \frac{4c}{21} \quad\Longrightarrow\quad c = \frac{21}{4}.$$

(1) For each $x \in [-1,1]$, y runs over $[x^2, 1]$: this is the only delicate step, and Figure 1.30 is there to make it visible.

(2) $\int_{x^2}^{1} y\,\mathrm{d}y = \frac{1}{2}(1-x^4)$.

(3) $\int_{-1}^{1} x^2\mathrm{d}x = \frac{2}{3}$ and $\int_{-1}^{1} x^6\mathrm{d}x = \frac{2}{7}$.

Now the probability that $X \geq Y$. Inside the support this needs $x \geq y \geq x^2$, possible only for $0 \leq x \leq 1$ (the hatched sliver in the figure):

$$\mathrm{P}(X \geq Y) \overset{(1)}{=} \frac{21}{4}\int_0^1\!\!\int_{x^2}^{x} x^2 y\,\mathrm{d}y\,\mathrm{d}x \overset{(2)}{=} \frac{21}{4}\int_0^1 x^2\,\frac{x^2-x^4}{2}\,\mathrm{d}x \overset{(3)}{=} \frac{21}{8}\Big(\frac{1}{5}-\frac{1}{7}\Big) = \frac{3}{20}.$$

(1) The event $\{X \geq Y\}$ intersected with the support is $\{0 \leq x \leq 1,\ x^2 \leq y \leq x\}$.
(2) $\int_{x^2}^{x} y\,\mathrm{d}y = \frac{1}{2}(x^2 - x^4)$.
(3) $\int_0^1 x^4 = \frac{1}{5}$, $\int_0^1 x^6 = \frac{1}{7}$, and $\frac{21}{8} \cdot \frac{2}{35} = \frac{3}{20}$.

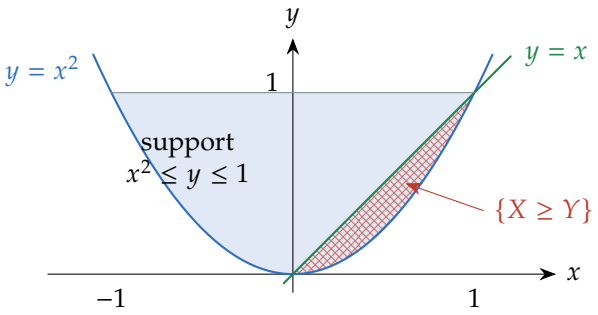


Figure 1.30: The density $\frac{21}{4}x^2y$ lives on the shaded region between the parabola $y = x^2$ and the line $y = 1$. The event $X \geq Y$ is the hatched sliver between the parabola and the diagonal $y = x$. To integrate, fix x and let y run up the vertical segment inside the region.

1.10.2 Marginal distributions: averaging out the nuisance

If we only care about X, the event "$X = x$" is the disjoint union of the events "$X = x$ and $Y = y$" over all values y. The law of total probability (first part of Chapter 1) adds them up.

Working definition 1.17 (marginal distribution)

The **marginal** mass function or density of X is

$$f_X(x) = \sum_y f(x,y) \quad \text{(discrete)}, \qquad f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y \quad \text{(continuous)},$$

and similarly $f_Y(y) = \sum_x f(x,y)$ or $\int f(x,y)\,\mathrm{d}x$ [sum or integrate the joint over every value of the variable we want to get rid of]. The variable summed over is *marginalised* or *integrated out*. *Operational test:* simulate pairs (x,y), throw away the y's, and histogram the x's.

We need a marginal distribution to isolate and understand the probability of a single variable by averaging out or removing "nuisance" variables from a complex joint distribution. The word "averaging" is meant literally. Write the joint as $f(x,y) = f(x \mid y) f_Y(y)$ (the product rule, derived below). Then the marginal $f_X(x) = \int f(x \mid y) f_Y(y) \mathrm{d}y$ is the average of the conditional densities $f(x \mid y)$, each weighted by how probable its nuisance value y is.

Why is the continuous marginal an integral over y [Cowan (1.20)–(1.21)]? Cut the vertical strip $[x, x + \mathrm{d}x]$ into small cells of height Δy at heights y_i:

$$\mathrm{P}\big(x \leq X \leq x + \mathrm{d}x\big) \overset{(1)}{=} \sum_i \mathrm{P}\big(x \leq X \leq x + \mathrm{d}x,\ y_i \leq Y \leq y_i + \Delta y\big)$$
$$\overset{(2)}{=} \sum_i f(x, y_i)\,\Delta y\,\mathrm{d}x \xrightarrow{\Delta y \to 0} \left[\int f(x,y)\,\mathrm{d}y\right]\mathrm{d}x.$$

(1) The cells are disjoint and together make up the strip (the law of total probability).
(2) The probability of a small cell is the joint density times its area.

The bracket is probability per unit x: the marginal density.

Example 1.38 (Margins of a table, and what they cannot tell you). For the table [Wasserman Ex. 2.24, p. 33]

	$Y=0$	$Y=1$	
$X=0$	$1/10$	$2/10$	$3/10$
$X=1$	$3/10$	$4/10$	$7/10$
	$4/10$	$6/10$	1

the marginal of X is the column of row sums, $f_X(0)=3/10$, $f_X(1)=7/10$, and the marginal of Y is the row of column sums. The name "marginal" literally means "written in the margin of the table".

Marginals lose information. The joint table $f(0,0)=\frac{1}{12}$, $f(1,0)=\frac{5}{12}$, $f(0,1)=f(1,1)=\frac{3}{12}$ has marginals $f_X=(\frac{1}{3},\frac{2}{3})$ and $f_Y=(\frac{1}{2},\frac{1}{2})$. So does the different table $f(0,0)=\frac{1}{6}$, $f(1,0)=\frac{1}{3}$, $f(0,1)=\frac{1}{6}$, $f(1,1)=\frac{1}{3}$ [CasellaBerger Ex. 4.1.9]. Knowing both marginals does not determine the joint. This is a small instance of a big theme: summarising discards information. A CMB power spectrum C_ℓ is a "marginal" summary that keeps the variance at each scale and throws away the phases, and Chapter 12 is about recovering some of what was thrown away.

Example 1.39 (Marginal densities by integration).

- $f(x,y)=e^{-(x+y)}$ for $x,y\ge 0$: $f_X(x)=e^{-x}\int_0^\infty e^{-y}\mathrm{d}y=e^{-x}$ [Wasserman Ex. 2.26].
- $f(x,y)=x+y$ on the unit square: $f_Y(y)=\int_0^1(x+y)\mathrm{d}x=\frac{1}{2}+y$ [Wasserman Ex. 2.27].
- $f(x,y)=\frac{21}{4}x^2y$ on $x^2\le y\le 1$ (Example 1.37):

$$f_X(x)\overset{(1)}{=}\frac{21}{4}x^2\int_{x^2}^1 y\,\mathrm{d}y=\frac{21}{8}x^2(1-x^4),\qquad -1\le x\le 1.$$

(1) At fixed x the density is nonzero only for $x^2\le y\le 1$, so those are the limits of the y integral [Wasserman Ex. 2.28, p. 34].

In the last case the limits of integration depend on x. Forgetting that is the commonest error in marginalisation.

1.10.3 The Gaussian case: marginalising by completing the square

The single most important joint density is the bivariate Gaussian. With zero means, standard deviations σ_x,σ_y and a correlation parameter $\rho\in(-1,1)$ (Section 1.13 shows that ρ is the correlation coefficient),

$$f(x,y)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{x^2}{\sigma_x^2}-\frac{2\rho xy}{\sigma_x\sigma_y}+\frac{y^2}{\sigma_y^2}\right]\right\}. \tag{1.28}$$

The contours of f are ellipses, tilted when $\rho\neq 0$. We need one integral, the Gaussian integral $\int_{-\infty}^{\infty}e^{-(y-m)^2/(2s^2)}\mathrm{d}y=\sqrt{2\pi}\,s$ (for any shift m and width $s>0$).

Example 1.40 (The marginal and the conditional of a bivariate Gaussian). The question: how is X distributed on its own, and how is Y distributed once $X=x$ is known? The plan is to rewrite the exponent so that y appears in a single square. Then the y integral becomes the Gaussian integral just recalled. First rearrange the bracket in (1.28):

$$\frac{x^2}{\sigma_x^2}-\frac{2\rho xy}{\sigma_x\sigma_y}+\frac{y^2}{\sigma_y^2}\overset{(1)}{=}\left(\frac{y}{\sigma_y}-\frac{\rho x}{\sigma_x}\right)^2+\frac{x^2}{\sigma_x^2}-\frac{\rho^2x^2}{\sigma_x^2}=\frac{\left(y-\rho\sigma_y x/\sigma_x\right)^2}{\sigma_y^2}+(1-\rho^2)\frac{x^2}{\sigma_x^2}.$$

(1) Expanding $(y/\sigma_y - \rho x/\sigma_x)^2$ gives $y^2/\sigma_y^2 - 2\rho xy/(\sigma_x\sigma_y) + \rho^2 x^2/\sigma_x^2$, so we must add back $(1-\rho^2)x^2/\sigma_x^2$ to restore the original expression.

Dividing by $2(1-\rho^2)$ and writing $m(x) = \rho\sigma_y x/\sigma_x$ and $s = \sigma_y\sqrt{1-\rho^2}$, the joint density factorises as

$$f(x,y) = \frac{1}{2\pi\sigma_x s}\exp\!\left[-\frac{(y-m(x))^2}{2s^2}\right]\exp\!\left[-\frac{x^2}{2\sigma_x^2}\right].$$

The marginal of X follows:

$$f_X(x) \overset{(1)}{=} \frac{e^{-x^2/2\sigma_x^2}}{2\pi\sigma_x s}\int_{-\infty}^{\infty} e^{-(y-m(x))^2/2s^2}\,\mathrm{d}y \overset{(2)}{=} \frac{e^{-x^2/2\sigma_x^2}}{2\pi\sigma_x s}\sqrt{2\pi}\,s = \frac{1}{\sqrt{2\pi}\,\sigma_x}e^{-x^2/2\sigma_x^2}.$$

(1) Definition of the marginal. The factor depending only on x comes out of the y integral.
(2) Gaussian integral with shift $m(x)$ and width s. The shift does not change the area, which is why the answer is so simple.

So $X \sim \mathcal{N}(0,\sigma_x^2)$ whatever ρ is, and by the symmetry of (1.28), $Y \sim \mathcal{N}(0,\sigma_y^2)$. The conditional density of Y at fixed $X = x$ is the joint divided by the marginal (defined in the next subsection):

$$f(y\mid x) = \frac{f(x,y)}{f_X(x)} = \frac{1}{\sqrt{2\pi}\,s}\exp\!\left[-\frac{(y-\rho\sigma_y x/\sigma_x)^2}{2s^2}\right], \qquad Y\mid X=x \;\sim\; \mathcal{N}\!\left(\rho\frac{\sigma_y}{\sigma_x}x,\; \sigma_y^2(1-\rho^2)\right).$$

Knowing X does two things to Y: it shifts the centre along the line $y = \rho\sigma_y x/\sigma_x$ (the "regression line"), and it shrinks the width by the factor $\sqrt{1-\rho^2}$. For $\rho = 0.7$ that factor is 0.71.

Watch out: Marginal width is not conditional width

In a two-parameter fit, the error bar on one parameter with the other held fixed at its best value is the *conditional* width $\sigma_y\sqrt{1-\rho^2}$. The error bar with the other parameter marginalised is σ_y. For strongly correlated parameters these differ a lot: $\rho = 0.95$ gives a factor 3.2. Quoting the conditional error as if it were the marginal one is a classic way to overstate a result. Chapter 10 meets the same distinction as $1/\sqrt{\mathbf{F}_{ii}}$ versus $\sqrt{(\mathbf{F}^{-1})_{ii}}$.

1.10.4 Conditional distributions: slicing and renormalising

Slicing the scatter plot at $Y = y$ and renormalising the slice gives the definition.

Working definition 1.18 (conditional mass function and density)

For discrete variables, the **conditional mass function** of X given $Y = y$ is

$$f_{X|Y}(x\mid y) = \mathrm{P}(X = x \mid Y = y) = \frac{f(x,y)}{f_Y(y)}, \qquad f_Y(y) > 0,$$

the ordinary conditional probability of the first part of Chapter 1. For continuous variables the **conditional density** is defined by the same formula [the joint along the line $Y = y$, divided by its total, so that it integrates to 1 over x], and $\mathrm{P}(X \in A \mid Y = y) = \int_A f_{X|Y}(x\mid y)\,\mathrm{d}x$. *Operational test:* keep only simulated pairs with y in a thin band around y_0, histogram their x values and normalise.

In the continuous case the event $\{Y = y\}$ has probability zero. So conditioning on it cannot literally mean $\mathrm{P}(A\cap B)/\mathrm{P}(B)$, which would be $0/0$. Why is the definition still right? Think of a thin band of width $\mathrm{d}y$ around y. Both numerator and denominator carry the same factor

$\mathrm{d}y$, which cancels and leaves the ratio of densities [Cowan (1.23)]. (Wasserman [Wasserman p. 36, footnote 6] calls conditioning on a probability-zero event "treading in deep water" and simply adopts the formula as the definition.) Rearranging the definition gives the product rule

$$f(x,y) = f_{X|Y}(x \mid y)\, f_Y(y) = f_{Y|X}(y \mid x)\, f_X(x), \tag{1.29}$$

and from it the continuous law of total probability and Bayes' theorem [Cowan (1.26)–(1.28)]:

$$f_X(x) = \int f_{X|Y}(x \mid y) f_Y(y)\, \mathrm{d}y, \qquad f_{Y|X}(y \mid x) = \frac{f_{X|Y}(x \mid y)\, f_Y(y)}{\int f_{X|Y}(x \mid y') f_Y(y')\, \mathrm{d}y'}. \tag{1.30}$$

The first is "sum over the branches of a tree", with a continuum of branches. The second reads the tree backwards. With y a parameter and x the data, $f_{X|Y}$ is the prediction and $f_{Y|X}$ the inference: this is the equation Chapter 5 is built on.

Example 1.41 (Conditional probabilities from a density).

(a) $f(x,y) = x + y$ on the unit square, with $f_Y(y) = y + \frac{1}{2}$ (Example 1.39). Then $f(x \mid y) = (x+y)/(y+\frac{1}{2})$ and [Wasserman Ex. 2.38, p. 37]

$$\mathrm{P}\big(X < \tfrac{1}{4} \mid Y = \tfrac{1}{3}\big) \overset{(1)}{=} \int_0^{1/4} \frac{x + \frac{1}{3}}{\frac{1}{3} + \frac{1}{2}}\, \mathrm{d}x \overset{(2)}{=} \frac{6}{5}\Big[\frac{x^2}{2} + \frac{x}{3}\Big]_0^{1/4} = \frac{6}{5}\Big(\frac{1}{32} + \frac{1}{12}\Big) = \frac{11}{80}.$$

(1) Integrate the conditional density over the event, at $y = \frac{1}{3}$.
(2) $1/(\frac{1}{3} + \frac{1}{2}) = \frac{6}{5}$, and the antiderivative is elementary. Then $\frac{1}{32} + \frac{1}{12} = \frac{11}{96}$.

(b) Build a joint from a marginal and a conditional: $X \sim \mathrm{Uniform}(0,1)$ and then $Y \mid X = x \sim \mathrm{Uniform}(x,1)$ [Wasserman Ex. 2.39]. Then $f(y \mid x) = 1/(1-x)$ for $x < y < 1$, and by (1.29) $f(x,y) = 1/(1-x)$ on $0 < x < y < 1$. The marginal of Y is

$$f_Y(y) \overset{(1)}{=} \int_0^y \frac{\mathrm{d}x}{1-x} \overset{(2)}{=} -\int_1^{1-y} \frac{\mathrm{d}u}{u} = -\ln(1-y), \qquad 0 < y < 1.$$

(1) At fixed y, the joint is nonzero for $0 < x < y$.
(2) Substitute $u = 1 - x$, $\mathrm{d}u = -\mathrm{d}x$; the limits $x = 0, y$ become $u = 1, 1-y$.

The density of Y rises towards $y = 1$: each stage pushes Y upward.

(c) For $f = \frac{21}{4}x^2 y$ on $x^2 \le y \le 1$, dividing by the marginal of Example 1.39 gives $f(y \mid x) = 2y/(1-x^4)$ on $x^2 \le y \le 1$. At $x = \frac{1}{2}$ this is $32y/15$, and [Wasserman Ex. 2.40, p. 38]

$$\mathrm{P}\big(Y \ge \tfrac{3}{4} \mid X = \tfrac{1}{2}\big) \overset{(1)}{=} \int_{3/4}^1 \frac{32y}{15}\, \mathrm{d}y = \frac{16}{15}\Big(1 - \frac{9}{16}\Big) = \frac{7}{15}.$$

(1) The conditional density at $x = \frac{1}{2}$ lives on $\frac{1}{4} \le y \le 1$, and the event asks for the top part.

1.10.5 Independence

The last question is whether slicing changes anything at all. If every slice looks the same, learning one variable teaches us nothing about the other.

Working definition 1.19 (independent random variables)

X and Y are **independent**, written $X \perp Y$, if for every pair of sets A, B

$$\mathrm{P}(X \in A,\ Y \in B) = \mathrm{P}(X \in A)\ \mathrm{P}(Y \in B)$$

[every event about X is independent of every event about Y]. Equivalently the joint factorises into the marginals, $f(x,y) = f_X(x) f_Y(y)$ for all x, y [Wasserman Thm 2.30, p. 35], and then $f(y \mid x) = f_Y(y)$: learning x does not change the distribution of Y. *Operational test:* slice the scatter plot at several x values. If the normalised y-histograms in the slices all look the same, and the same as the marginal, the variables are independent.

Example 1.42 (Checking independence).

(a) The table with all four entries $\frac{1}{4}$ is independent: $f_X = f_Y = (\frac{1}{2}, \frac{1}{2})$ and every cell equals $\frac{1}{2} \cdot \frac{1}{2}$. The table $f(0,0) = f(1,1) = \frac{1}{2}$, $f(0,1) = f(1,0) = 0$ has the *same* marginals but is not independent: $f_X(0) f_Y(1) = \frac{1}{4} \neq 0 = f(0,1)$ [Wasserman Ex. 2.31].

(b) X, Y independent with the same density $f(x) = 2x$ on $[0,1]$. The joint is $4xy$ on the unit square, and [Wasserman Ex. 2.32]

$$\mathrm{P}(X + Y \le 1) \overset{(1)}{=} 4 \int_0^1 x \Big[\int_0^{1-x} y \, \mathrm{d}y \Big] \mathrm{d}x \overset{(2)}{=} 2 \int_0^1 x(1-x)^2 \, \mathrm{d}x \overset{(3)}{=} 2\Big(\frac{1}{2} - \frac{2}{3} + \frac{1}{4}\Big) = \frac{1}{6}.$$

(1) The event is the triangle below the line $y = 1 - x$; for each x, y runs from 0 to $1 - x$.
(2) $\int_0^{1-x} y \, \mathrm{d}y = \frac{1}{2}(1-x)^2$.
(3) Expand $x(1-x)^2 = x - 2x^2 + x^3$ and integrate term by term.

(c) Shortcut [Wasserman Thm 2.33, p. 36]: if the joint is nonzero on a rectangle (possibly infinite) and has the form $g(x)h(y)$, the variables are independent. $f = 2e^{-(x+2y)}$ on $x, y > 0$ is $g(x)h(y)$ with $g = 2e^{-x}$, $h = e^{-2y}$, so $X \perp Y$.

(d) The rectangle condition matters. $\frac{21}{4} x^2 y$ on $x^2 \le y \le 1$ looks like $g(x)h(y)$, but the region is not a rectangle, and $f(y \mid x) = 2y/(1 - x^4)$ depends on x: learning x tells us $y \ge x^2$. X and Y are dependent.

For more than two variables the definitions carry over. A **random vector** $\boldsymbol{X} = (X_1, \ldots, X_n)$ has a joint density $f(x_1, \ldots, x_n)$; marginals integrate out any subset of components; $X_1, \ldots, X_n$ are independent if $f = \prod_i f_{X_i}(x_i)$. If they are independent *and* all have the same marginal F they are **iid** (independent and identically distributed), written $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$, and form a **random sample** of size n from F [Wasserman Def. 2.41, p. 39]. Repeated measurements, Monte Carlo draws and the pixels of a white-noise map are the standard iid examples.

Two multivariate models recur throughout the book [Wasserman §2.10]. The **multinomial** counts how n independent draws fall into k categories (histogram bins) with probabilities $p_1, \ldots, p_k$. The marginal of one count X_j is Binomial(n, p_j): lump all the other categories into "not j" and each draw becomes a Bernoulli(p_j) trial. The **multivariate normal** $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ has density

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{k/2} \, |\Sigma|^{1/2}} \exp\Big\{ -\tfrac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^{\mathsf{T}} \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \Big\}, \tag{1.31}$$

with Σ a symmetric positive-definite $k \times k$ matrix. For $k = 2$ and $\Sigma = \left(\begin{smallmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{smallmatrix} \right)$ it reduces to

(1.28):

$$|\Sigma| \overset{(1)}{=} \sigma_x^2\sigma_y^2(1-\rho^2), \qquad \Sigma^{-1} \overset{(2)}{=} \frac{1}{\sigma_x^2\sigma_y^2(1-\rho^2)}\begin{pmatrix}\sigma_y^2 & -\rho\sigma_x\sigma_y\\ -\rho\sigma_x\sigma_y & \sigma_x^2\end{pmatrix},$$

$$\boldsymbol{x}^{\mathsf{T}}\Sigma^{-1}\boldsymbol{x} \overset{(3)}{=} \frac{\sigma_y^2x^2 - 2\rho\sigma_x\sigma_y xy + \sigma_x^2y^2}{\sigma_x^2\sigma_y^2(1-\rho^2)} = \frac{1}{1-\rho^2}\left[\frac{x^2}{\sigma_x^2} - \frac{2\rho xy}{\sigma_x\sigma_y} + \frac{y^2}{\sigma_y^2}\right].$$

(1) Determinant of a 2×2 matrix, $ad - bc$.
(2) Inverse of a 2×2 matrix: swap the diagonal, negate the off-diagonal, divide by the determinant.
(3) Multiply out $(x,y)\,\Sigma^{-1}(x,y)^{\mathsf{T}}$, then divide numerator and denominator by $\sigma_x^2\sigma_y^2$.

The general marginal and conditional rules [Wasserman Thm 2.44, p. 40], $X_b \mid X_a = x_a \sim \mathcal{N}\big(\mu_b + \Sigma_{ba}\Sigma_{aa}^{-1}(x_a-\mu_a),\ \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}\big)$, reduce for $k=2$ to exactly what we derived in Example 1.40: $\Sigma_{ba}\Sigma_{aa}^{-1} = \rho\sigma_x\sigma_y/\sigma_x^2 = \rho\sigma_y/\sigma_x$ and $\Sigma_{bb} - \Sigma_{ba}^2/\Sigma_{aa} = \sigma_y^2(1-\rho^2)$. The general proofs are in Chapter 2.

1.10.6 Seeing it: projections, slices, and the smeared spectrum

The first script samples the bivariate Gaussian with $\sigma_x = 1$, $\sigma_y = 2$, $\rho = 0.7$, projects, and slices at $x_0 = 1$.

Listing 1.12: Marginal by projection, conditional by slicing (`code/ch01/12_marginal_gaussian.py`)

```
23 sx, sy, rho = 1.0, 2.0, 0.7
24 C = np.array([[sx**2, rho * sx * sy], [rho * sx * sy, sy**2]])
25 N = 200_000
26 XY = rng.multivariate_normal([0.0, 0.0], C, size=N)
27 X, Y = XY[:, 0], XY[:, 1]
28
29 # thin band at x0
30 x0, half = 1.0, 0.05
31 band = np.abs(X - x0) < half
32 Yb = Y[band]
33 cond_mean = rho * sy / sx * x0
34 cond_sd = sy * np.sqrt(1 - rho**2)
```

Line 26 draws 2×10^5 pairs from $\mathcal{N}(\mathbf{0}, \mathbf{C})$. The marginal needs no code at all: `Y` on its own *is* a sample from f_Y (line 27 just splits the columns). The conditional is a boolean mask, the band $|x-1| < 0.05$ (line 31), exactly the "conditioning is selecting a subset" of the first part of Chapter 1.

The simulated marginal widths are 1.001 and 2.006 (exact 1 and 2). The band holds 4855 points, whose y values have mean 1.404 and standard deviation 1.426, against the exact conditional $\rho\sigma_y x_0/\sigma_x = 1.400$ and $\sigma_y\sqrt{1-\rho^2} = 1.428$ (the band has finite width, which adds a tiny extra spread).

Example 1.43 (The measured spectrum: marginalising the true energy). Return to the photon in the calorimeter from the start of this section. Its true energy E is never observed. Only the measured energy $E_{\text{meas}} = E + \text{noise}$ is recorded, and we want the spectrum of E_{meas}. Let the true energy be exponential, $f(E) = e^{-E/\beta}/\beta$ for $E > 0$ with $\beta = 20\,\text{GeV}$ (a falling spectrum), and the detector Gaussian, $f(x \mid E) = \exp[-(x-E)^2/2\sigma^2]/(\sqrt{2\pi}\sigma)$ with $\sigma = 5\,\text{GeV}$, where $x = E_{\text{meas}}$. The joint is the product (1.29), and the spectrum we histogram is the marginal

$$f(x) = \int_0^\infty \frac{e^{-E/\beta}}{\beta}\,\frac{e^{-(x-E)^2/2\sigma^2}}{\sqrt{2\pi}\,\sigma}\,\mathrm{d}E. \tag{1.32}$$

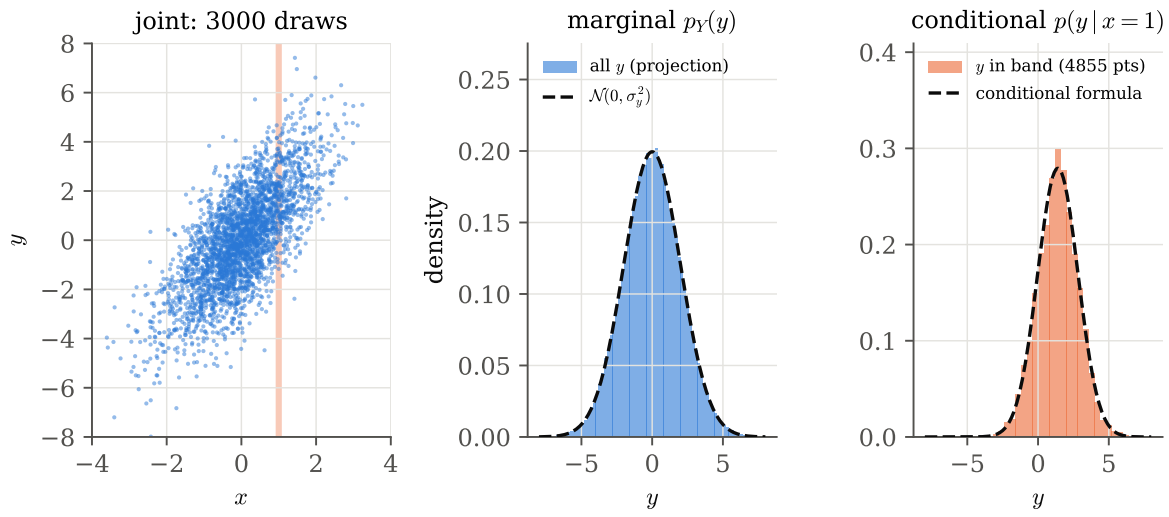


Figure 1.31: Left: 3000 draws from the correlated Gaussian, with the thin band at $x = 1$ shaded. Middle: all y values, projected onto the y axis, follow the marginal $\mathcal{N}(0, 4)$. Right: only the y values from inside the band, renormalised, follow the narrower and shifted conditional density.

To do the integral, collect the E-dependence of the exponent into one square, so that what is left is a Gaussian integral:

$$-\frac{E}{\beta} - \frac{(x-E)^2}{2\sigma^2} \overset{(1)}{=} -\frac{E^2 - 2E\left(x - \sigma^2/\beta\right) + x^2}{2\sigma^2} \overset{(2)}{=} -\frac{\left(E-m\right)^2}{2\sigma^2} + \frac{m^2 - x^2}{2\sigma^2}$$
$$\overset{(3)}{=} -\frac{\left(E-m\right)^2}{2\sigma^2} - \frac{x}{\beta} + \frac{\sigma^2}{2\beta^2},$$

with $m = x - \sigma^2/\beta$.

(1) Put both terms over $2\sigma^2$: $E/\beta = 2\sigma^2 E/(2\sigma^2\beta)$, and expand $(x-E)^2$.
(2) Complete the square, $E^2 - 2Em = (E-m)^2 - m^2$.
(3) $m^2 - x^2 = -2x\sigma^2/\beta + \sigma^4/\beta^2$; divide by $2\sigma^2$.

The last two terms do not involve E and leave the integral. What remains is a Gaussian in E centred at m, integrated over $E > 0$ only:

$$f(x) \overset{(1)}{=} \frac{1}{\beta}\, e^{\sigma^2/2\beta^2 - x/\beta} \int_0^\infty \frac{e^{-(E-m)^2/2\sigma^2}}{\sqrt{2\pi}\,\sigma}\,\mathrm{d}E \overset{(2)}{=} \frac{1}{\beta}\, e^{\sigma^2/2\beta^2 - x/\beta}\, \Phi\!\left(\frac{x - \sigma^2/\beta}{\sigma}\right).$$

(1) Insert the rearranged exponent into (1.32).
(2) The remaining integral is $\mathrm{P}(W > 0)$ for $W \sim \mathcal{N}(m, \sigma^2)$, which is $\mathrm{P}\big((W-m)/\sigma > -m/\sigma\big) = 1 - \Phi(-m/\sigma) = \Phi(m/\sigma)$ by the symmetry of the standard Gaussian.

The formula shows two things. First, far above the resolution scale, $x \gg \sigma^2/\beta + \sigma$, we have $\Phi \to 1$. There the measured spectrum is the true exponential with the same slope, multiplied by the constant $e^{\sigma^2/2\beta^2} > 1$. Smearing "feeds" events from low to high energy, and so raises the tail. Second, near and below zero the Φ factor cuts in. There is a finite probability of measuring a *negative* energy, which no true photon has.

Listing 1.13: Simulating the nuisance, then forgetting it (`code/ch01/13_detector_smearing.py`)

```
beta, sigma = 20.0, 5.0
N = 500_000
```

```
26 E_true = rng.exponential(beta, size=N)                    # the nuisance: never observed
27 E_meas = E_true + rng.normal(0.0, sigma, size=N)          # what the detector reports
28
29
30 def p_meas(x):
31     """Marginal density of the measured energy (true energy integrated out)."""
32     return (1 / beta) * np.exp(sigma**2 / (2 * beta**2) - x / beta) * \
33         stats.norm.cdf((x - sigma**2 / beta) / sigma)
```

Lines 26–27 are the whole physics: draw the true energy, add the noise. Marginalising over the nuisance then costs nothing: we simply never look at `E_true` again. The function `p_meas` (lines 30–33) is the closed form just derived.

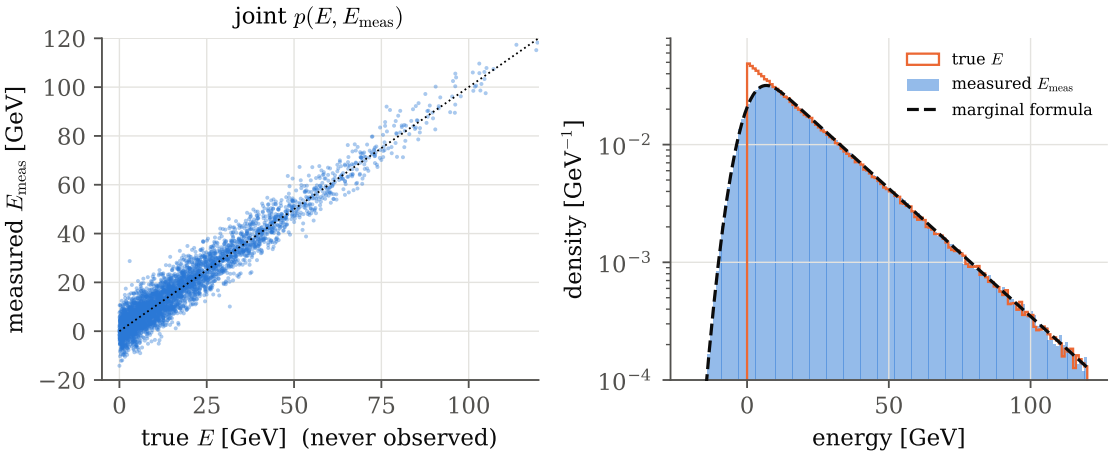


Figure 1.32: Left: the joint distribution of (true, measured) energy, scattered around the diagonal. Only the vertical coordinate is ever recorded. Right, on a log scale: the true spectrum (orange outline) is a straight line starting at zero. The measured spectrum (blue) is its marginal with the true energy integrated out. It leaks below zero and runs parallel to, and slightly above, the true line in the tail. The dashed curve is the closed form.

With 5×10^5 photons the simulation gives a mean measured energy of 19.99 GeV (exact $\beta = 20$), a variance of 425.0 GeV^2 (exact $\beta^2 + \sigma^2 = 425.0$, derived in Section 1.12), and a fraction 0.0866 of negative measured energies (exact $\int_{-\infty}^{0} f = 0.0860$). A χ^2 comparison of the histogram with the bin-integrated formula gives 125 for 136 bins, which is what agreement looks like (Chapter 4 makes this precise).

What this bought us. Marginalising a nuisance variable is, in simulation, the act of not looking at it. Analytically it is an integral over the joint, $f(x) = \int f(x \mid E) f(E)\, \mathrm{d}E$. The same integral with E a parameter is the Bayesian evidence of Chapter 5, and with E the true sky it is the forward model of Chapter 8.

How this is used in real research: Verification here, forward modelling in research

The closed form above *verifies* the simulation. In a real calorimeter the response $f(x \mid E)$ has non-Gaussian tails, depends on position and is known only through test-beam data and detector simulation. Then the marginal $\int f(x \mid E) f(E \mid \theta)\, \mathrm{d}E$ has no closed form and is computed by exactly the two lines of code above, repeated for each theory parameter θ. That forward-folded spectrum is what a fit compares with data. Going the other way, from x back to E, is *unfolding*, a conditional distribution $f(E \mid x)$ computed with (1.30).

1.11 Functions of random variables: change of variables and Jacobians

A classical harmonic oscillator swings as $x(t) = A\cos\omega t$. We take a photograph at a random moment, with no idea of the phase. Where is the bob most likely to be caught? The motion is completely deterministic; the only randomness is *when* we look, and the time t is uniform over a period. The position is a function of that uniform time, $X = A\cos\omega T$, so its distribution is fixed by the distribution of T and the function. We will find that the bob is most likely near the turning points, where it moves slowly, and least likely at the centre, where it is fastest.

The same question appears whenever an observable is a function of something whose distribution we know: the angle of a decay product, the transverse momentum computed from that angle, a flux from a magnitude, the logarithm of a likelihood. In each case we know the density of X and want the density of $Y = r(X)$. The tool is the Jacobian, and it works for several variables at once. The oscillator has a second use: histogramming $x(t)$ at random times is the same act as histogramming samples from a posterior (Chapter 5).

The key idea is that probability is conserved when we relabel. Think of probability as a fluid of total mass 1 spread along the x axis. The map $y = r(x)$ moves each drop of fluid to a new place but creates and destroys nothing. Whatever mass sat in $[x, x + \mathrm{d}x]$ now sits in the image interval $[y, y + \mathrm{d}y]$. If the map stretches that interval ($|\mathrm{d}y| > |\mathrm{d}x|$), the same mass is spread thinner and the density drops. If it compresses it, the density rises. For the oscillator the "map" is from time to position: near the turning points a long stretch of time maps onto a short stretch of position, so the density of position is high there. This is the same bookkeeping as a number density $n(\boldsymbol{x})$ under a change of coordinates, or a phase-space density under a canonical transformation.

In practice we find the distribution of $Y = r(X)$ in three steps.

1. For a value y, which values of x produce $Y \le y$? Draw it.
2. The probability of $Y \le y$ is the probability of that set of x's. That is the CDF of Y.
3. Differentiate to get the density. If r is one-to-one, this collapses to "old density times $|\mathrm{d}x/\mathrm{d}y|$". If several x's give the same y, add their contributions.

1.11.1 Discrete variables: collect the preimages

For a discrete X, $Y = r(X)$ is discrete and $\mathrm{P}(Y = y) = \mathrm{P}\big(X \in r^{-1}(y)\big)$, where $r^{-1}(y)$ is the set of all x with $r(x) = y$. If $\mathrm{P}(X = -1) = \mathrm{P}(X = 1) = \frac{1}{4}$, $\mathrm{P}(X = 0) = \frac{1}{2}$ and $Y = X^2$, then $\mathrm{P}(Y = 0) = \mathrm{P}(X = 0) = \frac{1}{2}$ and $\mathrm{P}(Y = 1) = \mathrm{P}(X = 1) + \mathrm{P}(X = -1) = \frac{1}{2}$ [Wasserman Ex. 2.45, p. 41]. Y takes fewer values than X because r is not one-to-one.

1.11.2 Continuous variables: the CDF method and the Jacobian

For a continuous variable the same three steps work with integrals instead of sums, and when r is one-to-one they collapse into a single formula.

Working definition 1.20 (change of variables for one variable)

Let X have density f_X and $Y = r(X)$. The **CDF method** computes

$$F_Y(y) = \mathrm{P}\big(r(X) \le y\big) = \int_{A_y} f_X(x)\,\mathrm{d}x, \qquad A_y = \{x : r(x) \le y\}, \qquad f_Y = F_Y'.$$

If r is strictly monotone with inverse $x = s(y)$, this gives

$$f_Y(y) = f_X\big(s(y)\big)\left|\frac{\mathrm{d}s}{\mathrm{d}y}\right|, \tag{1.33}$$

[the old density evaluated at the x that maps to y, times the factor by which the map squeezes lengths near that point]. If several x_k map to the same y (with r monotone near each), the contributions add: $f_Y(y) = \sum_k f_X(x_k)\,|\mathrm{d}x_k/\mathrm{d}y|$. *Operational test:* simulate X, apply r, histogram Y. The histogram must follow the formula.

Why does (1.33) hold [CasellaBerger Thm 2.1.5]? Take r increasing first. Then s is increasing too, and $r(X) \le y \iff X \le s(y)$:

$$f_Y(y) \overset{(1)}{=} \frac{\mathrm{d}}{\mathrm{d}y}\,\mathrm{P}\big(X \le s(y)\big) = \frac{\mathrm{d}}{\mathrm{d}y}F_X\big(s(y)\big) \overset{(2)}{=} f_X\big(s(y)\big)\,\frac{\mathrm{d}s}{\mathrm{d}y}.$$

(1) $F_Y(y) = \mathrm{P}(r(X) \le y)$, and for increasing r the event is $\{X \le s(y)\}$.
(2) Chain rule, with $F_X' = f_X$.

If r is decreasing, $r(X) \le y \iff X \ge s(y)$, so $F_Y(y) = 1 - F_X(s(y))$ and $f_Y = -f_X(s)\,s'$. Now $s' < 0$, so $-s' = |s'|$. Both cases are (1.33). The absolute value is there because densities are positive whichever way the map runs [Cowan (1.30)–(1.32)]. In the fluid picture: $f_Y\,|\mathrm{d}y| = f_X\,|\mathrm{d}x|$, "mass in = mass out".

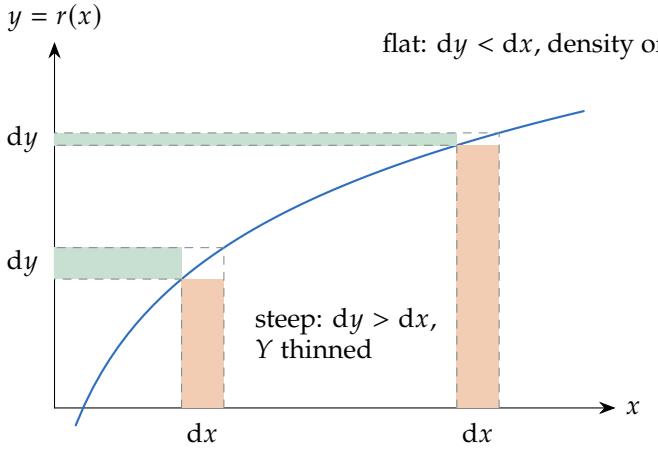


Figure 1.33: The Jacobian rule. Two strips of equal width $\mathrm{d}x$ carry probability $f_X\mathrm{d}x$ up to the curve and across to the y axis. Where the curve is steep (left) the strip lands on a wide interval $\mathrm{d}y$ and the density of Y is diluted. Where it is flat (right) the same width lands on a narrow interval and the density is concentrated. The factor is $|\mathrm{d}x/\mathrm{d}y| = 1/|r'(x)|$.

Example 1.44 (Standardising a Gaussian, and changing units). Let $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Z = (X - \mu)/\sigma$. The inverse is $x = s(z) = \mu + \sigma z$ with $\mathrm{d}s/\mathrm{d}z = \sigma > 0$, so

$$f_Z(z) \overset{(1)}{=} f_X(\mu + \sigma z)\,\sigma \overset{(2)}{=} \frac{\sigma}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(\sigma z)^2}{2\sigma^2}\right] = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}.$$

(1) (1.33) with $|s'| = \sigma$.
(2) Substitute $x - \mu = \sigma z$ in the Gaussian density.

So $Z \sim \mathcal{N}(0,1)$, the fact used in Example 1.33. Reversed, $X = \mu + \sigma Z$ is $\mathcal{N}(\mu, \sigma^2)$ [Wasserman p. 28]. A change of units is the same calculation: $E_{\mathrm{MeV}} = 1000\,E_{\mathrm{GeV}}$ has $|\mathrm{d}E_{\mathrm{GeV}}/\mathrm{d}E_{\mathrm{MeV}}| = 10^{-3}$, so every density value is divided by 1000, which is what we found by hand in Example 1.31.

Example 1.45 (The logarithm of an exponential, and what does not transform). Let $X \sim \mathrm{Exp}(1)$, $F_X(x) = 1 - e^{-x}$, and $Y = \ln X$ [Wasserman Ex. 2.46, p. 41]. By the CDF method,

$$F_Y(y) \overset{(1)}{=} \mathrm{P}(\ln X \le y) \overset{(2)}{=} \mathrm{P}(X \le e^y) = 1 - e^{-e^y}, \qquad f_Y(y) \overset{(3)}{=} e^y e^{-e^y}, \quad y \in \mathbb{R}.$$

(1) Definition of the CDF of Y.
(2) ln is increasing, so $\ln X \le y \iff X \le e^y$. Then use F_X.
(3) Differentiate: $\frac{\mathrm{d}}{\mathrm{d}y}(-e^{-e^y}) = e^{-e^y} \cdot e^y$.

The monotone formula gives the same thing directly: $s(y) = e^y$, $s' = e^y$, $f_X(e^y) = e^{-e^y}$.

Now compare summaries. The point: a change of variables carries the quantiles along, but it moves the mode. The *median* of X is $\ln 2$. The median of Y solves $1 - e^{-e^y} = \frac{1}{2}$, that is $e^y = \ln 2$, so it is $\ln(\ln 2)$, the log of the median of X. This holds for every quantile and any increasing map, because $\mathrm{P}(Y \le r(x)) = \mathrm{P}(X \le x)$. The *mode* behaves differently. The mode of X is 0 (the density e^{-x} is largest there), whose log is $-\infty$. The mode of Y is where $\frac{\mathrm{d}}{\mathrm{d}y}(y - e^y) = 0$, at $y = 0$, that is $x = 1$. The peak moves because the Jacobian reshapes the curve. Maximum a posteriori estimates therefore depend on the parametrisation, and quantile-based intervals do not (Chapter 5).

Example 1.46 (A square: two branches and two regimes). Let $X \sim \mathrm{Uniform}(-1, 3)$, density $\frac{1}{4}$, and $Y = X^2$, with values in $(0, 9)$ [Wasserman Ex. 2.47, p. 42]. The set $A_y = \{x : x^2 \le y\}$ is $[-\sqrt{y}, \sqrt{y}]$, intersected with $(-1, 3)$.

- $0 < y < 1$: $A_y = [-\sqrt{y}, \sqrt{y}]$ lies inside $(-1, 3)$, so $F_Y(y) = 2\sqrt{y}/4 = \frac{1}{2}\sqrt{y}$ and $f_Y = \frac{1}{4\sqrt{y}}$.
- $1 \le y < 9$: $-\sqrt{y} \le -1$, so $A_y \cap (-1, 3) = (-1, \sqrt{y}]$ and $F_Y(y) = (\sqrt{y} + 1)/4$, $f_Y = \frac{1}{8\sqrt{y}}$.

The density halves at $y = 1$ because below 1 two values of x (one negative, one positive) feed each y, and above 1 only the positive one does. The general version is the sum over branches [CasellaBerger Thm 2.1.8]: for $Y = X^2$ with any density, $f_Y(y) = \frac{1}{2\sqrt{y}}\left[f_X(\sqrt{y}) + f_X(-\sqrt{y})\right]$. For $X \sim \mathcal{N}(0, 1)$,

$$f_Y(y) \overset{(1)}{=} \frac{1}{2\sqrt{y}} \cdot 2 \cdot \frac{e^{-y/2}}{\sqrt{2\pi}} = \frac{1}{\sqrt{2\pi y}}\, e^{-y/2}, \qquad y > 0,$$

(1) Both branches $x = \pm\sqrt{y}$ have $|\mathrm{d}x/\mathrm{d}y| = 1/(2\sqrt{y})$ and, by symmetry, the same density $e^{-y/2}/\sqrt{2\pi}$.

which is the χ^2 density with one degree of freedom. It diverges at $y = 0$ (but integrably), and the sum of ν such squares gives the χ^2_ν of Chapter 2.

1.11.3 The oscillator: histogramming a deterministic motion

A mass m on a spring moves as $x(t) = A\cos\omega t$, $p(t) = -m\omega A \sin\omega t$, with period $T = 2\pi/\omega$. Nothing about this motion is random: once A is fixed, $x(t)$ is known at every instant. Yet there is an everyday question about it whose answer is a probability density:

If I look at the oscillator at a randomly chosen instant, where will I find it?

The probability cannot come from the equation of motion, which has none in it. It comes from four words of the question, "a randomly chosen instant". They sound innocent, but they carry

an assumption: the act of looking is equally likely to happen at every instant of the period, no part of the swing preferred,

$$P(t)\,\mathrm{d}t = \frac{\mathrm{d}t}{T}, \qquad 0 \le t < T.$$

This rule, which says *what* is chosen at random and how, is the **sampling prescription**. The rule is uniform; the motion is not. Why does the question make the instant uniform, and not something else? Because choosing the instant is what picks the state. At fixed energy the possible states of the oscillator are the points (x, p) on its phase-space orbit, so here the orbit is the sample space. The motion passes through each point of the orbit at exactly one instant of the period, so a random instant picks a random point of the orbit (Figure 1.35b). Once the instant is chosen, the dynamics carry it to its point, and that step has nothing random in it. The only probabilistic input is therefore the uniform t. Everything else (Method 2 below) is a change of variables from t to x. This is also why sampling is easy here. We never have to build the distribution on phase space: the motion walks us over it, and we only choose the time. To find the distribution of position, we histogram x for a large set of random values of t. We now find the density of x in three ways.

Method 1: the CDF. For $-A < x < A$, $\cos\omega t \le x/A$ holds for ωt in $[\arccos(x/A),\, 2\pi - \arccos(x/A)]$, an interval of length $2\pi - 2\arccos(x/A)$ out of 2π. So

$$F_X(x) \overset{(1)}{=} 1 - \frac{1}{\pi}\arccos\frac{x}{A} \overset{(2)}{=} \frac{1}{2} + \frac{1}{\pi}\arcsin\frac{x}{A}, \qquad f_X(x) \overset{(3)}{=} \frac{1}{\pi}\,\frac{1/A}{\sqrt{1 - x^2/A^2}} = \frac{1}{\pi\sqrt{A^2 - x^2}}.$$

(1) ωt is uniform on $[0, 2\pi)$, so the probability is the fraction of the circle: $(2\pi - 2\arccos(x/A))/2\pi$.
(2) $\arccos u = \frac{\pi}{2} - \arcsin u$.
(3) $\frac{\mathrm{d}}{\mathrm{d}u}\arcsin u = 1/\sqrt{1-u^2}$ and the chain rule with $u = x/A$.

Method 2: uniform time, then change variable from t to x. The sampling prescription of the question says that the time is uniformly distributed over one period,

$$P(t)\,\mathrm{d}t = \frac{\mathrm{d}t}{T}, \qquad 0 \le t < T. \tag{1.34}$$

Everything else is a change of variable. The map $t \mapsto x = A\cos\omega t$ is two-to-one: every $x \in (-A, A)$ is passed twice per period, once on the way out ($p < 0$) and once on the way back ($p > 0$). The probability of landing in $[x, x + \mathrm{d}x]$ is the probability of the two time intervals that map onto it, each of length $\mathrm{d}t = \mathrm{d}x/|v|$:

$$P(x)\,\mathrm{d}x \overset{(1)}{=} 2 \times \frac{1}{T}\left|\frac{\mathrm{d}t}{\mathrm{d}x}\right|\mathrm{d}x \overset{(2)}{=} \frac{2}{T}\,\frac{\mathrm{d}x}{\omega\sqrt{A^2 - x^2}} \overset{(3)}{=} \frac{\mathrm{d}x}{\pi\sqrt{A^2 - x^2}}.$$

(1) (1.34) applied to the two branches, which contribute equally; this is the Jacobian rule of Eq. (1.33) summed over preimages.
(2) $|\mathrm{d}x/\mathrm{d}t| = |v| = A\omega\,|\sin\omega t| = \omega\sqrt{A^2 - x^2}$, using $\sin^2 + \cos^2 = 1$ (equivalently energy conservation, $\frac{1}{2}mv^2 + \frac{1}{2}m\omega^2x^2 = \frac{1}{2}m\omega^2A^2$).
(3) $T = 2\pi/\omega$.

The probability of finding the mass in $\mathrm{d}x$ is the fraction of the period it spends there, $\propto \mathrm{d}x/|v|$. It is slow near the turning points, so it is found there most often.

Change the rule, change the answer. The arcsine law is the answer to the question about a random *instant*. Ask a different question about the same oscillator: pick a *position* at random, uniformly on $(-A, A)$, so that $P(x) = 1/(2A)$, and ask at which instant of the period the oscillator is there. Each x is passed twice per period, once moving left ($p < 0$, $0 < t < T/2$) and once moving right ($p > 0$, $T/2 < t < T$), so the rule must also say which passage we mean; we let a fair coin decide. Now the map runs the other way, from x to t, and the pair (position, branch) fixes the instant uniquely. The instant t in $(0, T/2)$ comes only from the position $x(t)$ on the left-moving branch, and an instant in $(T/2, T)$ only from $x(t)$ on the right-moving one, so

$$P(t) \overset{(1)}{=} \frac{1}{2}\, P\big(x(t)\big) \left|\frac{\mathrm{d}x}{\mathrm{d}t}\right| \overset{(2)}{=} \frac{1}{2} \cdot \frac{1}{2A} \cdot A\omega\, |\sin \omega t| = \frac{\omega}{4}\, |\sin \omega t|\,, \qquad 0 \le t < T. \tag{1.35}$$

(1) The Jacobian rule (1.33) for the map $x \mapsto t$ on the branch that contains t, times the probability $\frac{1}{2}$ that the coin chose that branch. Each instant has exactly one preimage, so there is one term.

(2) $P(x) = 1/(2A)$ is the new sampling prescription, and $|\mathrm{d}x/\mathrm{d}t| = A\omega\,|\sin \omega t|$ from $x = A\cos \omega t$.

It is normalised: $\int_0^T \frac{\omega}{4}\,|\sin \omega t|\,\mathrm{d}t = \frac{\omega}{4} \cdot 2 \cdot \frac{2}{\omega} = 1$. The time is no longer uniform. The sampled instants crowd around $t = T/4$ and $3T/4$, where the mass crosses the centre at full speed, and they avoid the turning points. This reverses the arcsine picture because each $\mathrm{d}x$ now carries the same probability. Where the swing is fast it covers more such intervals per unit time, so more sampled instants fall on that part of the swing. The fraction of instants in the first tenth of the period is $(1 - \cos \frac{\pi}{5})/4 = 0.0477$, not 0.1. The script `14b_sampling_rule.py` finds 0.0478, and its histogram of t matches (1.35) with $\chi^2 = 54.7$ over 50 bins (Figure 1.34). The equation of motion is the same in both questions. Only the sampling rule changed, and the distributions of both x and t changed with it.

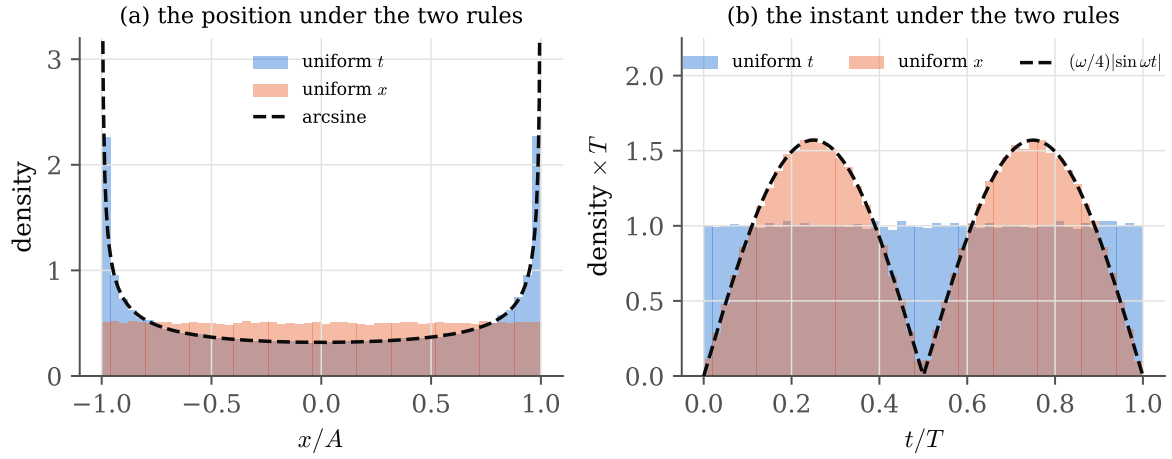


Figure 1.34: One oscillator, two sampling rules, 2×10^5 samples each. (a) Choosing the instant uniformly gives the arcsine law for the position (blue); choosing the position uniformly gives, by construction, a flat histogram (orange). (b) The instants behind the same samples: flat under the first rule, and under the second rule the two humps of $(\omega/4)\,|\sin \omega t|$, highest where the mass passes the centre and zero at the turning points $t = 0$, $T/2$, T.

Key idea: Equations of motion + sampling prescription → probability distribution

A deterministic motion has no distribution until we say what is chosen at random. The same orbit gives the arcsine law for x if the instant is uniform, and $(\omega/4)\,|\sin \omega t|$ for t if the position is uniform. Conservation laws tell us which states are *accessible* (energy

conservation confines the oscillator to its ellipse), not that they are equally probable. Being canonically conjugate does not make a variable uniform either: the phase $\theta = \omega t$, conjugate to the action E/ω, is uniform under the first rule only because that rule chose the instant uniformly, and under the second rule it has density $\frac{1}{4}|\sin\theta|$.

Method 3: from first principles, if we did not know that t is uniform. Method 2 assumed (1.34). Suppose we did not know that the instant is uniformly distributed, and asked instead what statistical mechanics says: in the **microcanonical ensemble** all phase-space points with energy between E and $E + \Delta E$ are equally probable. We then only have to measure areas in phase space. The energy is

$$H(x,p) = \frac{p^2}{2m} + \frac{1}{2}m\omega^2x^2 = E, \qquad E = \frac{1}{2}m\omega^2A^2,$$

so the orbit is the ellipse $x^2/A^2 + p^2/(m\omega A)^2 = 1$, with the two branches $p_\pm(x) = \pm m\omega\sqrt{A^2 - x^2}$. Thicken it into a shell $E \to E + \Delta E$. At fixed x the shell has thickness in the p direction

$$\Delta p \overset{(1)}{=} \sqrt{2m(E+\Delta E) - m^2\omega^2x^2} - \sqrt{2mE - m^2\omega^2x^2} \overset{(2)}{=} \frac{m\,\Delta E}{\sqrt{2mE - m^2\omega^2x^2}} \overset{(3)}{=} \frac{\Delta E}{\omega\sqrt{A^2 - x^2}}.$$

(1) Solve $H = E$ for $|p|$: $|p| = \sqrt{2mE - m^2\omega^2x^2}$, at the outer and inner energies.
(2) First order in ΔE: $\sqrt{a+\epsilon} - \sqrt{a} \simeq \epsilon/(2\sqrt{a})$ with $\epsilon = 2m\Delta E$.
(3) $2mE = m^2\omega^2A^2$, so the square root is $m\omega\sqrt{A^2 - x^2}$.

The vertical strip between x and $x + \mathrm{d}x$ meets the shell twice, at p_+ and at p_- (the two dark patches in Figure 1.35b), so the shell area inside the strip and the total shell area are

$$\mathrm{d}\mathcal{A} \overset{(1)}{=} 2\,\Delta p\,\mathrm{d}x = \frac{2\Delta E}{\omega\sqrt{A^2 - x^2}}\,\mathrm{d}x,$$
$$\mathcal{A}_{\text{shell}} \overset{(2)}{=} \mathcal{A}(E+\Delta E) - \mathcal{A}(E) \overset{(3)}{=} \frac{2\pi}{\omega}\,\Delta E.$$

(1) Two patches, each of height Δp and width $\mathrm{d}x$.
(2) The shell is the region between the two ellipses.
(3) The ellipse of energy E has semi-axes $\sqrt{2E/m\omega^2}$ and $\sqrt{2mE}$, so it encloses $\mathcal{A}(E) = \pi\sqrt{2E/m\omega^2}\sqrt{2mE} = 2\pi E/\omega$, which is linear in E.

Equal probability for equal phase-space area then gives

$$P(x)\,\mathrm{d}x \overset{(1)}{=} \frac{\mathrm{d}\mathcal{A}}{\mathcal{A}_{\text{shell}}} \overset{(2)}{=} \frac{2\Delta E\,\mathrm{d}x/\left(\omega\sqrt{A^2 - x^2}\right)}{2\pi\Delta E/\omega} = \frac{\mathrm{d}x}{\pi\sqrt{A^2 - x^2}}, \qquad |x| < A,$$

(1) Microcanonical ensemble: probability is proportional to phase-space area within the shell.
(2) The two results above; ΔE and ω cancel.

and it is normalised, $\frac{1}{\pi}\int_{-A}^{A}\mathrm{d}x/\sqrt{A^2 - x^2} = \frac{1}{\pi}[\arcsin(x/A)]_{-A}^{A} = 1$.

All three methods agree, and the agreement is not a coincidence. Methods 2 and 3 say the same thing from two ends: the *time* average over one orbit equals the *ensemble* average over the energy shell. To see the connection, compare the time spent in a patch with the patch's area. The time is its width divided by the speed, $\mathrm{d}x/|v|$. The patch area is $2\Delta p\,\mathrm{d}x$, and $\Delta p \propto 1/|v|$, so the area is proportional to the same thing. A uniformly random instant therefore lands in each

patch in proportion to its area. This equality of time and ensemble averages is **ergodicity**, and it is the property that makes Markov-chain sampling work in Chapter 9.

The density diverges at the turning points, where $v = 0$, but integrably. For example

$$\mathrm{P}\big(|X| > 0.9A\big) \overset{(1)}{=} 1 - \big[F_X(0.9A) - F_X(-0.9A)\big] \overset{(2)}{=} 1 - \frac{2}{\pi}\arcsin 0.9 = 0.2871,$$

(1) Complement of $-0.9A < X \le 0.9A$, and (1.27).
(2) arcsin is odd, so $F(0.9A) - F(-0.9A) = \frac{2}{\pi}\arcsin 0.9$.

so more than a quarter of the time is spent in the outer 10% of the range, both sides combined. Averaging x^2 over this density gives $\langle x^2\rangle = A^2/2$ (Section 1.12), the statement that the time-averaged potential energy is half the total.

Listing 1.14: The oscillator: random instants, deterministic map onto the phase-space orbit (`code/ch01/14_oscillator_arcsine.py`)

```
23 omega = 2 * np.pi / T
24 N = 200_000
25 t = rng.uniform(0.0, T, size=N)      # the random ingredient: WHEN we look
26 x = A * np.cos(omega * t)            # the deterministic map t -> (x, p): a point on the
       orbit
27 p = -m * omega * A * np.sin(omega * t)
```

Line 26 is the only random line: *when* we look. Lines 27–28 apply the deterministic map to a point (x, p) of the orbit, which is the sampled element of the sample space. Everything else is comparison. The script also computes the quantum probability density $|\psi_{30}(x)|^2$ of the oscillator eigenstate with $n = 30$, using the stable recursion for Hermite functions (lines 39–46), at the same energy $E = (n + \frac{1}{2})\hbar\omega = \frac{1}{2}m\omega^2A^2$, i.e. $A = \sqrt{2n+1}$ in units $\hbar = m = \omega = 1$.

The script finds $\mathrm{P}(|X| > 0.9A) = 0.2872$ (exact 0.2871), $\langle x^2\rangle = 0.4995\,A^2$ (exact 0.5), and a χ^2 of 65.4 over 60 bins against the bin-integrated arcsine law, computed from CDF differences (lines 105–109 of the script).

How research uses it. The oscillator experiment is: pick random t, compute $x(t)$, histogram. The histogram converges to the density of x without our ever writing that density down. Chapters 5 and 9 do exactly this with a posterior: an MCMC sampler produces points $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots$ and the histogram of any function $g(\boldsymbol{\theta})$ of them (a derived parameter such as H_0 computed from $\Omega_m h^2$ and Ω_m) is its posterior density, Jacobian included automatically. Here the Monte Carlo *verifies* a known arcsine law. There it *is* the answer, because the posterior density of a derived parameter has no closed form.

1.11.4 Why every point of the energy shell is equally likely: maximum entropy

Method 3 for the oscillator (Section 1.11.3) found the arcsine law without ever mentioning time. It used one sentence instead: all points of the energy shell are equally probable. Where does that sentence come from? Not from energy conservation, which only confines the oscillator to its ellipse and says nothing about how probable each point of it is (the key idea of Section 1.11.3). It is an extra assumption, and it can be derived from a cleaner one. Three assumptions are needed.

1. *A probabilistic description.* Cut phase space into small cells $i = 1, 2, \ldots$ of equal area $\mathrm{d}x\,\mathrm{d}p$. The state of the system is described by probabilities $p_i \ge 0$ with $\sum_i p_i = 1$. Equal area is the natural cell size because phase-space area is what the Hamiltonian motion preserves (Liouville's theorem).

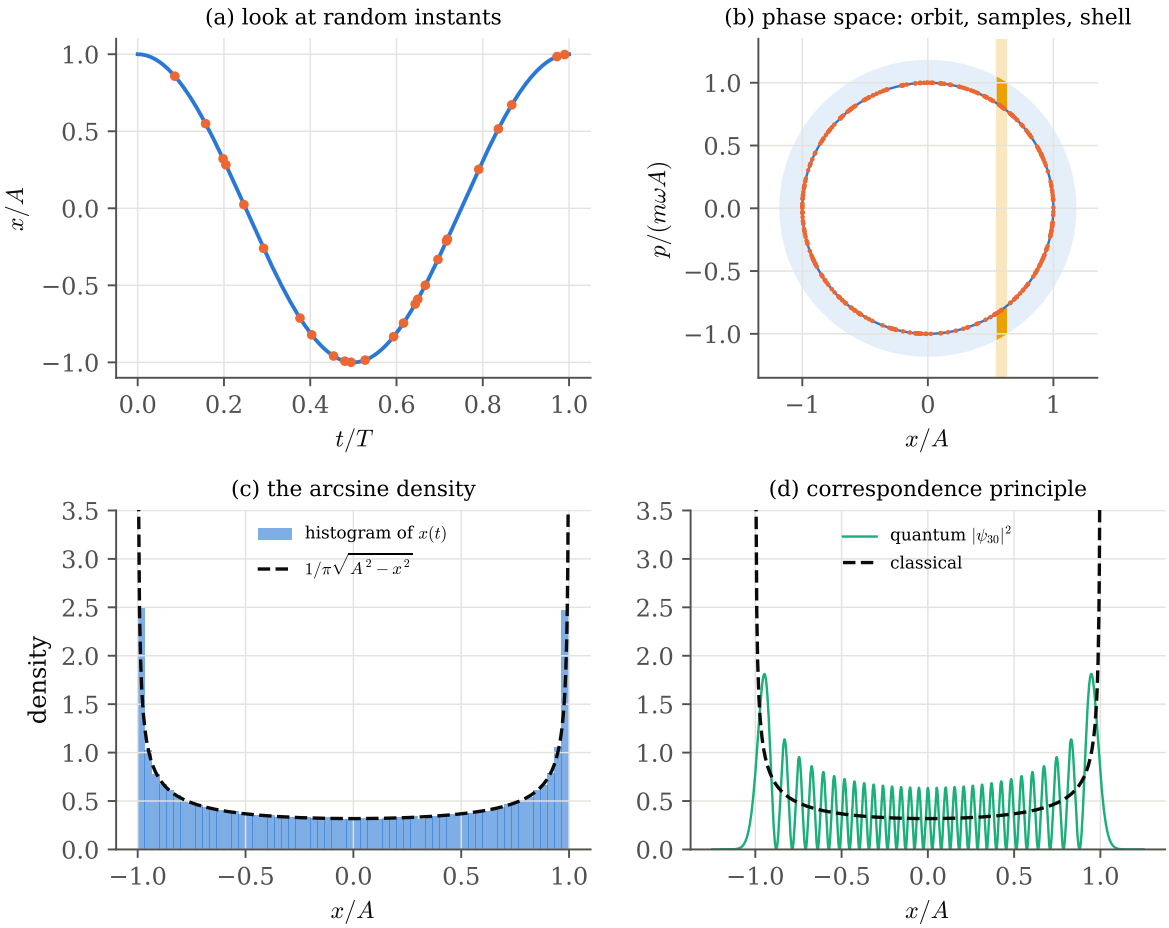


Figure 1.35: (a) The trajectory, looked at in random instants (dots). (b) The same instants as points (x, p) of phase space: they all lie on the energy ellipse (a circle in the scaled units), the sample space of this problem. The pale ring is a thickened energy shell; the vertical strip between x and $x + \mathrm{d}x$ meets it in two dark patches, at p_+ and p_-, whose area is what Method 3 computes. (c) The histogram of x at 2×10^5 random instants follows the arcsine density, lowest at the centre and rising steeply towards the turning points. (d) The quantum density for $n = 30$ oscillates rapidly, but its local average follows the same classical curve: the correspondence principle seen as a statement about probability densities.

2. *Constraints.* What we know about the system is written as equations the p_i must obey.
3. *Equilibrium.* Among all distributions that obey the constraints, the equilibrium one maximises the Gibbs–Shannon entropy

$$S[p] = -k_B \sum_i p_i \ln p_i . \tag{1.36}$$

Only normalisation: the microcanonical distribution. Keep only the M cells of the shell between E and $E + \Delta E$, the accessible states, and impose nothing but normalisation. With a Lagrange multiplier $k_B\alpha$ for $\sum_i p_i = 1$, the maximum has

$$0 \overset{(1)}{=} \frac{\partial}{\partial p_j}\Big[-k_B \sum_i p_i \ln p_i - k_B\alpha\Big(\sum_i p_i - 1\Big)\Big] \overset{(2)}{=} -k_B\big(\ln p_j + 1 + \alpha\big),$$

$$p_j \overset{(3)}{=} e^{-1-\alpha} \overset{(4)}{=} \frac{1}{M} \qquad \text{for every cell } j \text{ of the shell.}$$

(1) The stationary point of the entropy with the constraint added through its multiplier.
(2) Only the term $i = j$ depends on p_j, and $\frac{\mathrm{d}}{\mathrm{d}p}(p \ln p) = \ln p + 1$.
(3) Solve for p_j.
(4) The right side of (3) is the same number for every j; normalisation over the M cells fixes it to $1/M$.

It is a maximum, because $\partial^2 S/\partial p_j^2 = -k_B/p_j < 0$, and its value is $S = k_B \ln M$, Boltzmann's entropy of Section 1.3. Equal probability per cell of equal area means probability proportional to phase-space area: Method 3's "equal probability for equal phase-space area" was this assumption.

Normalisation and a fixed mean energy: the canonical distribution. Now let the system exchange energy with its surroundings, so that cells of every energy E_i are accessible, and suppose all we know is the mean energy, $\sum_i p_i E_i = U$. A second multiplier λ enters:

$$0 \overset{(1)}{=} \frac{\partial}{\partial p_j}\Big[S[p] - k_B\alpha\Big(\sum_i p_i - 1\Big) - \lambda\Big(\sum_i p_i E_i - U\Big)\Big] \overset{(2)}{=} -k_B\big(\ln p_j + 1 + \alpha\big) - \lambda E_j,$$

$$p_j \overset{(3)}{=} e^{-1-\alpha}\, e^{-\beta E_j} \overset{(4)}{=} \frac{e^{-\beta E_j}}{Z}, \qquad \beta \equiv \frac{\lambda}{k_B}, \quad Z = \sum_i e^{-\beta E_i}.$$

(1) The stationary point with both constraints.
(2) As before, plus $\partial(\sum_i p_i E_i)/\partial p_j = E_j$.
(3) Solve for $\ln p_j$ and exponentiate.
(4) Normalisation fixes $e^{-1-\alpha} = 1/Z$; Z, the **partition function**, is simply the normaliser.

What is β? Put the solution back into the entropy and ask how the maximum entropy changes with U:

$$S \overset{(1)}{=} -k_B \sum_i p_i\big(-\beta E_i - \ln Z\big) \overset{(2)}{=} k_B\big(\beta U + \ln Z\big),$$

$$\frac{\mathrm{d}S}{\mathrm{d}U} \overset{(3)}{=} k_B\Big(\beta + U\frac{\mathrm{d}\beta}{\mathrm{d}U} + \frac{\partial \ln Z}{\partial \beta}\frac{\mathrm{d}\beta}{\mathrm{d}U}\Big) \overset{(4)}{=} k_B\beta.$$

(1) $\ln p_i = -\beta E_i - \ln Z$.
(2) $\sum_i p_i E_i = U$ and $\sum_i p_i = 1$.
(3) β depends on U through the constraint; differentiate both terms.
(4) $\partial_\beta \ln Z = -\sum_i E_i e^{-\beta E_i}/Z = -U$, so the last two terms cancel.

Thermodynamics defines the temperature by $1/T = \mathrm{d}S/\mathrm{d}U$. So $\lambda = k_B\beta = 1/T$, that is $\beta = 1/k_B T$, and the maximum-entropy distribution at fixed mean energy is the Boltzmann distribution $p_i = e^{-E_i/k_B T}/Z$.

No sampling rule, and a definition that closes on itself. In Section 1.11.3 a probability needed a sampling prescription: choose the instant, or choose the position, and the answers differed. Here we never said what is chosen at random, and we did not need to. Why not? The entropy (1.36) is a function of the probabilities themselves, so maximising it is a statement about the p_i directly. The third assumption takes the place of the sampling rule. That assumption is also a definition: equilibrium *is* the maximum-entropy state compatible with the constraints. So the argument closes on itself. The microcanonical and canonical distributions apply exactly when the system is in that state, and the derivation says nothing about whether, or how fast, a given system gets there. For the one-dimensional oscillator the dynamics answer that question at

once: a uniform instant lands on the shell in proportion to area (the ergodicity paragraph of Section 1.11.3). The same variational argument with a fixed variance gives the Gaussian as the maximum-entropy density, in Chapter 2.

1.11.5 Time averages of a bounded motion: the virial theorem

The oscillator's $\langle x^2\rangle = A^2/2$ of Section 1.11.3 is a time average: under the uniform-instant rule, averaging over the arcsine density is averaging over one period. It says that the mean potential energy is $\langle V\rangle = \frac{1}{2}m\omega^2\langle x^2\rangle = \frac{1}{4}m\omega^2A^2 = E/2$, so the mean kinetic energy is $\langle K\rangle = E - \langle V\rangle = E/2$ as well: $2\langle K\rangle = 2\langle V\rangle$. This is the virial theorem for a potential $V \propto x^2$, and the general version takes three lines.

Take particles with positions $\boldsymbol{r}_a$ and momenta $\boldsymbol{p}_a$, potential $V(\boldsymbol{r}_1, \boldsymbol{r}_2, \dots)$, and kinetic energy $K = \sum_a \boldsymbol{p}_a^2/2m_a$. Let $G = \sum_a \boldsymbol{r}_a \cdot \boldsymbol{p}_a$ (for one particle in one dimension, $G = xp$). Then

$$\frac{\mathrm{d}G}{\mathrm{d}t} \overset{(1)}{=} \sum_a \dot{\boldsymbol{r}}_a \cdot \boldsymbol{p}_a + \sum_a \boldsymbol{r}_a \cdot \dot{\boldsymbol{p}}_a \overset{(2)}{=} 2K - \sum_a \boldsymbol{r}_a \cdot \nabla_a V,$$

$$\frac{G(\tau) - G(0)}{\tau} \overset{(3)}{=} 2\langle K\rangle_\tau - \Big\langle \sum_a \boldsymbol{r}_a \cdot \nabla_a V \Big\rangle_\tau \overset{(4)}{=} 0 \quad \text{in the limit } \tau \to \infty.$$

(1) Product rule.
(2) $\dot{\boldsymbol{r}}_a = \boldsymbol{p}_a/m_a$, so the first sum is $\sum_a \boldsymbol{p}_a^2/m_a = 2K$; Newton's law $\dot{\boldsymbol{p}}_a = -\nabla_a V$ in the second.
(3) Average both sides over a time τ, $\langle f\rangle_\tau = \frac{1}{\tau}\int_0^\tau f\,\mathrm{d}t$.
(4) For a bounded motion G stays bounded, so the left side goes to zero as $\tau \to \infty$ (for a periodic motion it is exactly zero over one period).

Hence the time-averaged **virial theorem**

$$2\langle K\rangle = \Big\langle \sum_a \boldsymbol{r}_a \cdot \nabla_a V \Big\rangle \qquad \Longrightarrow \qquad 2\langle K\rangle = n\,\langle V\rangle \quad \text{for } V \propto r^n, \tag{1.37}$$

the second form because a potential homogeneous of degree n, $V(\lambda\boldsymbol{r}_1, \lambda\boldsymbol{r}_2, \dots) = \lambda^n V$, obeys Euler's relation $\sum_a \boldsymbol{r}_a \cdot \nabla_a V = nV$ (differentiate in λ at $\lambda = 1$). For the oscillator $n = 2$ and $\langle K\rangle = \langle V\rangle$, as found above.

Gravity: negative heat capacity. For self-gravitating masses, $V = -\sum_{a<b} Gm_a m_b/|\boldsymbol{r}_a - \boldsymbol{r}_b|$ is homogeneous of degree $n = -1$. Then

$$E \overset{(1)}{=} \langle K\rangle + \langle V\rangle \overset{(2)}{=} \langle K\rangle - 2\langle K\rangle = -\langle K\rangle \overset{(3)}{=} -\tfrac{3}{2}Nk_BT, \qquad \frac{\mathrm{d}E}{\mathrm{d}T} = -\tfrac{3}{2}Nk_B < 0.$$

(1) E is conserved, so it equals its own time average.
(2) The virial theorem (1.37) with $n = -1$: $\langle V\rangle = -2\langle K\rangle$.
(3) The kinetic-gas definition of temperature for N particles, $\langle K\rangle = \frac{3}{2}Nk_BT$.

A self-gravitating system has a negative heat capacity: when it loses energy it heats up. A star radiates, contracts, its potential energy falls by twice the energy lost, and half of that goes into faster thermal motion. Star clusters behave the same way: the core loses energy to the halo and gets hotter.

The canonical distribution cannot have $C < 0$. The canonical distribution of Section 1.11.4 fixes the sign of the heat capacity, so it cannot describe such a system. With $Z(\beta) = \sum_i e^{-\beta E_i}$,

$$\frac{\partial \ln Z}{\partial \beta} \overset{(1)}{=} -\frac{\sum_i E_i e^{-\beta E_i}}{Z} = -\langle E \rangle, \qquad \frac{\partial^2 \ln Z}{\partial \beta^2} \overset{(2)}{=} \frac{\sum_i E_i^2 e^{-\beta E_i}}{Z} - \left(\frac{\sum_i E_i e^{-\beta E_i}}{Z}\right)^2 = \mathrm{Var}(E),$$

$$C \overset{(3)}{=} \frac{\mathrm{d}\langle E \rangle}{\mathrm{d}T} \overset{(4)}{=} \frac{\mathrm{d}\beta}{\mathrm{d}T}\,\frac{\mathrm{d}\langle E \rangle}{\mathrm{d}\beta} \overset{(5)}{=} \left(-\frac{1}{k_B T^2}\right)\left(-\frac{\partial^2 \ln Z}{\partial \beta^2}\right) = \frac{\mathrm{Var}(E)}{k_B T^2} \;\geq\; 0.$$

(1) Differentiate $\ln Z$; the ratio is the canonical average of E.
(2) Differentiate once more (quotient rule): $\partial_\beta (Z'/Z) = Z''/Z - (Z'/Z)^2$, which is $\langle E^2 \rangle - \langle E \rangle^2$ (the generating-function view of $\ln Z$ is in Section 1.14).
(3) The heat capacity is the change of mean energy with temperature.
(4) Chain rule through $\beta = 1/k_B T$.
(5) $\mathrm{d}\beta/\mathrm{d}T = -1/k_B T^2$ and the two results of the first line.

A variance is never negative, so a system described by the canonical distribution has $C \geq 0$. A self-gravitating system, with $C < 0$, cannot be described by the canonical distribution, and physically it cannot sit in equilibrium with a heat bath: if it is a little hotter than the bath it loses energy, gets hotter still, and runs away. Only the microcanonical description, an isolated system at fixed E with maximum entropy on its shell, can apply to it. The assumptions of Section 1.11.4 say exactly when each distribution is legitimate: fixed energy gives the microcanonical, a fixed mean energy (which in practice means contact with a bath) gives the canonical, and the second needs $C \geq 0$ to be consistent.

1.11.6 Isotropic decays: flat in $\cos\theta$, peaked in p_T

Example 1.47 (Angles and transverse momentum of an isotropic decay). A particle at rest decays into two daughters, each of momentum p^* in opposite directions. "Isotropic" means the daughter direction is uniform on the unit sphere: the probability of a patch is its solid angle over 4π, $\mathrm{d}\,\mathrm{P} = \mathrm{d}\Omega/4\pi = \sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi/4\pi$ with $\theta \in [0, \pi]$ the polar angle to the beam axis z. Then:

$$f(\theta) \overset{(1)}{=} \int_0^{2\pi} \frac{\sin\theta}{4\pi}\,\mathrm{d}\phi = \frac{\sin\theta}{2}, \qquad f_{\cos\theta}(u) \overset{(2)}{=} \frac{\sin\theta}{2} \cdot \frac{1}{\sin\theta} = \frac{1}{2}, \quad u \in [-1, 1].$$

(1) The joint density of (θ, ϕ) is $\sin\theta/4\pi$. Marginalise the azimuth ϕ.
(2) $u = \cos\theta$ is decreasing on $[0, \pi]$ with $|\mathrm{d}\theta/\mathrm{d}u| = 1/\sin\theta$. Apply (1.33).

So $\cos\theta$ is flat, θ is not (few decays go along the beam, because the poles have little solid angle), and $p_z = p^* \cos\theta$ is uniform on $[-p^*, p^*]$. The transverse momentum $p_T = p^* \sin\theta$ is different: two angles, θ and $\pi - \theta$, give the same p_T. With $\sin\theta = p_T/p^*$ and $|\cos\theta| = \sqrt{1 - p_T^2/p^{*2}}$,

$$\left|\frac{\mathrm{d}\theta}{\mathrm{d}p_T}\right| \overset{(1)}{=} \frac{1}{p^*\,|\cos\theta|} = \frac{1}{\sqrt{p^{*2} - p_T^2}},$$

$$f(p_T) \overset{(2)}{=} 2 \times \frac{\sin\theta}{2} \times \frac{1}{\sqrt{p^{*2} - p_T^2}} = \frac{p_T}{p^*\sqrt{p^{*2} - p_T^2}}, \qquad 0 < p_T < p^*.$$

(1) $\mathrm{d}p_T/\mathrm{d}\theta = p^* \cos\theta$.
(2) Two branches ($\theta < \pi/2$ and $\theta > \pi/2$) with the same density $\sin\theta/2$ and the same Jacobian.

The density diverges at $p_T = p^*$, the **Jacobian peak**: decays near $\theta = \pi/2$ all pile up at the maximum p_T, because p_T is stationary there. For $W \to e\nu$ with the W at rest, $p^* = m_W/2 \approx 40\,\text{GeV}$, and the sharp edge of the electron p_T spectrum near 40 GeV is what makes the p_T distribution sensitive to m_W. As a check, $\mathrm{P}(p_T > 0.9p^*) = \mathrm{P}(|\cos\theta| < \sqrt{1-0.81}) = \sqrt{0.19} = 0.4359$, because $\cos\theta$ is uniform.

Listing 1.15: Isotropic directions and the transverse momentum (`code/ch01/15_decay_angles.py`)

```
22 m_W = 80.4                       # GeV
23 pstar = m_W / 2
24 N = 400_000
25 # uniform direction on the sphere: normalise a 3-D standard Gaussian vector
26 v = rng.normal(size=(N, 3))
27 v /= np.linalg.norm(v, axis=1, keepdims=True)
28 cos_t = v[:, 2]
29 theta = np.arccos(cos_t)
30 pT = pstar * np.sqrt(v[:, 0]**2 + v[:, 1]**2)
```

Lines 26–27 generate isotropic directions by normalising a vector of three independent standard Gaussians. This works because the joint density of the three Gaussians, $(2\pi)^{-3/2}e^{-|v|^2/2}$, depends only on the length $|v|$, so it is unchanged by rotations and the direction must be uniform. Lines 28–30 compute the three random variables of the example from the same simulated directions.

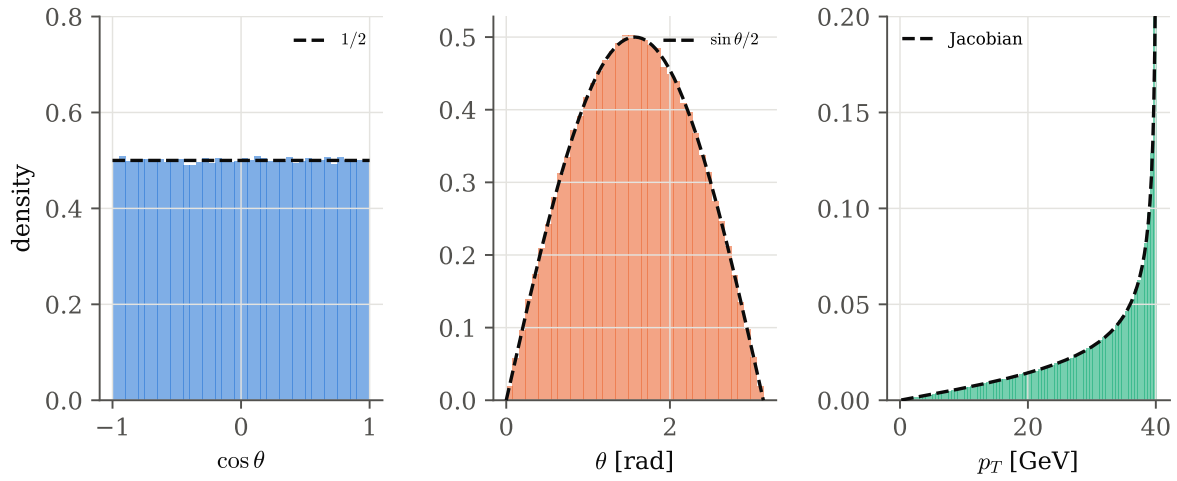


Figure 1.36: 4×10^5 isotropic decays of a particle at rest with $p^* = m_W/2$. Left: $\cos\theta$ is flat at $\frac{1}{2}$. Middle: θ itself follows $\sin\theta/2$, vanishing along the beam. Right: the transverse momentum rises towards the kinematic endpoint p^* and piles up there (the Jacobian peak).

The simulation gives $\langle\cos\theta\rangle = 0.0006$ (exact 0), $\langle p_T\rangle = 31.566\,\text{GeV}$ against the exact $p^*\langle\sin\theta\rangle = \pi p^*/4 = 31.573\,\text{GeV}$, and $\mathrm{P}(p_T > 0.9p^*) = 0.4353$ against 0.4359.

1.11.7 Several variables: the Jacobian determinant

Everything above extends to several variables at once, with lengths replaced by areas or volumes.

Working definition 1.21 (change of variables in several dimensions)

Let (X, Y) have joint density $f_{X,Y}$ and let $(U, V) = g(X, Y)$ be one-to-one on the region where $f_{X,Y} > 0$, with inverse $x = h_1(u, v)$, $y = h_2(u, v)$. Then

$$f_{U,V}(u, v) = f_{X,Y}\big(h_1(u, v), h_2(u, v)\big)\ |J|\,, \qquad J = \det \begin{pmatrix} \partial x/\partial u & \partial x/\partial v \\ \partial y/\partial u & \partial y/\partial v \end{pmatrix}, \tag{1.38}$$

[the old density at the preimage point, times the factor by which the inverse map scales areas near it]. The same holds in n dimensions with the $n \times n$ Jacobian determinant [Cowan (1.37)–(1.38)]. *Operational test:* a small rectangle $\mathrm{d}u\,\mathrm{d}v$ is the image of a parallelogram of area $|J|\,\mathrm{d}u\,\mathrm{d}v$, and both must carry the same probability.

The multidimensional formula is (1.33) with "length" replaced by "area": the inverse map sends the rectangle with sides $\mathrm{d}u, \mathrm{d}v$ to a parallelogram spanned by the vectors $(\partial_u x, \partial_u y)\mathrm{d}u$ and $(\partial_v x, \partial_v y)\mathrm{d}v$, whose area is $|J|\,\mathrm{d}u\,\mathrm{d}v$, the familiar volume element of a coordinate change.

Example 1.48 (Sum and difference of two Gaussians; polar coordinates). Let $X, Y \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$ and $U = X + Y$, $V = X - Y$ [CasellaBerger Ex. 4.3.4]. The inverse is $x = (u + v)/2$, $y = (u - v)/2$, so

$$J \overset{(1)}{=} \det \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{pmatrix} = -\tfrac{1}{2}, \qquad x^2 + y^2 \overset{(2)}{=} \frac{(u+v)^2 + (u-v)^2}{4} = \frac{u^2 + v^2}{2},$$

$$f_{U,V}(u, v) \overset{(3)}{=} \frac{1}{2\pi} e^{-(u^2+v^2)/4} \cdot \frac{1}{2} = \left[\frac{e^{-u^2/4}}{\sqrt{2\pi \cdot 2}}\right]\left[\frac{e^{-v^2/4}}{\sqrt{2\pi \cdot 2}}\right].$$

(1) Partial derivatives of x and y with respect to u, v.
(2) The cross terms $\pm 2uv$ cancel.
(3) (1.38) with $f_{X,Y} = \frac{1}{2\pi} e^{-(x^2+y^2)/2}$ and $|J| = \frac{1}{2}$.

The joint factorises: U and V are independent, each $\mathcal{N}(0, 2)$. The sum and difference of two equally noisy channels are statistically independent, a fact Section 1.13 will explain through their covariance.

In polar coordinates, $x = r\cos\varphi$, $y = r\sin\varphi$, $J = r$, and the same density becomes $f(r, \varphi) = \frac{1}{2\pi} \cdot r e^{-r^2/2}$: the angle is uniform on $[0, 2\pi)$, independent of the radius, and R has density $r e^{-r^2/2}$ (the Rayleigh distribution). Read backwards this is the Box–Muller recipe for making Gaussians from uniforms (Chapter 6).

For a single function $Z = r(X, Y)$ the three-step CDF method still works [Wasserman p. 42]: find $A_z = \{(x, y) : r(x, y) \le z\}$, integrate the joint over it, differentiate. Alternatively, invent a second variable to make the map one-to-one, use (1.38), and marginalise the invented variable [CasellaBerger §4.3]. For $Z = X + Y$ take $(x, y) \to (x, z)$ with $z = x + y$, so $y = z - x$ and $|J| = 1$. For independent X, Y with densities g, h:

$$f_Z(z) \overset{(1)}{=} \int f_{X,Z}(x, z)\,\mathrm{d}x \overset{(2)}{=} \int g(x)\,h(z - x)\,\mathrm{d}x, \tag{1.39}$$

(1) Marginalise the auxiliary variable x.
(2) (1.38) with $|J| = 1$ and independence, $f_{X,Y}(x, z - x) = g(x)h(z - x)$.

the **convolution** $g * h$: the density of a sum is the convolution of the densities [Cowan (1.36)]. For a product $Z = XY$ take $y = z/x$, $|J| = 1/|x|$, and get $f_Z(z) = \int g(x)h(z/x)\,\mathrm{d}x/|x|$, the Mellin convolution [Cowan (1.35)].

Example 1.49 (The sum of two uniforms is a triangle). Let $X_1, X_2 \overset{\text{iid}}{\sim} \mathrm{Uniform}(0, 1)$ and $Y = X_1 + X_2$ [Wasserman Ex. 2.48, p. 43]. *By geometry*: $A_y = \{x_1 + x_2 \le y\}$ within the unit square

is a triangle of area $y^2/2$ for $0 \le y < 1$, and the square minus a corner triangle of area $(2-y)^2/2$ for $1 \le y < 2$ (Figure 1.37). So $F_Y = y^2/2$ and then $1-(2-y)^2/2$, and $f_Y(y) = y$ on $[0,1]$, $2-y$ on $[1,2]$. *By convolution* (1.39), with $g = h = \mathbb{1}_{[0,1]}$ (the indicator of $[0,1]$),

$$f_Y(y) \overset{(1)}{=} \int_0^1 \mathbb{1}\left[0 \le y - x \le 1\right] dx \overset{(2)}{=} \min(y,1) - \max(0, y-1) = \begin{cases} y, & 0 \le y \le 1, \\ 2-y, & 1 \le y \le 2. \end{cases}$$

(1) $g(x) = 1$ restricts x to $[0,1]$, and $h(y-x) = 1$ needs $0 \le y - x \le 1$.
(2) Both conditions together: $\max(0, y-1) \le x \le \min(1, y)$. The integral is the length of that interval.

Adding two flat distributions gives a peaked one. Adding more gives something ever closer to a Gaussian, the central limit theorem of Chapter 2; Section 1.14 shows why the convolution becomes simple after a Fourier transform.

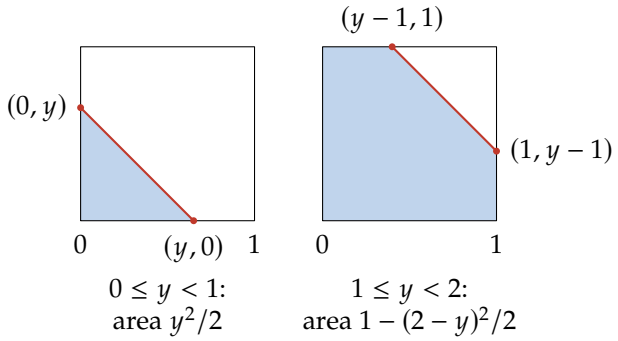


Figure 1.37: The event $X_1 + X_2 \le y$ inside the unit square is everything below the red line $x_2 = y - x_1$. For $y < 1$ it is a small triangle. For $y > 1$ it is the whole square minus the unshaded corner. The shaded area is the CDF of the sum.

The worked cases used every method in turn. The table says which one to reach for in each situation.

Taxonomy: Which method for which transformation

situation	method
X discrete	add probabilities of all preimages
r monotone	$f_Y = f_X(s(y))\,\lvert s'(y)\rvert$
r many-to-one (e.g. x^2, cos, sin)	sum the monotone formula over branches, or use the CDF
awkward region, piecewise	CDF method: draw A_y, integrate, differentiate
n variables to n variables	Jacobian determinant (1.38)
one function of n variables	add auxiliary variables, Jacobian, marginalise; sums $\to$ convolution
sums of many independent variables	characteristic functions (Section 1.14)
anything else	simulate and histogram (Monte Carlo, Chapter 6)

Three mistakes recur in all of these methods.

Watch out: Transforming the density's value instead of its argument

Writing $f_Y(y) = f_X(s(y))$ and forgetting the Jacobian is the commonest error. The result is not normalised, so a quick check $\int f_Y = 1$ catches it. A second error is forgetting branches

(θ and $\pi - \theta$ both give the same p_T). A third is expecting the peak to transform like a point: modes move under reparametrisation, quantiles do not (Example 1.45).

1.12 Expectation and variance: the long-run average and the spread

A slot machine pays \$100 with probability 0.001, pays \$5 with probability 0.14, and otherwise takes our \$1. Should we play? No single pull answers that, because a single pull is random. What we can predict is the average over many pulls. In N pulls we expect about $0.001N$ big wins, $0.14N$ small wins and $0.859N$ losses, so the average payoff per pull is

$$0.001 \times 100 + 0.14 \times 5 + 0.859 \times (-1) = -0.059$$

dollars. In the long run we lose about six cents per pull [Kruschke §4.3.3]. Each possible value was weighted by how often it occurs. That weighted average is the expectation.

Physicists use this object all the time under other names. The ensemble average $\langle A \rangle = \sum_i p_i A_i$ of statistical mechanics, with $p_i = e^{-\beta E_i}/Z$, is an expectation. So is the quantum expectation value $\langle \psi | \hat{A} | \psi \rangle = \sum_a a \, |\langle a | \psi \rangle|^2$, where the Born probabilities $|\langle a | \psi \rangle|^2$ weight the eigenvalues a. Two numbers summarise a whole distribution: a location (the mean) and a spread (the variance). There are rules for computing both for sums, for functions and for conditional situations, without ever writing down the density of the quantity being averaged. Three questions organise the topic.

1. What is the average value of X, and of any function $r(X)$?
2. How far from that average does X typically fall?
3. If we learn the value of another variable, how do the average and the spread change, and how do the conditional pieces reassemble into the whole?

1.12.1 The expectation

Start from data. Draw X independently N times and let n_x be the number of draws that gave the value x. The average of the draws is

$$\frac{1}{N} \sum_{j=1}^{N} X_j = \sum_x x \, \frac{n_x}{N}.$$

The frequencies n_x/N approach the probabilities $f(x) = \mathrm{P}(X = x)$ as N grows (the frequency interpretation of Section 1.2; the law of large numbers in Chapter 2 makes this precise). So the long-run average is $\sum_x x \, f(x)$. For a continuous variable, sort the draws into bins of width Δx. The fraction in the bin at x is about $f(x)\,\Delta x$, where f is the density, and the sum over bins becomes an integral as $\Delta x \to 0$.

Working definition 1.22 (expectation, mean)

The **expectation** (or **mean**, or first moment) of X is

$$\mathbb{E}(X) = \begin{cases} \sum_x x \, f(x) & X \text{ discrete}, \\ \int_{-\infty}^{\infty} x \, f(x) \, \mathrm{d}x & X \text{ continuous}, \end{cases} \tag{1.40}$$

the values weighted by the pmf [$f(x) = \mathrm{P}(X = x)$, the mass at x] or by the density [probability per unit x]. It is written μ, μ_X, $\mathbb{E}\,X$ or $\langle X \rangle$ and has the units of X. It **exists** only

if $\sum_x |x|\, f(x)$ (or $\int |x|\, f(x)\,\mathrm{d}x$) is finite. *Operational test:* the average of many independent draws of X settles down to $\mathbb{E}(X)$.

[Wasserman Def. 3.1, p. 47] writes both cases as $\int x\,\mathrm{d}F(x)$, a single notation for "sum against the masses or integrate against the density". We will use the two explicit forms.

Example: simple means. For a Bernoulli trial, $\mathbb{E}(X) = 0\cdot(1-p) + 1\cdot p = p$: the mean of an indicator is the probability of the event it indicates. For the number of heads in two fair flips, $\mathbb{E}(X) = 0\cdot\frac{1}{4} + 1\cdot\frac{1}{2} + 2\cdot\frac{1}{4} = 1$ [Wasserman Exs. 3.2–3.3, p. 48]. For a fair die, $\mathbb{E}(X) = \frac{1}{6}(1+2+\cdots+6) = \frac{21}{6} = 3.5$ [SeeingTheory p. 10], a value the die can never show. The mean is a long-run average, not a typical outcome. For $X \sim \mathrm{Uniform}(-1,3)$ the density is $\frac{1}{4}$ on $(-1,3)$ and

$$\mathbb{E}(X) \overset{(1)}{=} \int_{-1}^{3} x\cdot\tfrac{1}{4}\,\mathrm{d}x \overset{(2)}{=} \tfrac{1}{4}\Big[\tfrac{x^2}{2}\Big]_{-1}^{3} = \tfrac{1}{4}\cdot\tfrac{9-1}{2} = 1.$$

(1) (1.40) with $f = \frac{1}{4}$ inside the interval and 0 outside.
(2) Antiderivative of x.

The mean is the midpoint, as symmetry demands. The same happens for the parabola $f(x) = 6x(1-x)$ on $[0,1]$ [Kruschke (4.7)]: $\mathbb{E}(X) = 6\int_0^1 (x^2 - x^3)\,\mathrm{d}x = 6\big(\frac{1}{3} - \frac{1}{4}\big) = \frac{1}{2}$.

Example: the mean lifetime. A lifetime with density $f(x) = e^{-x/\beta}/\beta$ for $x > 0$ (exponential with mean β, Section 1.9) has

$$\mathbb{E}(X) \overset{(1)}{=} \int_0^\infty x\,\frac{e^{-x/\beta}}{\beta}\,\mathrm{d}x \overset{(2)}{=} \beta\int_0^\infty u\,e^{-u}\,\mathrm{d}u$$
$$\overset{(3)}{=} \beta\Big(\big[-u\,e^{-u}\big]_0^\infty + \int_0^\infty e^{-u}\,\mathrm{d}u\Big) \overset{(4)}{=} \beta\,(0+1) = \beta.$$

(1) Definition (1.40); the density vanishes for $x < 0$.
(2) Substitute $u = x/\beta$, so $x = \beta u$ and $\mathrm{d}x = \beta\,\mathrm{d}u$.
(3) Integrate by parts with u differentiated and e^{-u} integrated.
(4) $ue^{-u} \to 0$ at both ends, and $\int_0^\infty e^{-u}\mathrm{d}u = 1$.

So β really is the mean lifetime, and the median $\beta\ln 2$ found in Example 1.34 sits below it.

The balance point. If we read $f(x)$ as a mass density along a rod (or as point masses $f(x_i)$ at positions x_i), then $\mathbb{E}(X) = \sum_i x_i f(x_i)$ with $\sum_i f(x_i) = 1$ is the centre of mass: the rod balances on a fulcrum placed at the mean (Figure 1.38). A small mass far out in a tail pulls the balance point a long way. This is why the mean of a skewed distribution is dragged towards its long tail, and why a single outlier can move a sample mean a lot.

Example: a density with no mean. The Cauchy density $f(x) = 1/[\pi(1+x^2)]$ is the Breit–Wigner (Lorentzian) line shape of a resonance with unit half-width, centred at 0. It is symmetric, so one is tempted to say its mean is 0. But the mean exists only if $\int |x|\, f(x)\,\mathrm{d}x$ is finite, and

$$\int_{-\infty}^{\infty} \frac{|x|}{\pi(1+x^2)}\,\mathrm{d}x \overset{(1)}{=} \frac{2}{\pi}\int_0^\infty \frac{x}{1+x^2}\,\mathrm{d}x \overset{(2)}{=} \frac{1}{\pi}\Big[\ln(1+x^2)\Big]_0^\infty = \infty.$$

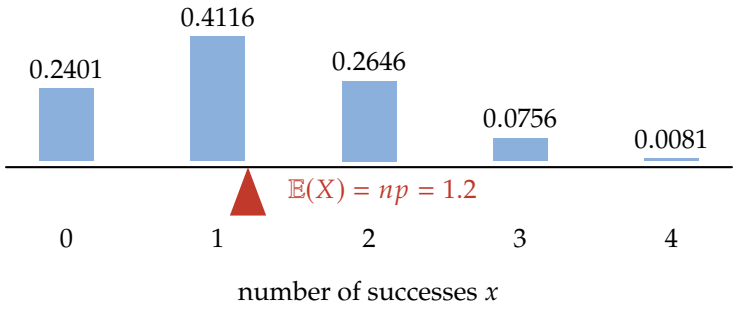


Figure 1.38: The pmf of a binomial with $n = 4$, $p = 0.3$, drawn as masses on a rod. The rod balances on the red fulcrum at $x = 1.2$, which is the mean. The mean is not one of the possible values, and it lies to the right of the most probable value $x = 1$ because the tail extends to the right.

(1) The integrand is even, so the integral is twice the integral over $x > 0$.
(2) $\frac{\mathrm{d}}{\mathrm{d}x} \ln(1 + x^2) = \frac{2x}{1+x^2}$, and the logarithm grows without bound.

([Wasserman Ex. 3.5, p. 48] reaches the same conclusion by parts.) The tails fall only like $1/x^2$, so $xf(x) \sim 1/(\pi x)$ and the integral diverges logarithmically. The positive and negative halves are each infinite, and "$\infty - \infty$" has no value. This is not a mathematical curiosity. In a simulation the running average never settles down: every so often a draw from far out in a tail arrives and moves the average by a finite amount, however many draws came before (Figure 1.39, the computer experiment of [Wasserman Exercise 3.9, p. 59]). After 10^4 draws the four Gaussian averages are all within 0.022 of zero, while the four Cauchy averages lie between 0.28 and 1.21 in absolute value and are still jumping. Section 1.14 will show the reason: the average of n Cauchy draws has exactly the same Cauchy distribution as a single draw.

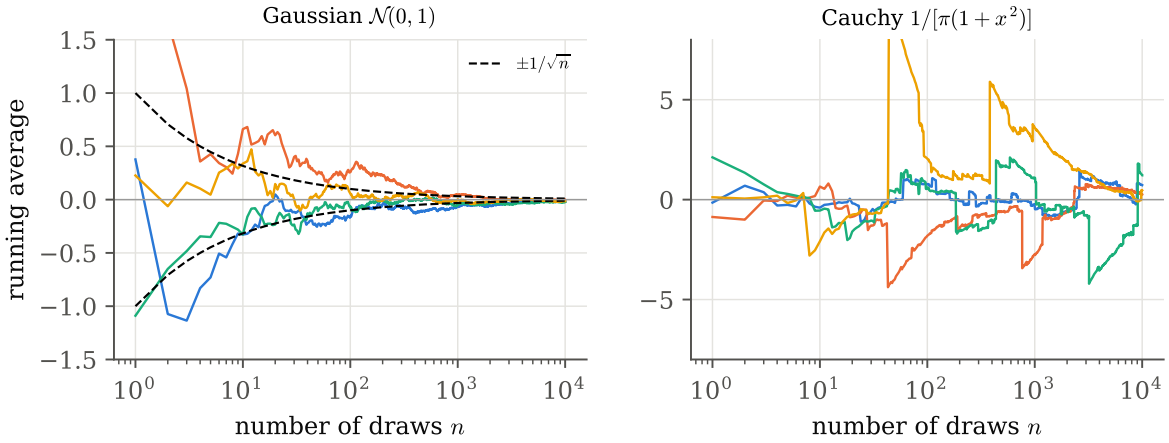


Figure 1.39: Running averages of four independent sequences of draws, on a logarithmic axis in n. Left: Gaussian draws; the averages close in on zero inside the dashed $\pm 1/\sqrt{n}$ envelope. Right: Cauchy draws; the averages wander and jump whenever a very large draw arrives, and do not converge.

From here on, whenever we write an expectation we assume it exists.

1.12.2 The average of a function: the lazy statistician

Often we want the mean of $Y = r(X)$: the kinetic energy $\frac{1}{2}mv^2$ of a molecule with random velocity, the flux $10^{-0.4m}$ of a star of random magnitude, the p_T of a decay product at a random angle. One route is to find f_Y by the change-of-variables rules of Section 1.11 and average y

against it. There is a much easier way [Wasserman Thm. 3.6, p. 48]:

$$\mathbb{E}\big(r(X)\big) = \sum_x r(x)\, f_X(x) \quad \text{or} \quad \mathbb{E}\big(r(X)\big) = \int r(x)\, f_X(x)\, \mathrm{d}x. \tag{1.41}$$

It is called the rule of the lazy statistician because it looks like what a lazy person would write without thinking, and it is right. Think of a game in which X is drawn and we are paid $r(X)$: our average income is each payment $r(x)$ times the chance of that x, summed.

Why does it hold? For discrete X [Wasserman Exercise 3.6, p. 59], group the values of x according to the value $y = r(x)$ they produce:

$$\mathbb{E}(Y) \overset{(1)}{=} \sum_y y\ \mathrm{P}(Y = y) \overset{(2)}{=} \sum_y y \sum_{x:\, r(x)=y} f_X(x) \overset{(3)}{=} \sum_y \sum_{x:\, r(x)=y} r(x)\, f_X(x) \overset{(4)}{=} \sum_x r(x)\, f_X(x).$$

(1) Definition (1.40) applied to Y.
(2) $\mathrm{P}(Y = y)$ is the total mass of all preimages of y (Section 1.11).
(3) Inside the inner sum $r(x) = y$, so y may be replaced by $r(x)$.
(4) Every x lies in exactly one set $\{x : r(x) = y\}$, so the double sum visits each x once.

For continuous X and increasing r with inverse s, the same thing happens with a Jacobian:

$$\mathbb{E}(Y) \overset{(1)}{=} \int y\, f_X\big(s(y)\big)\, s'(y)\, \mathrm{d}y \overset{(2)}{=} \int r(x)\, f_X(x)\, \mathrm{d}x.$$

(1) Definition of $\mathbb{E}(Y)$ with f_Y from (1.33).
(2) Substitute $x = s(y)$, so $y = r(x)$ and $\mathrm{d}x = s'(y)\, \mathrm{d}y$.

The Jacobian in the density is exactly cancelled by the Jacobian of the measure. In the fluid picture: $f_Y(y)\, \mathrm{d}y = f_X(x)\, \mathrm{d}x$ is the same drop of probability, and we multiply it by the same number $y = r(x)$ whichever coordinate we use [Cowan (1.40)]. A decreasing r or several branches give the same result branch by branch.

The most important special case is an indicator. Let $r = I_A$, the indicator of a set A ($I_A(x) = 1$ if $x \in A$ and 0 otherwise). Then

$$\mathbb{E}\big(I_A(X)\big) = \int I_A(x) f_X(x)\, \mathrm{d}x = \int_A f_X(x)\, \mathrm{d}x = \mathrm{P}(X \in A). \tag{1.42}$$

Probability is a special case of expectation. This is how every Monte Carlo estimate of a probability works (Chapter 6): the fraction of simulated draws that land in A is the sample mean of an indicator.

Example: $\mathbb{E}(e^X)$ two ways. Let $X \sim \mathrm{Uniform}(0, 1)$ and $Y = e^X$ [Wasserman Ex. 3.7, p. 49]. By (1.41), $\mathbb{E}(Y) = \int_0^1 e^x\, \mathrm{d}x = e - 1$. The long way first finds f_Y: the inverse is $s(y) = \ln y$ with $s'(y) = 1/y$, so (1.33) gives $f_Y(y) = 1/y$ on $(1, e)$, and $\mathbb{E}(Y) = \int_1^e y \cdot \frac{1}{y}\, \mathrm{d}y = e - 1$. Same answer, more work. A simulation with 10^6 draws gives 1.7189 (exact 1.7183). Note that $e^{\mathbb{E}(X)} = e^{1/2} = 1.649$ is *not* the answer.

Example: the broken stick. Break a stick of unit length at a uniformly random point X and keep the longer piece, $Y = \max\{X, 1 - X\}$ [Wasserman Ex. 3.8, p. 49]. On $x < \frac{1}{2}$ the longer piece is $1 - x$, on $x \geq \frac{1}{2}$ it is x:

$$\mathbb{E}(Y) \overset{(1)}{=} \int_0^{1/2} (1 - x)\, \mathrm{d}x + \int_{1/2}^1 x\, \mathrm{d}x \overset{(2)}{=} \left(\tfrac{1}{2} - \tfrac{1}{8}\right) + \left(\tfrac{1}{2} - \tfrac{1}{8}\right) = \tfrac{3}{4}.$$

(1) (1.41) with the density 1 on $(0,1)$, split where r changes formula.
(2) $\int_0^{1/2}(1-x)\mathrm{d}x = \frac{1}{2} - \frac{1}{8}$, and $\int_{1/2}^1 x\,\mathrm{d}x = \frac{1}{2} - \frac{1}{8}$.

The simulation gives 0.7498.

Example: time average and ensemble average for the oscillator. For the oscillator $x(t) = A\cos\omega t$ of Section 1.11.3 we can compute $\langle x^2\rangle$ in two ways. As an average over the position density $f(x) = 1/(\pi\sqrt{A^2-x^2})$:

$$\mathbb{E}(X^2) \overset{(1)}{=} \int_{-A}^{A} \frac{x^2}{\pi\sqrt{A^2-x^2}}\,\mathrm{d}x \overset{(2)}{=} \frac{1}{\pi}\int_{-\pi/2}^{\pi/2} \frac{A^2\sin^2\varphi}{A\cos\varphi}\,A\cos\varphi\,\mathrm{d}\varphi \overset{(3)}{=} \frac{A^2}{\pi}\cdot\frac{\pi}{2} = \frac{A^2}{2}.$$

(1) (1.41) with $r(x) = x^2$.
(2) Substitute $x = A\sin\varphi$: $\mathrm{d}x = A\cos\varphi\,\mathrm{d}\varphi$ and $\sqrt{A^2-x^2} = A\cos\varphi \geq 0$ on $[-\frac{\pi}{2}, \frac{\pi}{2}]$.
(3) $\int_{-\pi/2}^{\pi/2}\sin^2\varphi\,\mathrm{d}\varphi = \frac{\pi}{2}$ (the average of $\sin^2$ over a half period is $\frac{1}{2}$).

As an average over the uniform time T on one period $P = 2\pi/\omega$, again by (1.41) but now with $X = r(T)$: $\mathbb{E}(X^2) = \frac{1}{P}\int_0^P A^2\cos^2\omega t\,\mathrm{d}t = \frac{A^2}{2}$. The ensemble average over positions equals the time average over one period, a small instance of the ergodic idea that returns with Markov chains in Chapter 9. Multiplied by $\frac{1}{2}m\omega^2$ it says that the mean potential energy is half the total energy. The simulation with $A = 1$ gives 0.4995.

The rule extends to several variables [Wasserman (3.4), p. 49],

$$\mathbb{E}\big(r(X,Y)\big) = \iint r(x,y)\,f(x,y)\,\mathrm{d}x\,\mathrm{d}y.$$

For (X,Y) uniform on the unit square, $\mathbb{E}(X^2+Y^2) = \int_0^1\int_0^1 (x^2+y^2)\,\mathrm{d}x\,\mathrm{d}y = \frac{1}{3}+\frac{1}{3} = \frac{2}{3}$ [Wasserman Ex. 3.9, p. 49].

Watch out: The average of a function is not the function of the average

In general $\mathbb{E}\big(r(X)\big) \neq r\big(\mathbb{E}\,X\big)$: above, $\mathbb{E}(e^X) = 1.718$ but $e^{\mathbb{E}\,X} = 1.649$. Equality holds for linear r only. For a convex r (curving upwards, like x^2 or e^x) the inequality always runs one way, $\mathbb{E}\,r(X) \geq r(\mathbb{E}\,X)$ (Jensen's inequality, [CasellaBerger Thm. 4.7.7]), which is why $\langle v^2\rangle \geq \langle v\rangle^2$. Averaging magnitudes and then converting to flux, or averaging $\ln\mathcal{L}$ and exponentiating, gives biased answers for the same reason.

1.12.3 Linearity, and products of independent variables

The single most useful property of the expectation is that it is linear, whether or not the variables are independent [Wasserman Thm. 3.11, p. 50]. For two continuous variables with joint density $f(x,y)$:

$$\begin{aligned}\mathbb{E}(aX+bY) &\overset{(1)}{=} \iint (ax+by)\,f(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &\overset{(2)}{=} a\int x\Big[\int f(x,y)\,\mathrm{d}y\Big]\mathrm{d}x + b\int y\Big[\int f(x,y)\,\mathrm{d}x\Big]\mathrm{d}y \\ &\overset{(3)}{=} a\int x\,f_X(x)\,\mathrm{d}x + b\int y\,f_Y(y)\,\mathrm{d}y \overset{(4)}{=} a\,\mathbb{E}(X) + b\,\mathbb{E}(Y).\end{aligned}$$

(1) The lazy statistician for two variables, with $r(x,y) = ax+by$.

(2) Split the integral into two and do the inner integral over the variable that does not appear in front.
(3) The inner integrals are the marginal densities (Section 1.10).
(4) Definition (1.40).

No independence was used: marginalisation did all the work. By induction, $\mathbb{E}\big(\sum_i a_i X_i\big) = \sum_i a_i \,\mathbb{E}(X_i)$ for any number of variables, and $\mathbb{E}(aX + b) = a\,\mathbb{E}(X) + b$ because a constant is a variable with all its mass at one point.

Example: the binomial mean by indicators. The definition gives $\mathbb{E}(X) = \sum_{x=0}^{n} x\binom{n}{x}p^x(1-p)^{n-x}$, which is not an inviting sum. Instead write the number of successes as a sum of indicators, $X = \sum_{i=1}^{n} X_i$ with $X_i = 1$ if trial i succeeds [Wasserman Ex. 3.12, p. 50]. Each $\mathbb{E}(X_i) = p$, so linearity gives

$$\mathbb{E}(X) = \sum_{i=1}^{n} \mathbb{E}(X_i) = np.$$

This is the exact value 5 that the simulation of Section 1.8 found as 4.997 for ten fair flips. Linearity turned a combinatorial sum into one line. The same trick gives the mean count of any multinomial bin, $\mathbb{E}(X_j) = np_j$.

Example: the mean of the smeared energy. In the detector example of Example 1.43 the measured energy is $E_{\text{meas}} = E + \varepsilon$, the true energy plus a Gaussian error with mean 0. Linearity gives $\mathbb{E}(E_{\text{meas}}) = \mathbb{E}(E) + \mathbb{E}(\varepsilon) = \beta + 0 = 20\,\text{GeV}$ without touching the convolution integral (1.32). The simulation found 19.99 GeV.

For products the situation is different. If X and Y are independent then $f(x, y) = f_X(x)f_Y(y)$ and [Wasserman Thm. 3.13, p. 50]

$$\mathbb{E}(XY) \overset{(1)}{=} \iint xy\, f_X(x)f_Y(y)\,\mathrm{d}x\,\mathrm{d}y \overset{(2)}{=} \Big(\int x\, f_X(x)\,\mathrm{d}x\Big)\Big(\int y\, f_Y(y)\,\mathrm{d}y\Big) = \mathbb{E}(X)\;\mathbb{E}(Y). \tag{1.43}$$

(1) Lazy statistician with $r = xy$, and the factorised joint density.
(2) The integrand is a function of x times a function of y, so the double integral is a product of single integrals.

The same argument gives, for any functions g and h [CasellaBerger Thm. 4.2.10],

$$\mathbb{E}\big(g(X)h(Y)\big) = \mathbb{E}\big(g(X)\big)\;\mathbb{E}\big(h(Y)\big)$$

([SeeingTheory p. 32] gives the discrete version). Without independence it fails: if $X = Y$ is a fair coin indicator then $\mathbb{E}(XY) = \mathbb{E}(X^2) = \frac{1}{2}$, while $\mathbb{E}(X)\,\mathbb{E}(Y) = \frac{1}{4}$.

Key idea: What expectation respects

$\mathbb{E}\big(\sum_i a_i X_i + b\big) = \sum_i a_i\,\mathbb{E}(X_i) + b$ always. The mean of a product is the product of the means when the variables are independent, and not in general.

1.12.4 The mean as the best single guess

Suppose we must summarise X by one number b and we are penalised by the squared miss $(X - b)^2$. Which b minimises the expected penalty? Adding and subtracting $\mu = \mathbb{E}\,X$ answers it

without calculus [CasellaBerger Ex. 2.2.6]:

$$\begin{aligned}\mathbb{E}(X-b)^2 &\overset{(1)}{=} \mathbb{E}\big[(X-\mu)+(\mu-b)\big]^2 \\ &\overset{(2)}{=} \mathbb{E}(X-\mu)^2 + 2(\mu-b)\;\mathbb{E}(X-\mu) + (\mu-b)^2 \\ &\overset{(3)}{=} \mathbb{E}(X-\mu)^2 + (\mu-b)^2. \end{aligned} \tag{1.44}$$

(1) Add and subtract μ inside the bracket.
(2) Expand the square and use linearity; $\mu - b$ is a constant and comes out of the expectation.
(3) $\mathbb{E}(X-\mu) = \mathbb{E}\,X - \mu = 0$.

The first term does not depend on b and the second is smallest, zero, at $b = \mu$. So the mean is the best guess under squared loss, and the smallest achievable expected squared miss is $\mathbb{E}(X-\mu)^2$, which is the variance we define next [Kruschke §4.3.3.1]. Equation (1.44) is a Pythagoras theorem: total squared miss = spread about the mean + squared offset of the guess. It returns in Chapter 3 as "mean squared error = variance + bias^2".

If the penalty is the absolute miss $|X - M|$, the best guess is the median instead. Write $\mathbb{E}\,|X-M| = \int_{-\infty}^{M}(M-x)f(x)\,\mathrm{d}x + \int_{M}^{\infty}(x-M)f(x)\,\mathrm{d}x$ and differentiate in M:

$$\frac{\mathrm{d}}{\mathrm{d}M}\,\mathbb{E}\,|X-M| \overset{(1)}{=} \int_{-\infty}^{M} f(x)\,\mathrm{d}x - \int_{M}^{\infty} f(x)\,\mathrm{d}x \overset{(2)}{=} F(M) - \big(1-F(M)\big) = 2F(M) - 1.$$

(1) Leibniz rule. The boundary terms vanish because each integrand is zero at $x = M$, leaving the derivatives $+1$ and -1 of the integrands.
(2) Definition of the CDF F (Section 1.9).

The derivative is negative for $F(M) < \frac{1}{2}$ and positive after, so the minimum is at $F(M) = \frac{1}{2}$, the median. With a penalty of 0 for an exact hit and 1 for any miss, the best guess is the mode. These three loss functions reappear in Chapter 5, where they turn a posterior into the posterior mean, median and mode. Figure 1.40 checks the first two for an exponential with $\beta = 1$: the simulated squared loss is smallest at $M = 1.00$ (exact 1), where it equals 0.996 (the variance, exact 1), and the absolute loss is smallest at $M = 0.69$ (exact $\ln 2 = 0.693$).

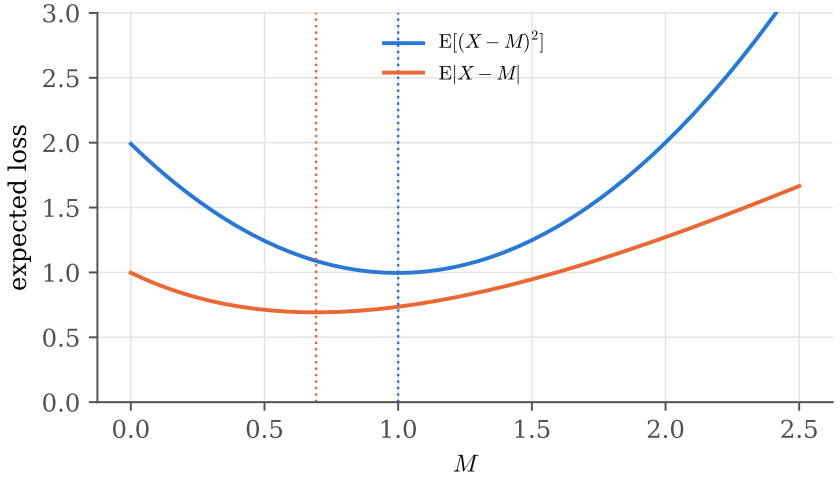


Figure 1.40: Expected squared miss and expected absolute miss of a guess M for exponential lifetimes with mean 1, computed by averaging over simulated draws. The squared miss bottoms out at the mean (blue dotted line), the absolute miss at the median $\ln 2$ (orange dotted line). The height of the blue minimum is the variance.

1.12.5 Variance and standard deviation

The mean says where a distribution sits. We also need how wide it is. The obvious candidate, the average deviation $\mathbb{E}(X-\mu)$, is useless: it is always $\mathbb{E}\,X-\mu=0$, because positive and negative deviations cancel by the definition of μ [Wasserman p. 50, footnote]. The average absolute deviation $\mathbb{E}\,|X-\mu|$ works but is awkward algebraically. The average *squared* deviation is the standard choice, and (1.44) has already shown why it is natural.

Working definition 1.23 (variance, standard deviation)

The **variance** of X is the mean squared deviation from the mean,

$$\operatorname{Var}(X)=\sigma^2=\mathbb{E}\big[(X-\mu)^2\big]=\int (x-\mu)^2 f(x)\,\mathrm{d}x, \tag{1.45}$$

[units of X squared]. The **standard deviation** is $\sigma=\sqrt{\operatorname{Var}(X)}$ [same units as X, a typical size of $X-\mu$]. *Operational test:* the average of $(X_j-\bar{X})^2$ over many draws.

Three properties do most of the work [Wasserman Thm. 3.15, p. 51]. First, the shortcut formula:

$$\operatorname{Var}(X)\overset{(1)}{=}\mathbb{E}\big[X^2-2\mu X+\mu^2\big]\overset{(2)}{=}\mathbb{E}(X^2)-2\mu\,\mathbb{E}(X)+\mu^2\overset{(3)}{=}\mathbb{E}(X^2)-\mu^2. \tag{1.46}$$

(1) Expand the square in (1.45).
(2) Linearity; μ is a constant.
(3) $\mathbb{E}(X)=\mu$, so $-2\mu^2+\mu^2=-\mu^2$.

Second, shifting does nothing and scaling scales: $\operatorname{Var}(aX+b)=\mathbb{E}\big[(aX+b-a\mu-b)^2\big]=a^2\,\mathbb{E}(X-\mu)^2=a^2\operatorname{Var}(X)$. Third, $\operatorname{Var}(X)\ge 0$, with equality only when X equals a constant with probability one, since a non-negative quantity $(X-\mu)^2$ can have zero mean only if it is zero with probability one [Wasserman Exercise 3.2, p. 59].

Example: Bernoulli, die, uniform, exponential. For a Bernoulli trial $X^2=X$, so $\mathbb{E}(X^2)=p$ and $\operatorname{Var}(X)=p-p^2=p(1-p)$, largest for a fair coin and zero when the outcome is certain. For a fair die

$$\operatorname{Var}(X)=\mathbb{E}(X^2)-\mu^2=\tfrac{1}{6}(1+4+9+16+25+36)-3.5^2=\tfrac{91}{6}-\tfrac{49}{4}=\tfrac{35}{12}\approx 2.92$$

[SeeingTheory p. 14].[4] For $X\sim\text{Uniform}(a,b)$:

$$\operatorname{Var}(X)\overset{(1)}{=}\frac{1}{b-a}\int_a^b x^2\,\mathrm{d}x-\left(\frac{a+b}{2}\right)^2\overset{(2)}{=}\frac{a^2+ab+b^2}{3}-\frac{a^2+2ab+b^2}{4}\overset{(3)}{=}\frac{(b-a)^2}{12}.$$

(1) Shortcut (1.46), with density $1/(b-a)$ and mean $(a+b)/2$.
(2) $\int_a^b x^2\mathrm{d}x=(b^3-a^3)/3=(b-a)(a^2+ab+b^2)/3$.
(3) Common denominator 12: $4a^2+4ab+4b^2-3a^2-6ab-3b^2=(b-a)^2$.

A physicist meets this constantly: a silicon strip detector of pitch w that records only which strip fired localises a track uniformly within the strip, so its position resolution is $w/\sqrt{12}$. For the exponential, $\mathbb{E}(X^2)=\beta^2\int_0^\infty u^2e^{-u}\mathrm{d}u=2\beta^2$ (parts twice), so $\operatorname{Var}(X)=2\beta^2-\beta^2=\beta^2$: the standard deviation of a lifetime equals its mean.

[4] Seeing Theory writes the variance as "$E[(X-EX)]^2$" in this exercise and as "$E(X^2)-E(X)$" in Proposition 0.0.1(c) on p. 12; both should read $E[(X-EX)^2]=E(X^2)-(EX)^2$, as in its own proof.

For a sum, expand the square. With $\mu_X = \mathbb{E}\,X$, $\mu_Y = \mathbb{E}\,Y$:

$$\begin{aligned}\operatorname{Var}(X+Y) &\overset{(1)}{=} \mathbb{E}\big[(X-\mu_X)+(Y-\mu_Y)\big]^2 \\ &\overset{(2)}{=} \operatorname{Var}(X)+\operatorname{Var}(Y)+2\,\mathbb{E}\big[(X-\mu_X)(Y-\mu_Y)\big] \\ &\overset{(3)}{=} \operatorname{Var}(X)+\operatorname{Var}(Y) \qquad (X, Y \text{ independent}).\end{aligned} \tag{1.47}$$

(1) The mean of $X+Y$ is $\mu_X+\mu_Y$ by linearity.
(2) Expand and use linearity.
(3) For independent variables (1.43) with $g(X) = X-\mu_X$, $h(Y) = Y-\mu_Y$ gives $\mathbb{E}(X-\mu_X)\,\mathbb{E}(Y-\mu_Y) = 0\cdot 0$.

The cross term is the covariance, the subject of Section 1.13. For independent variables it vanishes and variances add, so that $\operatorname{Var}\big(\sum_i a_i X_i\big) = \sum_i a_i^2 \operatorname{Var}(X_i)$ [Wasserman (3.8), p. 51]. Standard deviations do not add; they add *in quadrature*. Every "add the errors in quadrature" in a laboratory is this line, valid when the error sources are independent.

Example: the binomial variance, and the smeared spectrum. With $X = \sum_i X_i$ a sum of n independent indicators, each with variance $p(1-p)$, we get $\operatorname{Var}(X) = np(1-p)$ [Wasserman Ex. 3.16, p. 51]. It vanishes for $p = 0$ or $p = 1$, where every trial has a certain outcome. For the smeared energy $E_{\text{meas}} = E + \varepsilon$ the detector error does not depend on the true energy, so $\operatorname{Var}(E_{\text{meas}}) = \operatorname{Var}(E) + \operatorname{Var}(\varepsilon) = \beta^2 + \sigma^2 = 400 + 25 = 425\,\text{GeV}^2$, as the simulation of Example 1.43 found (425.0). The detector resolution adds in quadrature to the intrinsic width of the spectrum.

1.12.6 The sample mean and the sample variance

In practice we do not know μ and σ^2. We have n independent draws $X_1, \dots, X_n$ from the same distribution (an iid sample, Section 1.10) and form the **sample mean** and the **sample variance** [Wasserman (3.9)–(3.10), p. 51],

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\big(X_i - \bar{X}_n\big)^2. \tag{1.48}$$

They are functions of the data, so they are random variables with their own distributions. A function of the data is called a **statistic**, and its distribution over repeated experiments is its **sampling distribution** [Wasserman Exercise 3.19, p. 61]. In the chain from a physical model to inferred physics, this is the step "construct a summary S, then derive its sampling distribution $P(S \mid \theta)$", in its simplest form. The rules above already give the first two moments [Wasserman Thm. 3.17, p. 52]:

$$\begin{aligned}\mathbb{E}(\bar{X}_n) &\overset{(1)}{=} \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \mu, \\ \operatorname{Var}(\bar{X}_n) &\overset{(2)}{=} \frac{1}{n^2}\sum_{i=1}^{n}\operatorname{Var}(X_i) = \frac{\sigma^2}{n}.\end{aligned} \tag{1.49}$$

(1) Linearity.
(2) The draws are independent, so variances add, and each carries the factor $(1/n)^2$ from scaling.

The spread of the sample mean shrinks like $\sigma/\sqrt{n}$, the **standard error of the mean**: four times the data halves the error bar. For the sample variance we need $\mathbb{E}\big[\sum_i (X_i - \bar{X})^2\big]$:

$$\mathbb{E}\Big[\sum_{i=1}^{n}(X_i-\bar{X})^2\Big] \overset{(1)}{=} \mathbb{E}\Big[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\Big] \overset{(2)}{=} n\,\mathbb{E}(X_1^2) - n\,\mathbb{E}(\bar{X}^2)$$
$$\overset{(3)}{=} n(\sigma^2+\mu^2) - n\Big(\frac{\sigma^2}{n}+\mu^2\Big) = (n-1)\,\sigma^2.$$

(1) Expand: $\sum_i (X_i - \bar{X})^2 = \sum_i X_i^2 - 2\bar{X}\sum_i X_i + n\bar{X}^2$, and $\sum_i X_i = n\bar{X}$.
(2) Linearity; all X_i have the same distribution.
(3) The shortcut (1.46) read backwards, $\mathbb{E}(Z^2) = \mathrm{Var}(Z) + (\mathbb{E}\,Z)^2$, for $Z = X_1$ and for $Z = \bar{X}$ using (1.49).

Dividing by $n-1$ therefore gives $\mathbb{E}(S_n^2) = \sigma^2$. Why $n-1$ and not n? The deviations are measured from $\bar{X}$, which is computed from the same data. By (1.44) applied to the n data points, the sample mean is the point that makes $\sum_i (X_i - b)^2$ as small as possible. So deviations from $\bar{X}$ are systematically smaller than deviations from the true μ, by exactly one draw's worth of variance. In Python, `np.var(x)` divides by n unless one passes `ddof=1`, while `np.cov` divides by $n-1$ by default.

1.12.7 Conditional expectation and the law of total variance

What is the mean of Y among those experiments in which $X = x$? Compute the mean as before, but with the conditional density $f(y \mid x)$ of Section 1.10 in place of $f_Y(y)$ [Wasserman Def. 3.22, p. 54]:[5]

$$\mathbb{E}(Y \mid X = x) = \int y\, f(y \mid x)\,\mathrm{d}y \quad \Big(\text{or } \sum_y y\, f(y \mid x)\Big). \tag{1.50}$$

Similarly $\mathbb{E}\big(r(X,Y) \mid X = x\big) = \int r(x,y) f(y \mid x)\,\mathrm{d}y$. The result is a number for each x, that is, a function of x. Before X is observed we do not know which number it will be, so $\mathbb{E}(Y \mid X)$, the function evaluated at the random X, is itself a random variable. This is the subtle point of the definition [Wasserman p. 54, Warning]. $\mathbb{E}(Y)$ is a number, $\mathbb{E}(Y \mid X = x)$ is a function, and $\mathbb{E}(Y \mid X)$ is a random variable. Physically, $\mathbb{E}(Y \mid X = x)$ is the curve through the middle of the scatter plot of Y against X, the regression curve. Figure 1.44 traces it by binning.

Example: a two-stage draw. Draw $X \sim \mathrm{Uniform}(0,1)$ and then $Y \mid X = x \sim \mathrm{Uniform}(x,1)$ [Wasserman Ex. 3.23, p. 54]. The conditional density is $1/(1-x)$ on $(x,1)$, so

$$\mathbb{E}(Y \mid X = x) \overset{(1)}{=} \int_x^1 \frac{y}{1-x}\,\mathrm{d}y \overset{(2)}{=} \frac{1}{1-x}\cdot\frac{1-x^2}{2} \overset{(3)}{=} \frac{1+x}{2},$$

(1) (1.50) with the uniform conditional density.
(2) $\int_x^1 y\,\mathrm{d}y = (1-x^2)/2$.
(3) $1 - x^2 = (1-x)(1+x)$.

the midpoint of $(x,1)$, as it must be. So $\mathbb{E}(Y \mid X) = (1+X)/2$, a random variable.

Averaging the conditional mean over X gives back the unconditional mean. This is the **law of iterated expectations** [Wasserman Thm. 3.24, p. 55]:

$$\mathbb{E}\big[\mathbb{E}(Y \mid X)\big] \overset{(1)}{=} \int \Big[\int y\, f(y \mid x)\,\mathrm{d}y\Big] f_X(x)\,\mathrm{d}x \overset{(2)}{=} \iint y\, f(x,y)\,\mathrm{d}y\,\mathrm{d}x \overset{(3)}{=} \mathbb{E}(Y). \tag{1.51}$$

[5] [Wasserman (3.13)–(3.14), p. 54] has a stray "dx" after the two discrete sums; there is no integral in the discrete case.

(1) The outer expectation of the function $x \mapsto \mathbb{E}(Y \mid X = x)$, by the lazy statistician.
(2) Product rule $f(y \mid x) f_X(x) = f(x, y)$ (1.29).
(3) Lazy statistician with $r(x, y) = y$.

It is the law of total probability of Section 1.5 for averages: the overall mean is the mean in each branch weighted by the probability of the branch. For the two-stage draw, $\mathbb{E}(Y) = \mathbb{E}\big[(1 + X)/2\big] = (1 + \frac{1}{2})/2 = \frac{3}{4}$ [Wasserman Ex. 3.25, p. 55], with no joint density needed. A useful companion rule: anything already known given X comes out of the conditional expectation, $\mathbb{E}\big(r(X)s(Y) \mid X\big) = r(X)\, \mathbb{E}\big(s(Y) \mid X\big)$ [Wasserman Exercise 3.16, p. 60].

The spread splits in the same way. The **conditional variance** is [Wasserman Def. 3.26, p. 55]

$$\operatorname{Var}(Y \mid X = x) = \int \big(y - \mu(x)\big)^2 f(y \mid x)\, \mathrm{d}y, \qquad \mu(x) = \mathbb{E}(Y \mid X = x).$$

Then [CasellaBerger Thm. 4.4.7]

$$\begin{aligned} \operatorname{Var}(Y) &\overset{(1)}{=} \mathbb{E}\big[(Y - \mu(X)) + (\mu(X) - \mathbb{E}\,Y)\big]^2 \\ &\overset{(2)}{=} \mathbb{E}\big[(Y - \mu(X))^2\big] + \mathbb{E}\big[(\mu(X) - \mathbb{E}\,Y)^2\big] + 2\,\mathbb{E}\big[(Y - \mu(X))(\mu(X) - \mathbb{E}\,Y)\big] \\ &\overset{(3)}{=} \mathbb{E}\big[\operatorname{Var}(Y \mid X)\big] + \operatorname{Var}\big[\mathbb{E}(Y \mid X)\big] + 0. \end{aligned} \tag{1.52}$$

(1) Add and subtract $\mu(X) = \mathbb{E}(Y \mid X)$.
(2) Expand the square; linearity.
(3) First term: iterate, $\mathbb{E}[(Y - \mu(X))^2] = \mathbb{E}\big[\mathbb{E}\big((Y - \mu(X))^2 \mid X\big)\big] = \mathbb{E}[\operatorname{Var}(Y \mid X)]$. Second term: $\mu(X)$ has mean $\mathbb{E}\,Y$ by (1.51), so this is its variance. Third term: iterate again; given X the factor $\mu(X) - \mathbb{E}\,Y$ is known and comes out, leaving $\mathbb{E}\big(Y - \mu(X) \mid X\big) = \mu(X) - \mu(X) = 0$.

This is the **law of total variance** [Wasserman Thm. 3.27, p. 55]: the total spread is the average spread within each group plus the spread of the group means. A detector with a stable resolution but a calibration that drifts from run to run shows exactly this: within-run scatter plus run-to-run scatter of the mean response.

Example: counts with a random efficiency. A pixel counts how many of $n = 20$ photons it detects. Its efficiency Q is not known and varies from pixel to pixel, uniformly on $(0, 1)$ for the sake of the example; given $Q = q$ the count is $K \mid Q = q \sim \text{Binomial}(n, q)$. ([Wasserman Ex. 3.28, p. 56] tells the same story with counties and a disease rate.) A model built in stages like this is a **hierarchical model**. Then $\mathbb{E}(K \mid Q) = nQ$ and $\operatorname{Var}(K \mid Q) = nQ(1 - Q)$, so

$$\begin{aligned} \mathbb{E}(K) &\overset{(1)}{=} \mathbb{E}(nQ) = \frac{n}{2}, \\ \operatorname{Var}(K) &\overset{(2)}{=} \mathbb{E}\big[nQ(1 - Q)\big] + \operatorname{Var}(nQ) \overset{(3)}{=} n\Big(\frac{1}{2} - \frac{1}{3}\Big) + \frac{n^2}{12} = \frac{n}{6} + \frac{n^2}{12}. \end{aligned}$$

(1) Iterated expectations (1.51); $\mathbb{E}\,Q = \frac{1}{2}$.
(2) Law of total variance (1.52).
(3) $\mathbb{E}(Q) - \mathbb{E}(Q^2) = \frac{1}{2} - \frac{1}{3}$, and $\operatorname{Var}(nQ) = n^2 \operatorname{Var}(Q) = n^2/12$ from the uniform variance with $b - a = 1$.

For $n = 20$ this is 36.67, far larger than the variance $np(1 - p) = 5$ of a binomial with the same mean and a fixed efficiency of $\frac{1}{2}$. Most of the spread is the spread of the efficiencies, the $n^2/12$ term. The whole experiment takes three lines of Python:

Listing 1.16: A two-stage (hierarchical) draw (`code/ch01/16_expectation_checks.py`)

```
n = 20
Q = rng.uniform(size=N)
K = rng.binomial(n, Q)
K_plain = rng.binomial(n, 0.5, size=N)
var_exact = n / 6 + n**2 / 12
```

Line 31 draws one efficiency per pixel. Line 32 draws a binomial count *with that pixel's efficiency*: numpy broadcasts the array `Q`, so each count gets its own q. Line 33 is the comparison with a fixed efficiency. With 10^6 pixels the simulated mean is 9.997 (exact 10), the simulated variance is 36.67 (exact 36.67), and the fixed-efficiency variance is 4.998 (exact 5).

Figure 1.41 shows something more striking: the distribution of K is flat. To see why, marginalise the nuisance efficiency, exactly as in Section 1.10:

$$\mathrm{P}(K=k) \overset{(1)}{=} \int_0^1 \binom{n}{k} q^k (1-q)^{n-k}\,\mathrm{d}q \overset{(2)}{=} \binom{n}{k} \frac{k!\,(n-k)!}{(n+1)!} = \frac{1}{n+1}.$$

(1) Joint $= f(q)\ \mathrm{P}(K=k \mid q)$ with $f(q)=1$; integrate out q.

(2) The beta integral $I(a,b) = \int_0^1 q^a (1-q)^b\,\mathrm{d}q = \frac{a!\,b!}{(a+b+1)!}$. Integration by parts gives $I(a,b) = \frac{b}{a+1} I(a+1,b-1)$; repeating b times gives $I(a,b) = \frac{b!}{(a+1)\cdots(a+b)}\, I(a+b,0)$, and $I(a+b,0) = \frac{1}{a+b+1}$.

Every count from 0 to 20 has probability 1/21. The ignorance about Q has been passed on to K. In Chapter 5 the same calculation reappears as the *prior predictive* distribution of a binomial experiment with a flat prior on its success probability.

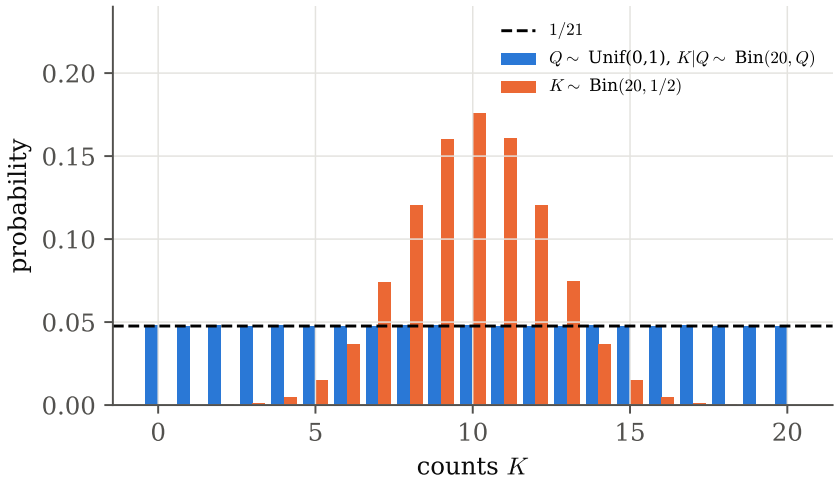


Figure 1.41: Counts out of 20 photons. Blue: each pixel has its own efficiency, drawn uniformly; the counts are spread evenly over $0, \ldots, 20$ at the height 1/21 of the dashed line. Orange: every pixel has efficiency $\frac{1}{2}$; the counts crowd around 10. Both have mean 10, but the variances are 36.7 and 5.

The cross term $\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]$ that appeared in the variance of a sum measures how two variables move together. Together with the averages and spreads above, it describes the *relation* between two variables at the level of second moments.

1.13 Covariance and correlation: how two variables move together

Two readout channels of a calorimeter sit on the same electronics board. Each channel records its own signal plus a common pedestal noise that the board adds to both: $X_1 = S_1 + N$ and $X_2 = S_2 + N$, with S_1, S_2 and N independent. Looked at one at a time, each channel is just a

noisy number. Looked at together, they tend to go up and down together, because whenever the pedestal N fluctuates high, both readings are pulled up. Nothing about the individual distributions reveals this. It is a property of the joint distribution, and the variance of a sum, (1.47), has already shown us the number that measures it: the cross term $\mathbb{E}\big[(X-\mu_X)(Y-\mu_Y)\big]$.

That number is the covariance. Its dimensionless version is the correlation coefficient. Both detect linear co-movement and miss everything else. For many variables the second moments collect into a covariance matrix, and a linear change of variables moves it by the rule $\mathbf{C}' = A\mathbf{C}A^{\mathsf{T}}$. Every error bar on a derived quantity, every Fisher forecast of Chapter 10 and every χ^2 of correlated band powers in Chapters 7 and 8 is built from this matrix. Three questions organise the topic.

1. How do we measure whether "X above its mean" goes with "Y above its mean"?
2. What does a zero value of that measure tell us, and what does it not?
3. For a vector of variables, how do the second moments change when we form linear combinations, and can we always find combinations that are uncorrelated?

1.13.1 Covariance and the correlation coefficient

Working definition 1.24 (covariance, correlation coefficient)

For X, Y with means μ_X, μ_Y and standard deviations σ_X, σ_Y, the **covariance** and the **correlation coefficient** are

$$\operatorname{Cov}(X,Y) = \mathbb{E}\big[(X-\mu_X)(Y-\mu_Y)\big], \qquad \rho_{XY} = \frac{\operatorname{Cov}(X,Y)}{\sigma_X\,\sigma_Y}. \tag{1.53}$$

Cov has the units of X times those of Y [positive if the two tend to be on the same side of their means]. ρ is a pure number between -1 and 1. *Operational test:* the average of $(x_j-\bar{x})(y_j-\bar{y})$ over many simulated pairs, `np.cov` and `np.corrcoef` in Python.

The definition is read off a scatter plot [Cowan pp. 18–19]. Put the origin at (μ_X, μ_Y) (Figure 1.42). A point in the upper-right or lower-left quadrant has deviations of the same sign, so its product $(x-\mu_X)(y-\mu_Y)$ is positive. A point in the other two quadrants contributes a negative product. The covariance is the probability-weighted balance of the two. A cloud tilted up to the right puts more mass in the positive quadrants and has Cov > 0. A cloud tilted down has Cov < 0. A cloud with as much mass in the negative quadrants as in the positive ones has Cov $= 0$, whatever its shape.

Expanding the product gives the form used in computation [Wasserman Thm. 3.19, p. 52]:

$$\begin{aligned}\operatorname{Cov}(X,Y) &\overset{(1)}{=} \mathbb{E}\big[XY - \mu_Y X - \mu_X Y + \mu_X\mu_Y\big] \\ &\overset{(2)}{=} \mathbb{E}(XY) - \mu_Y\mu_X - \mu_X\mu_Y + \mu_X\mu_Y = \mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y).\end{aligned} \tag{1.54}$$

(1) Multiply out $(X-\mu_X)(Y-\mu_Y)$.
(2) Linearity; μ_X, μ_Y are constants.

With $Y = X$ this is the variance shortcut (1.46): $\operatorname{Cov}(X,X) = \operatorname{Var}(X)$. Constants shifted onto either variable drop out, and the covariance is linear in each argument separately:

$$\operatorname{Cov}\Big(\sum_i a_i X_i,\ \sum_j b_j Y_j\Big) = \sum_i \sum_j a_i b_j\ \operatorname{Cov}(X_i, Y_j) \tag{1.55}$$

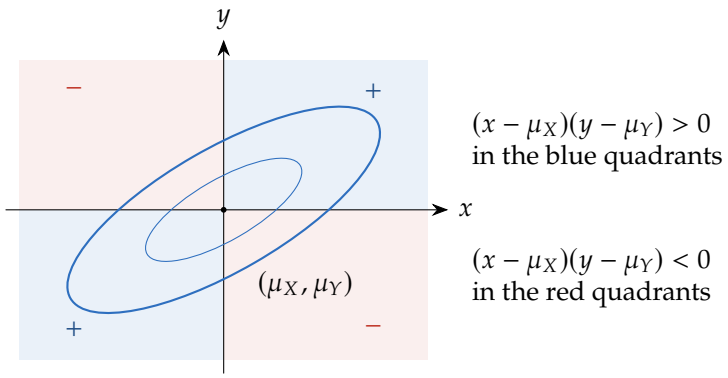


Figure 1.42: Reading the sign of a covariance. The ellipses are contours of a density tilted up to the right. Most of its probability sits in the blue quadrants, where the two deviations have the same sign, so the average of their product, the covariance, is positive.

[Wasserman Exercise 3.14, p. 60]. To see it, note that the deviation of $\sum_i a_i X_i$ from its mean is $\sum_i a_i(X_i - \mu_{X_i})$, do the same for the Y's, multiply the two sums term by term and take the expectation of each term. Setting both sums equal gives the variance of any linear combination [Wasserman Thm. 3.20, p. 52],

$$\operatorname{Var}\Big(\sum_i a_i X_i\Big) = \sum_i a_i^2 \operatorname{Var}(X_i) + 2\sum_{i<j} a_i a_j \operatorname{Cov}(X_i, X_j), \tag{1.56}$$

and in particular $\operatorname{Var}(X \pm Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) \pm 2\operatorname{Cov}(X, Y)$.

Example: the shared pedestal. Return to the two calorimeter channels of the opening, $X_1 = S_1 + N$ and $X_2 = S_2 + N$, which share the pedestal noise N. Take $\operatorname{Var}(S_1) = \operatorname{Var}(S_2) = \sigma_S^2$ and $\operatorname{Var}(N) = \sigma_N^2$. By (1.55),

$$\operatorname{Cov}(X_1, X_2) \overset{(1)}{=} \operatorname{Cov}(S_1, S_2) + \operatorname{Cov}(S_1, N) + \operatorname{Cov}(N, S_2) + \operatorname{Cov}(N, N) \overset{(2)}{=} \sigma_N^2,$$

(1) Bilinearity with $X_1 = S_1 + N$, $X_2 = S_2 + N$.
(2) Independent variables have zero covariance (shown below); $\operatorname{Cov}(N, N) = \operatorname{Var}(N)$.

so $\rho = \sigma_N^2/(\sigma_S^2 + \sigma_N^2)$, which is $\frac{1}{2}$ when the signal and the pedestal fluctuate equally. Now form the difference. By (1.56), $\operatorname{Var}(X_1 - X_2) = 2(\sigma_S^2 + \sigma_N^2) - 2\sigma_N^2 = 2\sigma_S^2$: the common noise has cancelled completely. The sum has $\operatorname{Var}(X_1 + X_2) = 2\sigma_S^2 + 4\sigma_N^2$: it is doubled. This is common-mode rejection, and it is why differential measurements are powerful. A gravitational-wave interferometer reads the *difference* of its two arm lengths, so the laser noise common to both arms cancels to first order.

1.13.2 Why $|\rho| \le 1$, and what $\rho = \pm 1$ means

The correlation coefficient removes the units by dividing by both standard deviations. That it then lies in $[-1, 1]$ follows from one observation [CasellaBerger Thm. 4.5.7]: the mean of a square is never negative. For every real t let

$$h(t) \overset{(1)}{=} \mathbb{E}\Big[\big(t(X - \mu_X) + (Y - \mu_Y)\big)^2\Big] \overset{(2)}{=} t^2\sigma_X^2 + 2t\ \operatorname{Cov}(X, Y) + \sigma_Y^2 \overset{(3)}{\ge} 0.$$

(1) Definition of h, the mean square of a combination of the two deviations.
(2) Expand the square and use linearity.

(3) The expectation of a non-negative quantity is non-negative.

A quadratic in t that is never negative has at most one real root, so its discriminant is not positive: $4\,\mathrm{Cov}(X,Y)^2 - 4\sigma_X^2\sigma_Y^2 \le 0$, that is, $|\mathrm{Cov}(X,Y)| \le \sigma_X\sigma_Y$, or $-1 \le \rho \le 1$. Equality, $|\rho| = 1$, means the discriminant is zero, so $h(t_0) = 0$ for one t_0, namely $t_0 = -\mathrm{Cov}(X,Y)/\sigma_X^2$. A non-negative variable with zero mean is zero with probability one, so $Y - \mu_Y = -t_0(X - \mu_X)$: the points lie exactly on a straight line with slope $a = -t_0 = \mathrm{Cov}/\sigma_X^2$, which has the sign of ρ. Conversely if $Y = aX + b$ then $\mathrm{Cov}(X,Y) = a\sigma_X^2$ and $\sigma_Y = |a|\,\sigma_X$, so $\rho = a/|a| = \pm 1$ [Wasserman Thm. 3.19, p. 52].

In the language of vectors this is the Cauchy–Schwarz inequality. Centred random variables form a vector space with the inner product $\langle U, V\rangle = \mathbb{E}(UV)$, the length of U is its standard deviation, and ρ is the cosine of the angle between $X - \mu_X$ and $Y - \mu_Y$. It is ± 1 when the two "vectors" are parallel, 0 when they are orthogonal. A quantum physicist will recognise the normalised overlap $\big|\langle\psi|\phi\rangle\big| \,/ (\|\psi\|\,\|\phi\|) \le 1$. Because it is a cosine, ρ is unchanged by shifting or by rescaling either variable with a positive factor: it does not care whether energies are in GeV or TeV.

Figure 1.43 calibrates the eye. With 1500 points from bivariate Gaussians of unit variances and $\rho = -0.8,\ 0,\ 0.5,\ 0.95$, the sample correlations are -0.795, 0.040, 0.486 and 0.949. A correlation of 0.5 is a visibly tilted but still fat cloud; it takes ρ close to 1 to make a thin line.

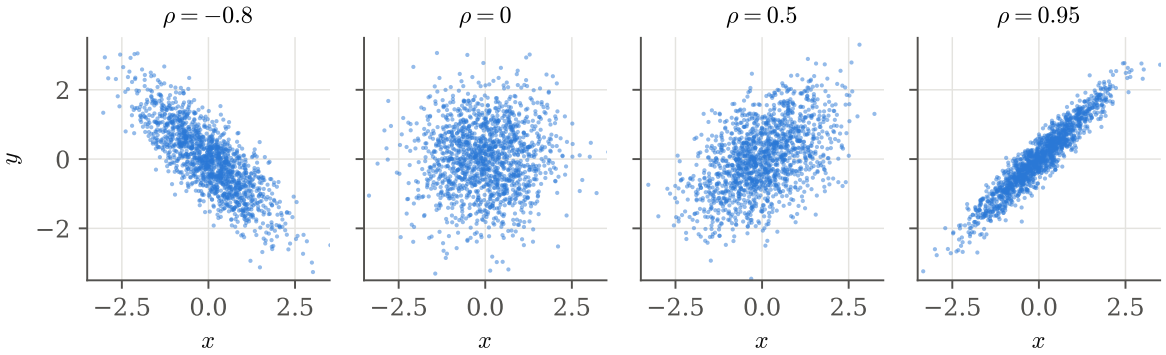


Figure 1.43: Clouds of 1500 draws from bivariate Gaussians with unit variances and four values of the correlation coefficient. The sign of ρ gives the direction of the tilt; its size says how tightly the points hug a straight line.

Why a nonlinear response can give a strong linear correlation. Suppose a calorimeter converts deposited energy X into a signal $Y = f(X)$. Its calibration curve may bend appreciably over the full energy range, yet a beam with a narrow energy spread samples only a short stretch of that curve. Let $\mu_X = \mathbb{E}(X)$ and write $\delta X = X - \mu_X$. Expanding the response about the beam mean and then subtracting its expectation gives

$$Y = f(\mu_X) + f'(\mu_X)\,\delta X + \tfrac{1}{2}f''(\mu_X)(\delta X)^2 + \cdots,$$
$$\delta Y \equiv Y - \mathbb{E}(Y) = f'(\mu_X)\,\delta X + \tfrac{1}{2}f''(\mu_X)\big[(\delta X)^2 - \sigma_X^2\big] + \cdots.$$

The subtraction matters because curvature can shift the mean signal away from $f(\mu_X)$. If the slope varies little over the sampled energies and is nonzero there, the first term dominates the fluctuations. Multiplying by δX and averaging then gives

$$\mathrm{Cov}(X,Y) = \mathbb{E}(\delta X\,\delta Y) \simeq f'(\mu_X)\;\mathbb{E}\big[(\delta X)^2\big] = f'(\mu_X)\,\sigma_X^2.$$

The ratio $\mathrm{Cov}(X,Y)/\sigma_X^2$ is the slope of the best mean-square straight-line fit, so in this narrow beam it approximates the local gain in signal per unit energy. Pearson's ρ gives the slope when

both variables are measured in their own standard deviations. An increase of one σ_X in energy therefore corresponds to an increase of $\rho\,\sigma_Y$ in the fitted signal, with the sign of ρ setting the direction. Independent, additive readout noise leaves the covariance with X unchanged but increases σ_Y, so it can reduce $|\rho|$ even when the gain is unchanged. The same local expansion underlies error propagation in Section 3.6.

If the beam spans a strongly curved part of the calibration, or is centred where $f'(\mu_X) = 0$, a single slope can miss an important part of the response. A quadratic fluctuation, for example, can be detected by examining how the mean signal varies with energy or by measuring its covariance with $(\delta X)^2$. Such higher-order summaries can require more data because they give unusually large fluctuations more weight.

1.13.3 Zero correlation is not independence

If X and Y are independent, $\mathbb{E}(XY) = \mathbb{E}(X)\,\mathbb{E}(Y)$ by (1.43), and (1.54) gives $\mathrm{Cov}(X, Y) = 0$. The converse is false [Wasserman Thm. 3.19, p. 52]. The standard counterexample shows why. Let $X \sim \mathcal{N}(0, 1)$ and $Y = X^2$. Then Y is completely determined by X, the strongest dependence there is. Yet

$$\mathrm{Cov}(X, Y) \overset{(1)}{=} \mathbb{E}(X \cdot X^2) - \mathbb{E}(X)\;\mathbb{E}(X^2) \overset{(2)}{=} \mathbb{E}(X^3) - 0 \cdot 1 \overset{(3)}{=} 0.$$

(1) Shortcut (1.54) with $Y = X^2$.
(2) $\mathbb{E}(X) = 0$ and $\mathbb{E}(X^2) = \mathrm{Var}(X) = 1$.
(3) $\mathbb{E}(X^3) = \int x^3 \varphi(x)\,\mathrm{d}x = 0$: the integrand is odd because the Gaussian density φ is even.

The covariance vanishes because the relation is symmetric: for every point (x, x^2) on the right branch of the parabola there is an equally likely point $(-x, x^2)$ on the left, and their products of deviations cancel [Cowan Fig. 1.10]. The correlation coefficient measures how well a *straight line* fits, and the best straight line through a symmetric parabola is flat.

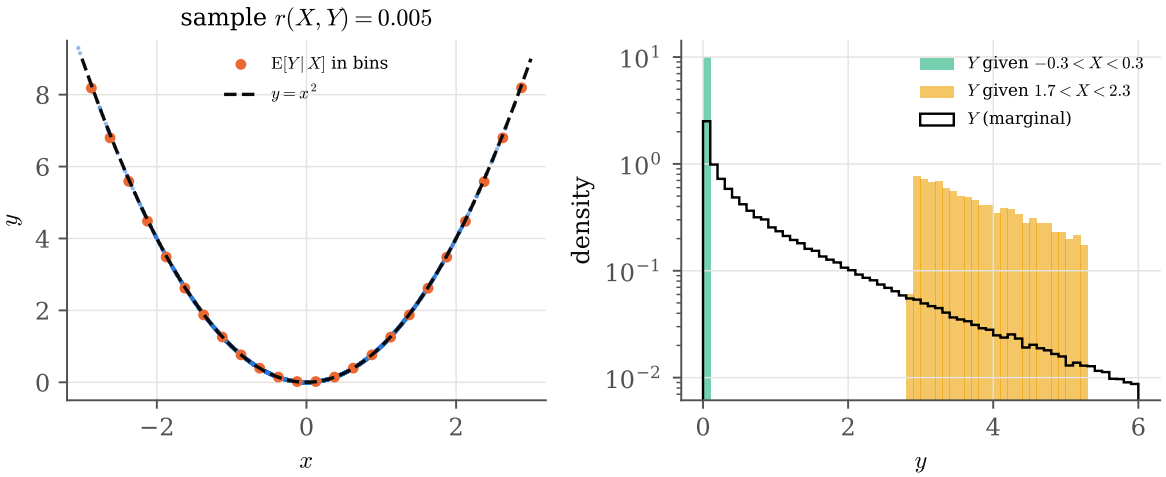


Figure 1.44: Left: draws of (X, X^2) with X Gaussian. The points lie exactly on the parabola, the binned conditional mean $\mathbb{E}(Y \mid X)$ (orange) follows it, and yet the sample correlation is essentially zero. Right: the distribution of Y for X near 0 and for X near 2 (logarithmic density axis). The two conditional distributions do not even overlap, so knowing X changes everything we know about Y, which is the opposite of independence.

Figure 1.44 shows it numerically. With 2×10^5 draws the sample correlation of X and Y is 0.0046. The dependence is all there: the correlation of Y with $|X|$ (a monotone function of X on each branch) is 0.9359, and the conditional distributions of Y in two slices of X are

completely different. [CasellaBerger Ex. 4.5.9] gives a version with noise, $Y = X^2 + Z$, with the same conclusion. Operationally: before concluding that two quantities are unrelated because their correlation is small, plot them.

There is a ladder of conditions between independence and zero correlation. If the conditional mean of X does not depend on Y, $\mathbb{E}(X \mid Y = y) = c$ for all y, then X and Y are uncorrelated [Wasserman Exercise 3.18, p. 61]:

$$\mathbb{E}(XY) \overset{(1)}{=} \mathbb{E}\big[\mathbb{E}(XY \mid Y)\big] \overset{(2)}{=} \mathbb{E}\big[Y\ \mathbb{E}(X \mid Y)\big] \overset{(3)}{=} c\ \mathbb{E}(Y) \overset{(4)}{=} \mathbb{E}(X)\ \mathbb{E}(Y).$$

(1) Iterated expectations (1.51).
(2) Given Y, the factor Y is known and comes out.
(3) $\mathbb{E}(X \mid Y) = c$.
(4) $\mathbb{E}(X) = \mathbb{E}[\mathbb{E}(X \mid Y)] = c$.

The same two steps show that if $\mathbb{E}(Y \mid X) = X$ then $\mathbb{E}(XY) = \mathbb{E}(X^2)$ and $\mathbb{E}(Y) = \mathbb{E}(X)$, so $\mathrm{Cov}(X, Y) = \mathrm{Var}(X)$ [Wasserman Exercise 3.21, p. 61]. So independence implies a constant conditional mean, which implies zero correlation, and neither implication reverses (in the parabola example $\mathbb{E}(Y \mid X) = X^2$ is far from constant).

For one family the ladder collapses. If (X, Y) is bivariate Gaussian (1.28) with $\rho = 0$, the cross term $-2\rho xy/(\sigma_x\sigma_y)$ in the exponent vanishes, the exponent splits into an x part plus a y part, and the density factorises into $f_X(x)f_Y(y)$. For jointly Gaussian variables, uncorrelated means independent. This is why the sum and the difference of two equally noisy Gaussian channels in Example 1.48 came out independent. They are jointly Gaussian, and the sum-and-difference example of the next subsection shows that they are uncorrelated.

Watch out: Gaussian marginals do not make a Gaussian pair

The shortcut "uncorrelated ⇒ independent" needs the variables to be *jointly* Gaussian, not just each Gaussian. Let $X \sim \mathcal{N}(0, 1)$ and let $S = \pm 1$ be an independent fair sign. Then $Y = SX$ is also $\mathcal{N}(0, 1)$ (flipping the sign of a symmetric variable changes nothing), and $\mathrm{Cov}(X, Y) = \mathbb{E}(S)\,\mathbb{E}(X^2) = 0$. But $|Y| = |X|$ always, so the two are strongly dependent. The pair lives on the two diagonals $y = \pm x$, which no bivariate Gaussian does.

1.13.4 The covariance matrix and how linear maps move it

With many variables we collect them into a column vector $\boldsymbol{X} = (X_1, \ldots, X_k)^{\mathsf{T}}$ with mean vector $\boldsymbol{\mu} = (\mathbb{E}\,X_1, \ldots, \mathbb{E}\,X_k)^{\mathsf{T}}$. All the variances and covariances go into one matrix, the **covariance matrix** ("variance–covariance matrix" in [Wasserman p. 53], "error matrix" in [Cowan p. 18]):

$$\mathbf{C}_{ij} = \mathrm{Cov}(X_i, X_j), \qquad \mathbf{C} = \mathbb{E}\big[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^{\mathsf{T}}\big] = \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \cdots \\ \vdots & & \ddots \end{pmatrix}. \tag{1.57}$$

The outer product $(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^{\mathsf{T}}$ is a $k \times k$ matrix whose (i, j) entry is $(X_i - \mu_i)(X_j - \mu_j)$, and the expectation is taken entry by entry. The matrix is symmetric, its diagonal holds the variances, and it is positive semi-definite: for any fixed vector $\boldsymbol{a}$,

$$\boldsymbol{a}^{\mathsf{T}}\mathbf{C}\,\boldsymbol{a} \overset{(1)}{=} \sum_{i,j} a_i a_j\,\mathrm{Cov}(X_i, X_j) \overset{(2)}{=} \mathrm{Var}\big(\boldsymbol{a}^{\mathsf{T}}\boldsymbol{X}\big) \overset{(3)}{\geq} 0.$$

(1) Write out the quadratic form.

(2) Bilinearity (1.55) with $b_j = a_j$ and $Y_j = X_j$.
(3) A variance is never negative.

So all eigenvalues of $\mathbf{C}$ are ≥ 0. A zero eigenvalue means some combination $\boldsymbol{a}^{\mathsf{T}}\boldsymbol{X}$ has zero variance, i.e. is a constant: the variables obey an exact linear constraint (as the multinomial counts below do). For variables with nonzero variance, dividing entry (i, j) by $\sigma_i\sigma_j$ gives the correlation matrix. Each off-diagonal entry describes the linear association of a pair on the same dimensionless scale, and each diagonal entry is one. For jointly Gaussian variables the mean vector and covariance matrix determine the entire joint distribution.

Now let $\boldsymbol{Y} = A\boldsymbol{X} + \boldsymbol{b}$ for a fixed $m \times k$ matrix A and vector $\boldsymbol{b}$. Linearity gives $\mathbb{E}\,\boldsymbol{Y} = A\boldsymbol{\mu} + \boldsymbol{b}$, and [Wasserman Lemma 3.21, p. 54]

$$\mathbf{C}_{\boldsymbol{Y}} \overset{(1)}{=} \mathbb{E}\big[(A\boldsymbol{X} - A\boldsymbol{\mu})(A\boldsymbol{X} - A\boldsymbol{\mu})^{\mathsf{T}}\big] \overset{(2)}{=} \mathbb{E}\big[A(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^{\mathsf{T}}A^{\mathsf{T}}\big] \overset{(3)}{=} A\,\mathbf{C}\,A^{\mathsf{T}}. \tag{1.58}$$

(1) Definition (1.57) for $\boldsymbol{Y}$; the shift $\boldsymbol{b}$ cancels in $\boldsymbol{Y} - \mathbb{E}\,\boldsymbol{Y}$.
(2) $(A\boldsymbol{v})^{\mathsf{T}} = \boldsymbol{v}^{\mathsf{T}}A^{\mathsf{T}}$.
(3) A is constant and comes out of the entry-by-entry expectation on both sides (linearity).

In components, $(\mathbf{C}_{\boldsymbol{Y}})_{ij} = \sum_{k,l} A_{ik}A_{jl}\,\mathrm{Cov}(X_k, X_l)$ [Cowan (1.62)], which is just (1.55). For a single combination, $A = \boldsymbol{a}^{\mathsf{T}}$ is one row and $\mathrm{Var}(\boldsymbol{a}^{\mathsf{T}}\boldsymbol{X}) = \boldsymbol{a}^{\mathsf{T}}\mathbf{C}\boldsymbol{a}$, the quadratic form above. The covariance matrix transforms like a rank-2 tensor, $C'_{ij} = A_{ik}A_{jl}C_{kl}$; the inertia tensor and the stress tensor transform the same way.

Key idea: Linear maps move covariance by sandwiching

If $\boldsymbol{Y} = A\boldsymbol{X} + \boldsymbol{b}$ then $\mathbb{E}\,\boldsymbol{Y} = A\ \mathbb{E}\,\boldsymbol{X} + \boldsymbol{b}$ and $\mathbf{C}_{\boldsymbol{Y}} = A\,\mathbf{C}_{\boldsymbol{X}}A^{\mathsf{T}}$, exactly, for any distribution.

Example: sum and difference of two channels. Take $\mathbf{C} = \left(\begin{smallmatrix}\sigma_1^2 & c\\ c & \sigma_2^2\end{smallmatrix}\right)$ and $A = \left(\begin{smallmatrix}1 & 1\\ 1 & -1\end{smallmatrix}\right)$, so $Y_1 = X_1 + X_2$ and $Y_2 = X_1 - X_2$. Then

$$A\,\mathbf{C}\,A^{\mathsf{T}} \overset{(1)}{=} \begin{pmatrix}\sigma_1^2 + c & c + \sigma_2^2\\ \sigma_1^2 - c & c - \sigma_2^2\end{pmatrix}\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix} \overset{(2)}{=} \begin{pmatrix}\sigma_1^2 + \sigma_2^2 + 2c & \sigma_1^2 - \sigma_2^2\\ \sigma_1^2 - \sigma_2^2 & \sigma_1^2 + \sigma_2^2 - 2c\end{pmatrix}.$$

(1) First product $A\mathbf{C}$, row by column; $A^{\mathsf{T}} = A$ here.
(2) Second product; e.g. the $(1,2)$ entry is $(\sigma_1^2 + c) - (c + \sigma_2^2)$.

The diagonal reproduces $\mathrm{Var}(X_1 \pm X_2) = \sigma_1^2 + \sigma_2^2 \pm 2c$. The new off-diagonal entry, $\mathrm{Cov}(X_1 + X_2, X_1 - X_2) = \sigma_1^2 - \sigma_2^2$, is zero whenever the two channels are equally noisy, whatever their correlation c. With $\mathbf{C} = \sigma^2\mathbb{1}$ we get $A\mathbf{C}A^{\mathsf{T}} = 2\sigma^2\mathbb{1}$, and since the pair is jointly Gaussian that is the independence of Example 1.48. For the numerical case $\sigma_1^2 = 4$, $\sigma_2^2 = 1$, $c = 1.8$ the prediction is $\left(\begin{smallmatrix}8.6 & 3\\ 3 & 1.4\end{smallmatrix}\right)$, and the sample covariance of 2×10^5 simulated pairs is $\left(\begin{smallmatrix}8.595 & 2.999\\ 2.999 & 1.401\end{smallmatrix}\right)$.

Example: the multinomial covariance. Throw n events into k histogram bins with probabilities $p_1, \ldots, p_k$ (Section 1.10), and let X_i be the count in bin i. Each X_i is binomial, $\mathrm{Var}(X_i) = np_i(1-p_i)$. To get the covariance of two bins, merge them: $X_i + X_j$ counts the events that fell in either, so it is Binomial$(n, p_i + p_j)$ [Wasserman p. 54]. Then

$$\begin{aligned} 2\,\mathrm{Cov}(X_i, X_j) &\overset{(1)}{=} \mathrm{Var}(X_i + X_j) - \mathrm{Var}(X_i) - \mathrm{Var}(X_j)\\ &\overset{(2)}{=} n(p_i + p_j)(1 - p_i - p_j) - np_i(1 - p_i) - np_j(1 - p_j) \overset{(3)}{=} -2np_ip_j, \end{aligned}$$

(1) (1.56) for $X_i + X_j$, solved for the covariance.

(2) Binomial variances of the merged bin and of each bin.
(3) Expand: the linear terms cancel, and $-(p_i+p_j)^2+p_i^2+p_j^2=-2p_ip_j$.

so $\mathrm{Cov}(X_i,X_j)=-np_ip_j$ for $i\neq j$. Histogram bins with a fixed total are anticorrelated: an extra event in one bin is an event missing from the others. The covariance matrix has a zero eigenvalue along $\boldsymbol{a}=(1,\dots,1)$, because $\sum_i X_i=n$ is exactly constant. (If the total itself is a Poisson count, as for a collider run of fixed luminosity, the bins become independent Poisson variables; see Chapter 2.)

1.13.5 Diagonalising: the principal axes of the error ellipse

Correlated variables are awkward: we cannot read off the uncertainty of one without the other. Can a linear change of variables always make them uncorrelated? Yes, and it is a rotation [Cowan §1.7]. Because $\mathbf{C}$ is real and symmetric, it has k orthonormal eigenvectors $\boldsymbol{r}_1,\dots,\boldsymbol{r}_k$, $\mathbf{C}\boldsymbol{r}_j=\lambda_j\boldsymbol{r}_j$, with $\lambda_j\geq 0$ by positive semi-definiteness. Put them as the columns of an orthogonal matrix R and take $\boldsymbol{W}=R^{\mathsf{T}}\boldsymbol{X}$, so $W_i=\boldsymbol{r}_i^{\mathsf{T}}\boldsymbol{X}$ is the component of $\boldsymbol{X}$ along $\boldsymbol{r}_i$. By (1.58) with $A=R^{\mathsf{T}}$,

$$(\mathbf{C}_W)_{ij}\overset{(1)}{=}\left(R^{\mathsf{T}}\mathbf{C}R\right)_{ij}\overset{(2)}{=}\boldsymbol{r}_i^{\mathsf{T}}\mathbf{C}\,\boldsymbol{r}_j\overset{(3)}{=}\lambda_j\,\boldsymbol{r}_i^{\mathsf{T}}\boldsymbol{r}_j\overset{(4)}{=}\lambda_j\,\delta_{ij}.$$

(1) The sandwich rule with $A=R^{\mathsf{T}}$.
(2) Row i of R^{T} is $\boldsymbol{r}_i^{\mathsf{T}}$ and column j of R is $\boldsymbol{r}_j$.
(3) Eigenvalue equation.
(4) Orthonormality of the eigenvectors.

The new variables are uncorrelated and their variances are the eigenvalues [Cowan (1.65)]. Geometrically, the contours of a Gaussian density with covariance $\mathbf{C}$ are ellipses (ellipsoids) whose axes point along the eigenvectors, with semi-axes $\sqrt{\lambda_j}$ at "1σ". The rotation simply turns the coordinate axes onto the principal axes of this **error ellipse**, exactly like rotating onto the principal axes of an inertia tensor or onto the normal modes of coupled oscillators.

In two dimensions everything is explicit. Write $\mathbf{C}=\begin{pmatrix}\sigma_1^2 & c\\ c & \sigma_2^2\end{pmatrix}$ with $c=\rho\sigma_1\sigma_2$. The eigenvalues solve $\det(\mathbf{C}-\lambda\mathbb{1})=\lambda^2-(\sigma_1^2+\sigma_2^2)\lambda+\sigma_1^2\sigma_2^2-c^2=0$, so

$$\lambda_\pm=\tfrac{1}{2}\left[\sigma_1^2+\sigma_2^2\pm\sqrt{(\sigma_1^2+\sigma_2^2)^2-4(1-\rho^2)\sigma_1^2\sigma_2^2}\right] \tag{1.59}$$

[Cowan (1.69)]. For the rotation angle, rotate by θ, $W_1=\cos\theta\,X_1+\sin\theta\,X_2$ and $W_2=-\sin\theta\,X_1+\cos\theta\,X_2$, and demand zero covariance:

$$\begin{aligned}\mathrm{Cov}(W_1,W_2)&\overset{(1)}{=}-\sin\theta\cos\theta\,\sigma_1^2+(\cos^2\theta-\sin^2\theta)\,c+\sin\theta\cos\theta\,\sigma_2^2\\&\overset{(2)}{=}\tfrac{1}{2}\sin 2\theta\,(\sigma_2^2-\sigma_1^2)+c\cos 2\theta,\end{aligned}$$

(1) Bilinearity (1.55).
(2) Double-angle formulas.

which vanishes for $\tan 2\theta=2c/(\sigma_1^2-\sigma_2^2)$ [Cowan (1.71)]. For $\sigma_1^2=4$, $\sigma_2^2=1$, $c=1.8$ (so $\rho=0.9$): $\tan 2\theta=1.2$, $\theta=25.1^\circ$, and (1.59) gives $\lambda=(5\pm\sqrt{21.96})/2=4.843$ and 0.157. The simulated rotated variables have variances 4.840 and 0.157 and covariance -0.001 (Figure 1.45, right). The long axis carries almost all the variance: nearly all the scatter of the two channels is in one combination.

The same algebra run backwards generates correlated variables. If $\boldsymbol{Z}$ has independent standard normal components, $\mathbf{C}_{\boldsymbol{Z}}=\mathbb{1}$, and $\boldsymbol{X}=L\boldsymbol{Z}$ has covariance $L\mathbb{1}L^{\mathsf{T}}=LL^{\mathsf{T}}$. So any

factorisation $\mathbf{C} = LL^{\mathsf{T}}$ gives a recipe. The standard one is the Cholesky factor, a lower triangular L; for 2×2, $L = \begin{pmatrix} \sigma_1 & 0 \\ \rho\sigma_2 & \sigma_2\sqrt{1-\rho^2} \end{pmatrix}$, which one checks by multiplying out. Here $L = \left(\begin{smallmatrix} 2.000 & 0 \\ 0.900 & 0.436 \end{smallmatrix}\right)$, and the script uses exactly this:

Listing 1.17: Correlated Gaussians, the sandwich rule and the eigenbasis (`code/ch01/18_covariance_transform.py`)

```
21 C = np.array([[4.0, 1.8], [1.8, 1.0]])
22 L = np.linalg.cholesky(C)
23 N = 200_000
24 Z = rng.normal(size=(2, N))
25 X = L @ Z
26 C_hat = np.cov(X)
27
28 A = np.array([[1.0, 1.0], [1.0, -1.0]])
29 Y = A @ X
30 CY_pred = A @ C @ A.T
31 CY_hat = np.cov(Y)
32
33 lam, R = np.linalg.eigh(C)          # columns of R are unit eigenvectors, lam ascending
34 W = R.T @ X                         # coordinates along the eigenvectors
35 CW_hat = np.cov(W)
```

Lines 22–25 build L and generate $\boldsymbol{X} = L\boldsymbol{Z}$ for 2×10^5 pairs; their sample covariance (line 26) is $\left(\begin{smallmatrix} 3.998 & 1.798 \\ 1.798 & 0.999 \end{smallmatrix}\right)$, the target $\mathbf{C}$. Lines 28–31 apply the sum-and-difference matrix and compare $A\mathbf{C}A^{\mathsf{T}}$ with the sample covariance of $\boldsymbol{Y}$. Lines 33–35 diagonalise: `np.linalg.eigh` returns the eigenvalues in ascending order and the unit eigenvectors as the columns of `R`, and $\boldsymbol{W} = R^{\mathsf{T}}\boldsymbol{X}$ is uncorrelated.

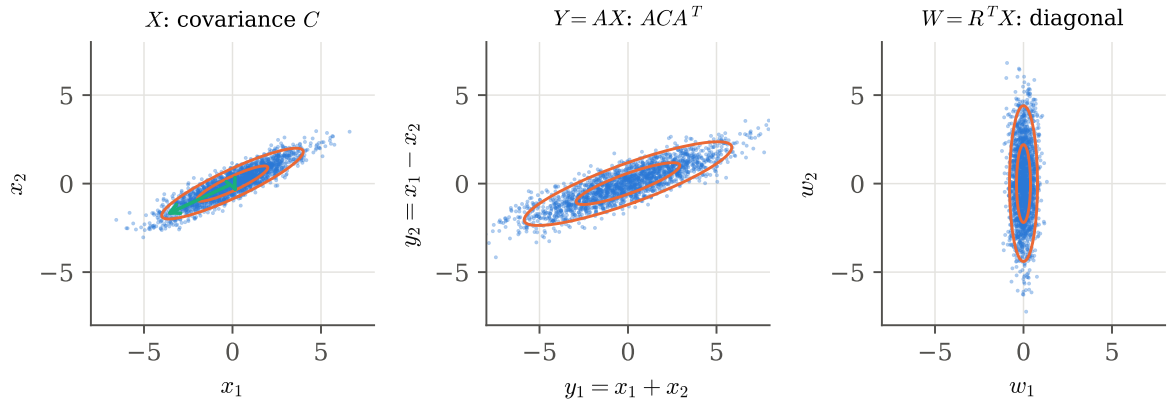


Figure 1.45: One cloud of correlated pairs seen in three coordinate systems, with the predicted 1σ and 2σ ellipses. Left: the original channels, with the two eigenvectors of $\mathbf{C}$ drawn as green arrows. Middle: sum and difference; the ellipse is the one predicted by $A\mathbf{C}A^{\mathsf{T}}$. Right: components along the eigenvectors; the ellipse is aligned with the axes, so the new variables are uncorrelated.

1.13.6 A first look at error propagation

Most derived quantities are not linear in the measured ones: a mass from two momenta, a ratio of two rates, a distance from a flux. If the input uncertainties are small, we can linearise [Cowan §1.6]. Expand $y(\boldsymbol{x})$ to first order about the means, $y(\boldsymbol{x}) \approx y(\boldsymbol{\mu}) + \sum_i A_i(x_i - \mu_i)$ with $A_i = \partial y/\partial x_i$ at $\boldsymbol{\mu}$. The constant does not affect the spread, and the rest is linear, so the sandwich

rule applies to the Jacobian matrix $A_{ki} = \partial y_k / \partial x_i$:

$$\mathbb{E}\, y_k \approx y_k(\boldsymbol{\mu}), \qquad \operatorname{Cov}(y_k, y_l) \approx \sum_{i,j} \frac{\partial y_k}{\partial x_i}\frac{\partial y_l}{\partial x_j}\, \mathbf{C}_{ij}, \qquad \mathbf{C}_y \approx A\, \mathbf{C}\, A^{\mathsf{T}} \tag{1.60}$$

[Cowan (1.53)–(1.55)]. For a linear map this is exact (1.58); otherwise it is the first term of an expansion.

Example: products add relative errors in quadrature. For $y = x_1 x_2$ the Jacobian is $A = (\mu_2,\ \mu_1)$, so

$$\sigma_y^2 \overset{(1)}{=} \mu_2^2\sigma_1^2 + \mu_1^2\sigma_2^2 + 2\mu_1\mu_2\, \mathbf{C}_{12}, \qquad \frac{\sigma_y^2}{y^2} \overset{(2)}{=} \frac{\sigma_1^2}{\mu_1^2} + \frac{\sigma_2^2}{\mu_2^2} + 2\frac{\mathbf{C}_{12}}{\mu_1\mu_2}.$$

(1) (1.60) with $\partial y/\partial x_1 = x_2$, $\partial y/\partial x_2 = x_1$ at the means.
(2) Divide by $y^2 \approx \mu_1^2\mu_2^2$.

For uncorrelated inputs, absolute errors add in quadrature for a sum and relative errors add in quadrature for a product [Cowan (1.59)–(1.60)]. The approximation fails when y curves appreciably over a range of one standard deviation; the classic case is $1/x$ with σ_x not small compared with μ_x. Then one simulates (Chapter 6). Error propagation, and its limits, is developed properly in Chapter 3.

How research uses it. In a real analysis the covariance matrix is rarely computed on paper. The band powers of a CMB or galaxy power spectrum measured on a masked, noisy sky (Chapters 7 and 8) are correlated in ways no formula captures, so one simulates N_s mock data sets and uses the sample covariance, the matrix version of S_n^2 with its $1/(N_s-1)$. The χ^2 likelihood needs the *inverse*, and the inverse of a noisy estimate is biased; it must be rescaled by $(N_s - p - 2)/(N_s - 1)$ for p data points [6]. The Fisher matrix of Chapter 10 is the inverse of a forecast parameter covariance, and its eigenvectors are the principal axes above: the best- and worst-constrained combinations of parameters. The simulations here *verified* $A\mathbf{C}A^{\mathsf{T}}$, a known result; in those analyses the same simulate-and-average machinery is the only way to obtain $\mathbf{C}$ at all.

1.14 Moments, generating functions and characteristic functions

A physicist has two standard ways of packing a whole function into something easier to handle. One is a list of moments: the total charge, the dipole moment and the quadrupole moment of a charge distribution describe it from far away, better and better as more terms are kept. The other is a transform: a Fourier or Laplace transform turns a messy operation on functions into a simple one on their transforms. Probability uses both, and statistical mechanics has used them all along. The partition function $Z(\beta) = \int g(E)\, e^{-\beta E}\, \mathrm{d}E$ is the Laplace transform of the density of states $g(E)$, and the mean energy and its fluctuations, $\langle E\rangle = -\partial_\beta \ln Z$ and $\langle E^2\rangle - \langle E\rangle^2 = \partial_\beta^2 \ln Z = k_B T^2 C_V$, are read off from its derivatives.

Higher moments describe the shape of a distribution. Two transforms, the moment generating function and the characteristic function, produce all the moments at once. Their more important use is a different one: sums. In Section 1.11 the density of a sum of independent variables was a convolution (1.39), an integral that gets worse with every term added. After a Fourier transform, convolution becomes multiplication. That single fact is what makes the central limit theorem of Chapter 2 provable in a few lines. Three questions organise the topic.

1. Which numbers beyond the mean and the variance describe the shape of a distribution, and do they determine it?
2. How do we get the distribution of a sum of independent variables without doing convolution integrals?
3. What do we do when the moments, or the transform itself, do not exist?

1.14.1 Moments and the shape of a distribution

The k-th **moment** of X is $\mathbb{E}(X^k)$ and the k-th **central moment** is $\mathbb{E}\big[(X-\mu)^k\big]$, assuming $\mathbb{E}\,|X|^k < \infty$ [Wasserman p. 49–50] ([Cowan (1.41)–(1.42)] writes them μ'_n and μ_n). The first moment is the mean and the second central moment the variance. The standardised third and fourth central moments describe shape:

$$\gamma_1 = \frac{\mathbb{E}(X-\mu)^3}{\sigma^3} \ \text{(skewness)}, \qquad \gamma_2 = \frac{\mathbb{E}(X-\mu)^4}{\sigma^4} - 3 \ \text{(excess kurtosis)}. \tag{1.61}$$

Both are pure numbers. The skewness is positive when a long tail extends to the right (the exponential has $\gamma_1 = 2$), zero for a symmetric distribution. The -3 is chosen so that a Gaussian has $\gamma_2 = 0$; distributions with heavier tails than the Gaussian have $\gamma_2 > 0$. In the multipole language the mean is the centre of charge, and the variance is the moment of inertia about it.

A higher moment existing guarantees the lower ones [Wasserman Thm. 3.10, p. 49]. For $j < k$,

$$\begin{aligned}\mathbb{E}\,|X|^j &\overset{(1)}{=} \int_{|x|\le 1} |x|^j f(x)\,\mathrm{d}x + \int_{|x|>1} |x|^j f(x)\,\mathrm{d}x \\ &\overset{(2)}{\le} \int_{|x|\le 1} f(x)\,\mathrm{d}x + \int_{|x|>1} |x|^k f(x)\,\mathrm{d}x \overset{(3)}{\le} 1 + \mathbb{E}\,|X|^k .\end{aligned}$$

(1) Split the range at $|x| = 1$.
(2) On $|x| \le 1$, $|x|^j \le 1$. On $|x| > 1$, $|x|^j \le |x|^k$ because $j < k$.
(3) The first integral is at most the total probability 1; the second is at most the full integral of $|x|^k f$.

The converse fails, and tails decide: the Cauchy density has no moments at all (Section 1.12), and a density falling like $1/|x|^4$ has a mean and a variance but no third moment.

Do the moments, if they all exist, determine the distribution? Not always. Here are two different densities on $x > 0$ [CasellaBerger Ex. 2.3.10]:

$$f_1(x) = \frac{1}{\sqrt{2\pi}\,x}\, e^{-(\ln x)^2/2}, \qquad f_2(x) = f_1(x)\big[1 + \sin(2\pi \ln x)\big],$$

with identical moments of every order. (f_1 is the lognormal: $\ln X$ is standard normal.) With $u = \ln x$, so that $f_1(x)\,\mathrm{d}x = \varphi(u)\,\mathrm{d}u$ with φ the standard normal density, the extra piece of the r-th moment of f_2 is

$$\begin{aligned}\int_0^\infty x^r f_1(x)\sin(2\pi\ln x)\,\mathrm{d}x &\overset{(1)}{=} \int_{-\infty}^{\infty} e^{ru}\varphi(u)\sin(2\pi u)\,\mathrm{d}u \overset{(2)}{=} e^{r^2/2}\int_{-\infty}^{\infty}\varphi(u-r)\sin(2\pi u)\,\mathrm{d}u \\ &\overset{(3)}{=} e^{r^2/2}\int_{-\infty}^{\infty}\varphi(y)\sin(2\pi y)\,\mathrm{d}y \overset{(4)}{=} 0.\end{aligned}$$

(1) Substitute $x = e^u$; $x^r = e^{ru}$.
(2) Complete the square: $ru - u^2/2 = r^2/2 - (u-r)^2/2$.

(3) Substitute $y = u - r$. For integer r, $\sin(2\pi(y + r)) = \sin(2\pi y)$.
(4) φ is even and sin is odd, so the integrand is odd.

So f_1 and f_2 share all moments, $\mathbb{E}\,X^r = e^{r^2/2}$ for both, though they are visibly different densities. The problem is again the tails: the moments grow so fast that they fail to pin down the function. When the support is bounded, or when the moment generating function below exists in an interval around $t = 0$, the moments do determine the distribution [CasellaBerger Thm. 2.3.11].

1.14.2 The moment generating function

The **moment generating function** (mgf) of X is [Wasserman Def. 3.29, p. 56]

$$\psi_X(t) = \mathbb{E}\big(e^{tX}\big) = \int e^{tx} f(x)\,\mathrm{d}x, \tag{1.62}$$

for the real t where the expectation is finite. It is the Laplace transform of the density (with $t \to -t$), and we assume it is finite on some interval around $t = 0$. The name comes from expanding the exponential: $e^{tX} = \sum_k t^k X^k/k!$, so

$$\psi_X(t) = \sum_{k=0}^{\infty} \frac{t^k}{k!}\,\mathbb{E}(X^k), \qquad \text{hence} \qquad \psi_X^{(k)}(0) = \mathbb{E}(X^k). \tag{1.63}$$

The moments are the Taylor coefficients of ψ at 0 (times $k!$). Moving $\mathbb{E}$ through the sum, or through $\mathrm{d}/\mathrm{d}t$ as [Wasserman p. 56] does, is legitimate precisely because ψ is finite near 0.

Example: the exponential. For $X \sim \mathrm{Exp}(1)$, $f(x) = e^{-x}$ on $x > 0$ [Wasserman Ex. 3.30, p. 57]:

$$\psi_X(t) \overset{(1)}{=} \int_0^{\infty} e^{tx} e^{-x}\,\mathrm{d}x \overset{(2)}{=} \int_0^{\infty} e^{-(1-t)x}\,\mathrm{d}x \overset{(3)}{=} \frac{1}{1-t} \overset{(4)}{=} \sum_{k=0}^{\infty} t^k \qquad (t < 1).$$

(1) Definition (1.62).
(2) Combine the exponents.
(3) The integral converges only for $t < 1$; for $t \geq 1$ it diverges.
(4) Geometric series, valid for $|t| < 1$.

Matching with (1.63), $\mathbb{E}(X^k)/k! = 1$, so $\mathbb{E}(X^k) = k!$: mean 1, $\mathbb{E}\,X^2 = 2$, variance $2 - 1 = 1$, all without an integral beyond the first. We can run the same computation on data. Replace ψ by the empirical mgf, the average of e^{tX_j} over simulated draws, and differentiate it numerically at 0:

Listing 1.18: Moments from the empirical mgf (`code/ch01/19_characteristic_function.py`)

```
23 X = rng.exponential(1.0, size=1_000_000)
24 h = 0.01
25 psi = lambda t: np.mean(np.exp(t * X))
26 m1 = (psi(h) - psi(-h)) / (2 * h)
27 m2 = (psi(h) - 2 * psi(0.0) + psi(-h)) / h**2
```

Line 25 defines $\hat{\psi}(t)$ as a sample mean over 10^6 draws; lines 26–27 are the central finite differences for $\psi'(0)$ and $\psi''(0)$ with step $h = 0.01$. They give 0.9993 and 1.9993 (exact 1 and 2).

Linear changes and sums. Two properties make the mgf useful [Wasserman Lemma 3.31, p. 57]. For $Y = aX + b$, $\psi_Y(t) = \mathbb{E}\big(e^{t(aX+b)}\big) = e^{bt}\psi_X(at)$. For a sum of independent variables,

$$\psi_{X_1+\cdots+X_n}(t) \overset{(1)}{=} \mathbb{E}\big(e^{tX_1}e^{tX_2}\cdots e^{tX_n}\big) \overset{(2)}{=} \mathbb{E}\big(e^{tX_1}\big)\cdots\mathbb{E}\big(e^{tX_n}\big) = \prod_{i=1}^{n} \psi_{X_i}(t). \tag{1.64}$$

(1) The exponential of a sum is the product of exponentials.
(2) Independence: the mean of a product of functions of independent variables is the product of the means (Section 1.12).

Combined with uniqueness (if two mgfs agree on an interval around 0, the distributions are equal, [Wasserman Thm. 3.33, p. 57]), this identifies the distribution of a sum by recognising its mgf.

Example: binomial and Poisson sums. A Bernoulli trial has $\psi(t) = (1-p) + pe^t$. A binomial count is a sum of n independent trials, so $\psi(t) = (pe^t + q)^n$ with $q = 1-p$ [Wasserman Ex. 3.32, p. 57]. Adding two independent binomials with the same p gives $(pe^t+q)^{n_1}(pe^t+q)^{n_2} = (pe^t+q)^{n_1+n_2}$, the mgf of Binomial$(n_1+n_2, p)$: pooling two runs of trials is one longer run [Wasserman Ex. 3.34, p. 57]. For a Poisson count,

$$\psi(t) \overset{(1)}{=} \sum_{k=0}^{\infty} e^{tk}\,\frac{e^{-\lambda}\lambda^k}{k!} \overset{(2)}{=} e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda e^t)^k}{k!} \overset{(3)}{=} e^{-\lambda}e^{\lambda e^t} = e^{\lambda(e^t-1)}.$$

(1) Lazy statistician with the Poisson pmf.
(2) Collect $e^{tk}\lambda^k = (\lambda e^t)^k$.
(3) Exponential series $\sum_k a^k/k! = e^a$.

Then independent $Y_1 \sim \text{Poisson}(\lambda_1)$ and $Y_2 \sim \text{Poisson}(\lambda_2)$ have $\psi_{Y_1+Y_2} = e^{\lambda_1(e^t-1)}e^{\lambda_2(e^t-1)} = e^{(\lambda_1+\lambda_2)(e^t-1)}$, so $Y_1 + Y_2 \sim \text{Poisson}(\lambda_1+\lambda_2)$ [Wasserman Ex. 3.35, p. 58].[6] Signal events and background events counted in the same window form one Poisson count with the summed rate, the starting point of every counting experiment in Chapter 4.

Example: the Gaussian, by completing the square. For $X \sim \mathcal{N}(\mu, \sigma^2)$ write $x = \mu + \sigma z$:

$$\psi_X(t) \overset{(1)}{=} e^{\mu t}\int \frac{e^{-z^2/2}}{\sqrt{2\pi}}\,e^{\sigma t z}\,\mathrm{d}z \overset{(2)}{=} e^{\mu t}\,e^{\sigma^2t^2/2}\int\frac{e^{-(z-\sigma t)^2/2}}{\sqrt{2\pi}}\,\mathrm{d}z \overset{(3)}{=} e^{\mu t+\sigma^2t^2/2}.$$

(1) $e^{tx} = e^{\mu t}e^{\sigma t z}$, and $f(x)\,\mathrm{d}x = \varphi(z)\,\mathrm{d}z$.
(2) Complete the square: $-z^2/2 + \sigma t z = -(z-\sigma t)^2/2 + \sigma^2t^2/2$.
(3) The remaining integral is a shifted standard normal density, which integrates to 1.

A sum of independent Gaussians therefore has mgf $\exp\left[(\sum_i \mu_i)t + (\sum_i \sigma_i^2)t^2/2\right]$: it is Gaussian, with the means added and the variances added. The table of [Wasserman p. 58] also lists Gamma(α,β), $\psi(t) = (1-\beta t)^{-\alpha}$ for $t < 1/\beta$. Since an exponential with mean β is Gamma$(1,\beta)$, the sum of n independent exponential lifetimes has mgf $(1-\beta t)^{-n}$ and is Gamma(n,β) [Wasserman Exercise 3.24, p. 61]. The named distributions and their mgfs are the subject of Chapter 2.

Cumulants. Because sums multiply mgfs, their logarithms add. The **cumulant generating function** $K(t) = \ln\psi(t)$ has Taylor coefficients $K(t) = \sum_n \kappa_n t^n/n!$ called **cumulants** [CasellaBerger §2.6.2]. The first two are

$$\kappa_1 = K'(0) \overset{(1)}{=} \frac{\psi'(0)}{\psi(0)} = \mu, \qquad \kappa_2 = K''(0) \overset{(2)}{=} \frac{\psi''(0)}{\psi(0)} - \left(\frac{\psi'(0)}{\psi(0)}\right)^2 = \sigma^2,$$

(1) Derivative of a logarithm; $\psi(0) = \mathbb{E}(1) = 1$ and $\psi'(0) = \mu$.

[6] The text writes the sum as "$Y = Y_1 + Y + 2$"; it should read $Y = Y_1 + Y_2$.

(2) Quotient rule, then $\psi''(0) = \mathbb{E}\,X^2$ and the shortcut (1.46).

and $\kappa_3 = \mathbb{E}(X-\mu)^3$. For independent X, Y every cumulant adds, $\kappa_n(X+Y) = \kappa_n(X) + \kappa_n(Y)$. For the Gaussian $K(t) = \mu t + \sigma^2 t^2/2$ exactly, so every cumulant beyond the second is zero. Non-zero higher cumulants therefore measure non-Gaussianity: the skewness is $\kappa_3/\kappa_2^{3/2}$. The primordial non-Gaussianity parameter f_{NL} of cosmology is a third cumulant of this kind, and the topological summaries of Chapter 12 are one way to detect such departures from Gaussianity.

Return to the partition function $Z(\beta) = \int g(E)\, e^{-\beta E}\, \mathrm{d}E$ of the opening. The mgf of the energy in the canonical ensemble is $\mathbb{E}\, e^{tE} = Z(\beta - t)/Z(\beta)$, so $K(t) = \ln Z(\beta - t) - \ln Z(\beta)$ and $\kappa_1 = -\partial_\beta \ln Z$, $\kappa_2 = \partial_\beta^2 \ln Z$: the free energy is a cumulant generating function.

1.14.3 The characteristic function

The mgf has a weakness: it may not exist. For the Cauchy density $e^{tx}/(1+x^2)$ grows without bound in one tail for every $t \neq 0$, so $\psi(t) = \infty$. The lognormal f_1 above has all its moments, yet $\psi(t) = \infty$ for every $t > 0$, which is how it escapes the uniqueness theorem. The cure is to put an i in the exponent.

Working definition 1.25 (characteristic function)

The **characteristic function** of X is

$$\phi_X(t) = \mathbb{E}\big(e^{itX}\big) = \int e^{itx} f(x)\, \mathrm{d}x, \qquad t \in \mathbb{R}, \tag{1.65}$$

[the Fourier transform of the density]. It always exists, since $\left|e^{itx}\right| = 1$ gives $\left|\phi(t)\right| \le \int f = 1$, and it determines the distribution uniquely. The density is recovered by the inverse transform

$$f(x) = \frac{1}{2\pi} \int \phi_X(t)\, e^{-itx}\, \mathrm{d}t \tag{1.66}$$

whenever ϕ_X is integrable. *Operational test:* the average of $\cos(tX_j) + i\sin(tX_j)$ over draws.

([Wasserman p. 56, footnote 2] mentions it in one line; [CasellaBerger §2.6.2] states the existence and uniqueness.) In the Fourier convention used throughout, $\tilde{h}(f) = \int h(x) e^{-2\pi i f x} \mathrm{d}x$, the characteristic function is $\phi(t) = \tilde{f}(-t/2\pi)$: the same transform with the sign of the exponent flipped and the frequency measured in radians. The properties follow directly from (1.65): $\phi(0) = 1$; $\phi(-t) = \overline{\phi(t)}$; ϕ is real if and only if the distribution is symmetric about 0; $\phi_{aX+b}(t) = e^{ibt}\phi_X(at)$; the product rule (1.64) holds with it in place of t; and when the moments exist, $\phi^{(k)}(0) = i^k\, \mathbb{E}(X^k)$, so the Taylor expansion near 0 is

$$\phi_X(t) = 1 + i\mu t - \tfrac{1}{2}\,\mathbb{E}(X^2)\, t^2 + o(t^2). \tag{1.67}$$

For the Gaussian, replacing t by it in the mgf gives $\phi(t) = e^{i\mu t - \sigma^2 t^2/2}$: the Fourier transform of a Gaussian is a Gaussian, with the width inverted, as in the uncertainty relation between position and momentum wave packets.

Example: the uniform and its sums. For $U \sim \mathrm{Uniform}(-1, 1)$,

$$\phi_U(t) \overset{(1)}{=} \frac{1}{2}\int_{-1}^{1} e^{itx}\, \mathrm{d}x \overset{(2)}{=} \frac{1}{2} \cdot \frac{e^{it} - e^{-it}}{it} \overset{(3)}{=} \frac{\sin t}{t}.$$

(1) Definition (1.65) with density $\frac{1}{2}$ on $(-1, 1)$.
(2) Antiderivative $e^{itx}/(it)$.
(3) $e^{it} - e^{-it} = 2i \sin t$.

It is real because U is symmetric, and it is the diffraction pattern of a single slit, since a uniform density is a slit aperture. The sum of n independent copies has $\phi(t) = (\sin t/t)^n$. For $n = 2$, the inverse transform (1.66) of $(\sin t/t)^2$ must be the triangle density $(1 - |s|/2)/2$ on $(-2, 2)$ found by convolution in Example 1.49. Figure 1.46 checks both directions numerically. With 10^5 draws, the empirical characteristic functions of U, $U_1 + U_2$ and $U_1 + U_2 + U_3$ agree with $(\sin t/t)^n$ to within 0.0073 at every t (the expected sampling noise is about $1/\sqrt{10^5} \approx 0.003$ per point). A numerical inverse Fourier transform of $(\sin t/t)^2$ reproduces the triangle to within 0.0004.

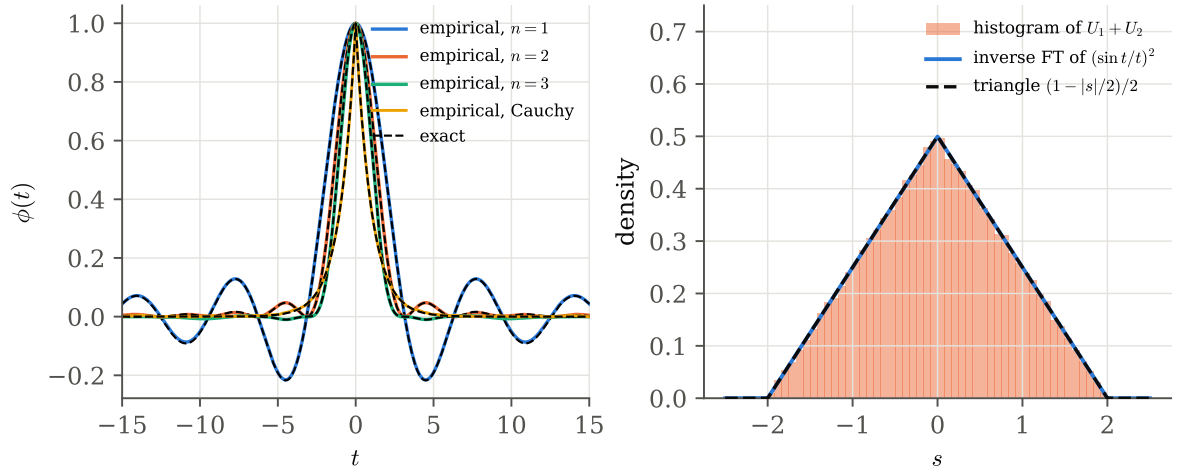


Figure 1.46: Left: characteristic functions estimated from simulated draws (coloured) against the exact ones (black dashed): $(\sin t/t)^n$ for sums of $n = 1, 2, 3$ uniforms, which narrow and flatten their side lobes as n grows, and $e^{-|t|}$ for the Cauchy, with its kink at $t = 0$. Right: the histogram of $U_1 + U_2$, the numerical inverse transform of $(\sin t/t)^2$, and the exact triangle lie on top of each other.

The picture behind this is the convolution theorem, drawn in Figure 1.47. For independent X and Y the density of $X + Y$ is the convolution $f_X * f_Y$ (1.39), and its characteristic function is the product $\phi_X \phi_Y$. Going around the square the long way, through Fourier space, replaces an integral by a multiplication.

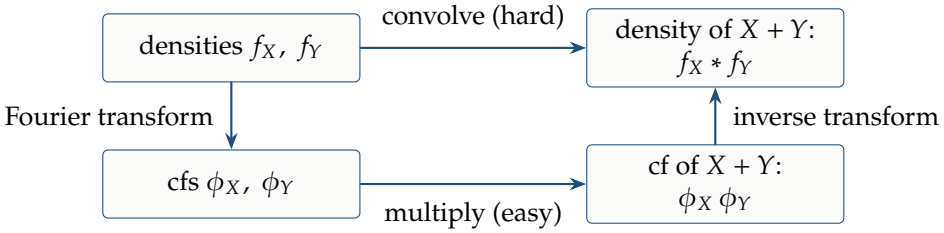


Figure 1.47: Two routes from the densities of two independent variables to the density of their sum. The top route is the convolution integral. The bottom route transforms each density, multiplies the transforms, and transforms back.

Key idea: Sums of independent variables

For independent $X_1, \dots, X_n$ the characteristic function of the sum is the product $\prod_i \phi_{X_i}(t)$, and the characteristic function determines the distribution. The same holds for the mgf wherever it exists.

Example: the Cauchy, and why its average never settles. The characteristic function of the Cauchy density is $e^{-|t|}$. The cleanest proof applies the inversion formula (1.66) to $e^{-|t|}$ and checks that the result is the Cauchy density:

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-|t|}e^{-itx}\,\mathrm{d}t \overset{(1)}{=} \frac{1}{2\pi}\left[\int_0^{\infty} e^{-t(1+ix)}\,\mathrm{d}t + \int_0^{\infty} e^{-t(1-ix)}\,\mathrm{d}t\right] \overset{(2)}{=} \frac{1}{2\pi}\left[\frac{1}{1+ix}+\frac{1}{1-ix}\right] \overset{(3)}{=} \frac{1}{\pi(1+x^2)}.$$

(1) Split at $t = 0$ and substitute $t \to -t$ in the negative half.
(2) $\int_0^{\infty} e^{-ct}\mathrm{d}t = 1/c$ for $\operatorname{Re} c > 0$.
(3) Common denominator $(1 + ix)(1 - ix) = 1 + x^2$, numerator 2.

By uniqueness, $\phi(t) = e^{-|t|}$. It has a kink at $t = 0$, so it has no derivative there, which by (1.67) is the transform-side sign that the mean does not exist. Now average n independent Cauchy draws:

$$\phi_{\bar{X}_n}(t) \overset{(1)}{=} \prod_{i=1}^{n} \phi_X\left(\frac{t}{n}\right) \overset{(2)}{=} \left(e^{-|t|/n}\right)^n = e^{-|t|}.$$

(1) $\bar{X}_n = \sum_i X_i/n$: scaling rule $\phi_{aX}(t) = \phi_X(at)$ with $a = 1/n$, and the product rule for the independent sum.
(2) The Cauchy characteristic function.

The average of n Cauchy draws has exactly the distribution of one draw, for every n. Averaging gains nothing, which is what the running averages of Figure 1.39 showed. A physicist measuring the centre of a resonance line with a pure Breit–Wigner shape by averaging event positions would find the same.

1.14.4 Setting up the central limit theorem

The characteristic function makes the most important theorem of the next chapter almost mechanical. Let $X_1, \dots, X_n$ be iid with mean 0 and variance σ^2, and standardise their sum, $Z_n = (X_1 + \cdots + X_n)/(\sigma\sqrt{n})$. Then

$$\phi_{Z_n}(t) \overset{(1)}{=} \left[\phi_X\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n \overset{(2)}{=} \left[1 - \frac{\sigma^2}{2}\,\frac{t^2}{\sigma^2 n} + o\left(\frac{1}{n}\right)\right]^n = \left[1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right]^n \xrightarrow[n\to\infty]{} e^{-t^2/2}.$$

(1) Scaling and product rules, as for the Cauchy average.
(2) Taylor expansion (1.67) with $\mu = 0$ and $\mathbb{E}\,X^2 = \sigma^2$, at the small argument $t/(\sigma\sqrt{n})$.

The last step is the limit $(1 + a/n)^n \to e^a$. The limit $e^{-t^2/2}$ is the characteristic function of $\mathcal{N}(0, 1)$, and convergence of characteristic functions implies convergence of the CDFs [CasellaBerger Thm. 2.6.1]. So the standardised sum of many iid variables with finite variance is approximately Gaussian, whatever their own distribution: the central limit theorem. Only the first two terms of (1.67) survived, which is why only the mean and the variance of the X_i matter in the limit. The Cauchy fails the hypothesis, having no finite variance and a kinked ϕ,

and is not Gaussianised. The proper statement, its conditions, the rate of convergence and its failures are the business of Chapter 2.

One more generating function is handy for counts. The **probability generating function** $G(s) = \mathbb{E}(s^X) = \sum_k \mathrm{P}(X = k)\, s^k$ of a variable taking values $0, 1, 2, \ldots$ is a power series whose coefficients are the probabilities themselves, $\mathrm{P}(X = k) = G^{(k)}(0)/k!$, and whose derivatives at $s = 1$ give the factorial moments $\mathbb{E}[X(X-1)\cdots]$ [CasellaBerger §2.6.2]. It is the mgf with $s = e^t$; for a Poisson count $G(s) = e^{\lambda(s-1)}$.

Recap

- A random variable is a map from outcomes to numbers. Its distribution is given by a CDF, and by a pmf (mass) or a pdf (probability per unit x, with units).
- Marginals average out nuisance variables; conditionals slice and renormalise; independence means the joint factorises.
- Change of variables: $f_Y\,|\mathrm{d}y| = f_X\,|\mathrm{d}x|$, summed over branches, with $|J|$ in several dimensions.
- $\mathbb{E}\, r(X) = \int r(x) f(x)\,\mathrm{d}x$. Expectation is linear always; products factorise under independence. The mean minimises the expected squared miss, the median the absolute miss.
- $\mathrm{Var}(X) = \mathbb{E}\, X^2 - \mu^2$; independent variances add; $\mathrm{Var}(\bar{X}_n) = \sigma^2/n$; S_n^2 divides by $n - 1$.
- $\mathbb{E}[\mathbb{E}(Y \mid X)] = \mathbb{E}\, Y$ and $\mathrm{Var}\, Y = \mathbb{E}\,\mathrm{Var}(Y \mid X) + \mathrm{Var}\,\mathbb{E}(Y \mid X)$.
- Covariance measures linear co-movement; $|\rho| \le 1$; zero correlation does not imply independence (except for jointly Gaussian variables).
- $\mathbf{C}_{AX+b} = A\mathbf{C}A^\mathsf{T}$; eigenvectors of $\mathbf{C}$ give uncorrelated combinations; $\boldsymbol{X} = L\boldsymbol{Z}$ with $LL^\mathsf{T} = \mathbf{C}$ generates them.
- mgf and cf generate moments; sums of independent variables multiply them; the cf always exists and leads straight to the central limit theorem.

Chapter summary

Every later statement of the form "the probability of this is that" rests on the ideas of this chapter. The first part worked with events: what has a probability, which rules the numbers obey, how conditioning shrinks the sample space, how trees chain conditions, and how reading a tree backwards turns a prediction $\mathrm{P}(E \mid H)$ into an inference $\mathrm{P}(H \mid E)$. The second part turned outcomes into numbers, gave them densities, and compressed the densities into means, variances, covariance matrices and transforms. A data analysis in physics runs along a chain,

$$\text{model} \to \text{data} \to \text{summary } S \to P(S \mid \theta) \to \text{inference :}$$

a physical model with parameters θ produces data, we compress the data into a summary S, we work out how S is distributed for each θ, and we infer θ. Probability is the language of every arrow. The fourth arrow has already been walked once. The sample mean $\bar{X}_n$ is a summary, and $\mathbb{E}\,\bar{X}_n = \mu$, $\mathrm{Var}\,\bar{X}_n = \sigma^2/n$ are the first two moments of its sampling distribution. The two tables list each concept, what it means operationally (how to compute or test it), and where it is used next.

Events and their probabilities

concept	operational meaning	used later in
sample space, event	list every outcome; an event is a subset, in code a boolean mask over simulated outcomes	every chapter
axioms of P	non-negative, total 1, additive over disjoint events; all other rules (complement, inclusion–exclusion, Boole) follow	Chapters 2 and 4
frequency vs degree of belief	long-run fraction of repetitions vs a coherent state of knowledge; same rules, different reading	Chapters 3 and 5
counting	equally likely outcomes: favourable over total, via multiplication principle and binomial coefficients	Chapter 2
conditional probability	shrink the sample space to where the evidence is true and re-measure: $\mathrm{P}(A \mid B) = \mathrm{P}(AB)/\mathrm{P}(B)$; in code, filter then count	Chapters 4 and 5
trees, total probability	multiply along a path, add across paths: $\mathrm{P}(E) = \sum_i \mathrm{P}(E \mid H_i)\,\mathrm{P}(H_i)$	Chapters 5 and 9
independence	$\mathrm{P}(AB) = \mathrm{P}(A)\,\mathrm{P}(B)$; learning B does not change $\mathrm{P}(A)$; pairwise is weaker than mutual	Chapters 2 and 6
Bayes' theorem	read the tree backwards: from $\mathrm{P}(E \mid H)$ and $\mathrm{P}(H)$ to $\mathrm{P}(H \mid E)$	Chapters 5, 9 and 10

Random variables and their summaries

concept	operational meaning	used later in
random variable	a function on outcomes; one simulated outcome gives every observable at once	every chapter
Bernoulli, binomial	one yes/no trial vs the number of successes in n independent identical trials, $\binom{n}{k}p^k(1-p)^{n-k}$	Chapters 2 and 4
CDF, quantiles	$F(x) = \mathrm{P}(X \le x)$, estimated by the empirical CDF; $F^{-1}(u)$ of a uniform u samples X	Chapters 4 and 6
pmf, pdf	mass at a point; density = probability per unit x (a normalised histogram of width $\to 0$)	Chapters 2 and 3
HDI, equal-tailed interval	shortest region of given probability vs equal tail areas	Chapter 5
joint, marginal	marginal = integrate (average) out the nuisance variables; in code, ignore the column	Chapters 5, 9 and 10
conditional density	slice at the observed value and renormalise; $f(x,y) = f(y \mid x)f(x)$	Chapters 3 and 5
independence of variables	joint density factorises; test by comparing conditional slices	Chapters 2 and 6
change of variables	$f_Y \lvert dy\rvert = f_X \lvert dx\rvert$ summed over branches; Jacobian determinant in n dimensions; check by histogramming	Chapters 2, 5 and 6
sampling prescription	say what is chosen at random: uniform instant gives the arcsine law for x, uniform x gives $(\omega/4)\lvert\sin\omega t\rvert$ for t; conservation laws give accessibility, not equal probability	Chapters 6 and 9
maximum entropy (microcanonical, canonical)	maximise $-k_B \sum p_i \ln p_i$ under the constraints: normalisation alone gives uniform on the energy shell, a fixed mean energy gives $e^{-E_i/k_BT}/Z$	Chapters 2 and 5
virial theorem, negative heat capacity	bounded motion: $2\langle K\rangle = n\langle V\rangle$ for $V \propto r^n$; gravity gives $E = -\frac{3}{2}Nk_BT$, $C < 0$, which the canonical $C = \mathrm{Var}\,E/k_BT^2 \ge 0$ forbids	when the canonical distribution applies
expectation	probability-weighted average; long-run mean of draws; $\mathbb{E}\,r(X) = \int rf$; linear always	every chapter
variance, standard deviation	mean squared deviation; independent variances add; errors add in quadrature	Chapters 2 and 3
sample mean and variance	$\bar{X}$ has variance σ^2/n; S^2 divides by $n-1$ to be unbiased	Chapters 3 and 4
conditional expectation, total variance	$\mathbb{E}\,Y = \mathbb{E}[\mathbb{E}(Y \mid X)]$; spread = mean within-group spread + spread of group means	Chapters 3 and 5
covariance, correlation	mean product of deviations; $\lvert\rho\rvert \le 1$ measures linear co-movement only; plot before trusting $\rho \approx 0$	Chapters 3 and 7
covariance matrix	$\mathbf{C}_{AX+b} = A\mathbf{C}A^{\mathsf{T}}$; eigenvectors give uncorrelated combinations; $\boldsymbol{X} = L\boldsymbol{Z}$ generates correlated Gaussians	Chapters 7, 8 and 10
error propagation	linearise: $\mathbf{C}_y \approx A\mathbf{C}A^{\mathsf{T}}$ with the Jacobian A; exact only for linear maps	Chapters 3 and 10
moments, cumulants	shape numbers; cumulants add for independent sums and vanish beyond order two for a Gaussian	Chapters 2 and 12
mgf, characteristic function	generate moments; sums of independent variables multiply them; cf always exists and gives the CLT	Chapter 2

Reading the literature

- **Main line.** [Wasserman Chs. 1–3] covers probability, random variables and expectation compactly, with many short examples and exercises; most of the examples of Chapter 1 are worked out from there in full.
- **The clearest first reading.** [Kruschke Ch. 4]: what a distribution is, mass versus density (with the density exceeding 1), means, variances, HDIs, and two-way tables, all in small steps.
- **Pictures.** [SeeingTheory Chs. 1–3]: interactive visual versions of events, expectation, variance and random variables.
- **The physicist's version.** [Cowan Ch. 1]: densities as normalised histograms, marginals as projections, conditionals as thin bands, functions of random variables, the covariance matrix, error propagation and the orthogonal transformation to uncorrelated variables.
- **Careful statements.** [CasellaBerger Chs. 1–2, 4]: transformations with full hypotheses, the minimising-distance property of the mean, non-unique moments, generating functions, conditional variance, and the correlation bounds.
- **Bayesian framing.** [Lambert Ch. 3] for probability trees and the two directions of probability, prediction and inference, which Chapter 5 develops.
- **Going further.** Characteristic functions and the continuity theorem are treated in any graduate probability text; for the physicist, [Cowan Ch. 10] applies them to sums of random variables.

Chapter 2

Probability Distributions from First Principles

Where this chapter is going

The way an experiment records data guides our choice of probability law. A detector paddle records whether it fired, a histogram counts how many events fell in each bin, and a photomultiplier can record how long it took for the tenth photon to arrive. Even when the particles arrive by the same random process, the count in a fixed interval and the waiting time for a fixed count have different distributions. For each example we state our assumptions, derive the distribution and its mean and variance, then simulate the experiment in Python to compare the observed frequencies with the prediction.

Building a distribution from its machine is the *prediction* direction of probability (Chapter 1). We fix a hypothesis, say "the source emits at rate λ", and ask which outcomes it produces and how often. Recall the book's pipeline (Chapter 0): a theory produces data, we compress the data into a statistic S, we work out how S scatters when the experiment is repeated, and from that we infer the physics, labelled by parameters θ. The distributions of counts, histograms and waiting times in this chapter supply the scatter step, $P(S \mid \theta)$, for the simplest statistics. Once we have $P(S \mid \theta)$, we can read it the other way round: hold the observed S fixed and let θ vary. Read that way it is the likelihood (Chapter 3), and Bayes' theorem turns the likelihood into a posterior (Chapter 5).

The counting distributions come first, from a single yes/no trial up to the waiting times between random events. The continuous distributions follow. With them come the reason the Gaussian appears everywhere, the distributions built from Gaussians (χ^2, Student-t, F), and the remaining common distributions.

A distribution is the fingerprint of a mechanism. Every counting distribution answers one forward question: *if the world works like this machine, how often does each outcome occur?* To use a distribution correctly, we must recognise the machine behind it. So we meet each distribution in the same four steps:

1. **The story:** which physical process are we modelling, and which assumptions (independence, constant probability, constant rate) does it satisfy?
2. **The derivation:** count the ways the outcome can happen and multiply their probabilities.
3. **The summary numbers:** mean and variance, derived rather than quoted.
4. **The check:** simulate the machine itself, not the formula, and compare the histogram with the derived law.

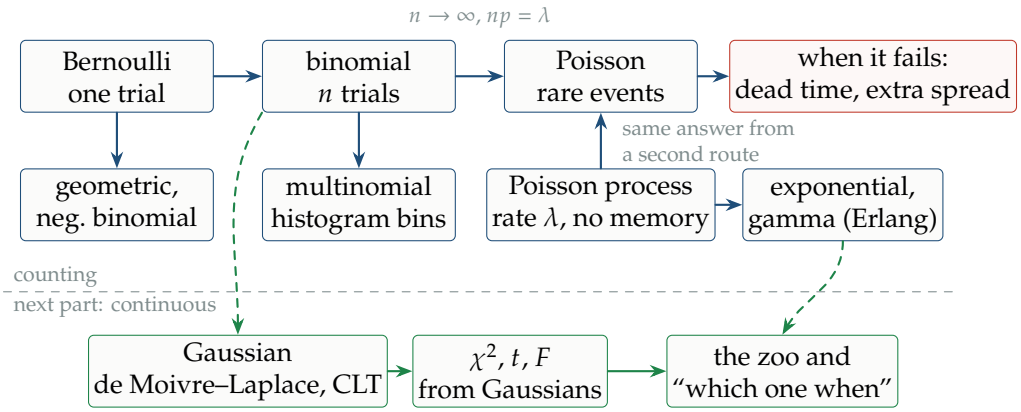


Figure 2.1: The road through the chapter. Blue boxes are the counting distributions, each from the one before it or from a second, independent route (the Poisson distribution appears both as a limit of the binomial and from three physical inputs about the arrivals). The red box collects the ways real detectors break the assumptions. Green boxes are the continuous distributions of the next part; the dashed arrows mark where the binomial and the waiting times feed into them.

2.1 One trial: the Bernoulli distribution

A cosmic-ray muon crosses a plastic scintillator paddle. Either the photomultiplier behind the paddle produces a pulse above threshold, or it does not. Nothing else can happen. The paddle has an *efficiency* ϵ: the probability that it fires when a muon really goes through. This single yes/no trial is the building block of every coincidence trigger, every tracking efficiency and every "did this event pass the cut?" in an analysis. One number describes the trial completely: the probability of a "yes", which we call p in general (for the paddle, $p = \epsilon$). Every counting distribution in this chapter is assembled from copies of this trial.

Physical picture: a two-level system measured once. A spin-$\frac{1}{2}$ particle in the state $\alpha|\uparrow\rangle + \beta|\downarrow\rangle$, measured along z, gives "up" with probability $|\alpha|^2$ and "down" with probability $|\beta|^2 = 1 - |\alpha|^2$. One measurement, two outcomes, one number. The paddle is the same situation without the Hilbert space. Its microscopic history is unknown (how many photons were made, how many reached the photocathode), and the threshold collapses all of it into a single bit: fired or not.

The question, step by step.

1. What can the outcome be? Two things. We code them as numbers: 1 ("success", fired) and 0 ("failure", did not fire).
2. How many numbers do we need to describe the trial? One: the probability p of a 1. The probability of a 0 is then forced to be $1 - p$.
3. Can we write both probabilities as one formula in the outcome x? Yes. That single formula is what makes products over many trials simple later.
4. What are its average and spread? Derive them, then check by simulation.

2.1.1 The notation of distributions

Three pieces of notation are used for the rest of the book: the symbol $\sim$, the probability mass function, and the parameter.

Working definition 2.1 (has distribution, pmf, parameter)

We write $X \sim F$ and read it "X *has distribution* F". The symbol $\sim$ does *not* mean "approximately" here. For a random variable X [a rule that assigns a number to each outcome of the experiment] that takes whole-number values, the **probability mass function** (pmf) is

$$f(x) \;=\; \mathrm{P}(X = x),$$

the probability that X comes out equal to the particular number x. A **parameter** [a fixed number, such as p, that selects one member of a family of distributions] is not random. It is written after a bar or a semicolon when we want to stress it: $f(x \mid p)$ or $f(x; p)$.

Watch out: three kinds of letter

In $\mathrm{P}(X = x) = p^x(1-p)^{1-x}$ there are three different kinds of object. X (capital) is the random variable: it has not been observed yet, and it has a distribution. x (lower case) is a possible value: an ordinary number we plug in. p is a parameter: fixed, usually unknown, and the thing statistical inference tries to learn from data. Mixing them up is the commonest source of nonsense in statistics [Wasserman p. 26]. The whole of Chapter 3 is about reading the same formula as a function of p for fixed x.

2.1.2 The two simplest distributions

Two very simple distributions mark the extreme cases of everything that follows. The first has no randomness at all; the second has no preferred outcome.

Working definition 2.2 (point mass and discrete uniform)

X has a **point mass** at a, written $X \sim \delta_a$, if $\mathrm{P}(X = a) = 1$ [the outcome is certain, there is no randomness at all]. Its cumulative distribution function (CDF) $F(x) = \mathrm{P}(X \le x)$ is a single step,

$$F(x) = \begin{cases} 0 & x < a, \\ 1 & x \ge a. \end{cases}$$

X has the **discrete uniform** distribution on $\{1, \dots, N\}$ if $f(x) = 1/N$ for $x = 1, \dots, N$ and 0 otherwise [all N outcomes equally likely, like a fair die with N faces].

The point mass is what any distribution becomes when its spread shrinks to zero: a perfectly stable voltage reference, a noise-free measurement. Its mean is a and its variance is 0. The discrete uniform is the other extreme: no outcome is preferred. Its mean and variance need two standard sums, which we recall and derive once here.

Example: mean and variance of a fair N-sided die. We need $\sum_{i=1}^{N} i = \frac{N(N+1)}{2}$ and $\sum_{i=1}^{N} i^2 = \frac{N(N+1)(2N+1)}{6}$. The first follows by writing the sum forwards and backwards and adding: each of the N columns sums to $N+1$. For the second, sum the identity $(i+1)^3 - i^3 = 3i^2 + 3i + 1$ from $i = 1$ to N. The left side telescopes (each i^3 cancels against the next term) to $(N+1)^3 - 1$. The right side is $3\sum i^2 + 3\sum i + N$. Solving, $3\sum i^2 = (N+1)^3 - 1 - \frac{3N(N+1)}{2} - N$, which simplifies to $\frac{N(N+1)(2N+1)}{2}$; divide by 3. With both sums in hand,

$$\mathbb{E}\,X \overset{(1)}{=} \sum_{x=1}^{N} x\,\frac{1}{N} \overset{(2)}{=} \frac{1}{N}\,\frac{N(N+1)}{2} = \frac{N+1}{2},$$

$$\mathbb{E}\,X^2 \overset{(1)}{=} \sum_{x=1}^{N} x^2\,\frac{1}{N} \overset{(2)}{=} \frac{(N+1)(2N+1)}{6},$$

$$\operatorname{Var} X \overset{(3)}{=} \mathbb{E}\,X^2 - (\mathbb{E}\,X)^2 = \frac{(N+1)(2N+1)}{6} - \frac{(N+1)^2}{4} \overset{(4)}{=} \frac{(N+1)(N-1)}{12}.$$

(1) The expectation of $g(X)$ is the sum of $g(x)$ times the probability of x (Chapter 1).
(2) The two power sums recalled above.
(3) The shortcut $\operatorname{Var} X = \mathbb{E}[(X-\mu)^2] = \mathbb{E}\,X^2 - 2\mu\,\mathbb{E}\,X + \mu^2 = \mathbb{E}\,X^2 - \mu^2$ with $\mu = \mathbb{E}\,X$.
(4) Take out $(N+1)/12$: $2(2N+1) - 3(N+1) = N - 1$.

For an ordinary die ($N = 6$) this gives mean 3.5 and variance $35/12 \approx 2.92$ [CasellaBerger §3.2, p. 70]. For a digitiser that reports one of N equally likely codes, the standard deviation is $\sqrt{(N^2-1)/12} \approx N/\sqrt{12}$, the famous "$1/\sqrt{12}$ of a bin" of quantisation noise.

2.1.3 The Bernoulli trial

Working definition 2.3 (Bernoulli distribution)

$X \sim \text{Bernoulli}(p)$ if X takes the value 1 with probability p and the value 0 with probability $1-p$, for some $p \in [0,1]$ [one trial with two outcomes, success coded as 1]. Its pmf is

$$f(x \mid p) \;=\; p^x(1-p)^{1-x}, \qquad x \in \{0,1\}. \tag{2.1}$$

Recognise it by: a single trial, exactly two outcomes, a fixed probability p.

The exponents in (2.1) act as switches that pick the right factor. To see this, put in the two possible values. At $x = 1$: $p^1(1-p)^0 = p$. At $x = 0$: $p^0(1-p)^1 = 1-p$. The switches pay off when we multiply many such factors, one per trial: the exponents simply add up.

A random variable that is 1 when something happens and 0 otherwise is also called an **indicator**. We write I_A for the indicator of an event A. Every Bernoulli variable is the indicator of its success event, and every indicator is Bernoulli with $p = \mathrm{P}(A)$.

Mean and variance. One fact does all the work: $0^2 = 0$ and $1^2 = 1$, so $X^2 = X$ for a Bernoulli variable.

$$\mathbb{E}\,X \overset{(1)}{=} 0\cdot(1-p) + 1\cdot p = p,$$

$$\mathbb{E}\,X^2 \overset{(2)}{=} \mathbb{E}\,X = p,$$

$$\operatorname{Var} X \overset{(3)}{=} \mathbb{E}\,X^2 - (\mathbb{E}\,X)^2 = p - p^2 = p(1-p). \tag{2.2}$$

(1) Sum of value times probability over the two values.
(2) $X^2 = X$ because the only values are 0 and 1.
(3) The variance shortcut used for the fair die above.

The variance $p(1-p)$ is zero at $p = 0$ and $p = 1$ (a certain outcome is a point mass) and largest, $1/4$, at $p = 1/2$. A fair coin is the most unpredictable yes/no trial there is.

2.1.4 Where does the randomness live? An explicit sample space

A random variable is a map from the sample space Ω of Chapter 1 (the set of all possible outcomes of the experiment) to numbers. In the formulas above Ω never appeared, but it is still

there. For the Bernoulli trial we can build it explicitly [Wasserman p. 27], and the construction is exactly how a computer makes a Bernoulli variable.

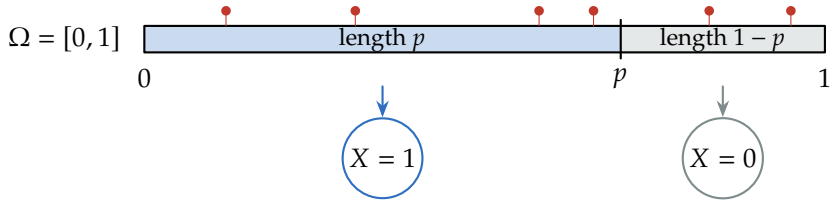


Figure 2.2: The Bernoulli trial as a map. A point ω is thrown uniformly on the unit interval (red dots are six such throws). Every ω in the blue part, of length p, is sent to $X = 1$; every ω in the grey part is sent to $X = 0$. The probability of $X = 1$ is just the length of the blue part.

Take $\Omega = [0, 1]$ with the "length" probability $\mathrm{P}([a, b]) = b - a$ for $0 \le a \le b \le 1$ [a point thrown uniformly on the unit interval]. Fix p and define the map

$$X(\omega) = \begin{cases} 1 & \omega \le p, \\ 0 & \omega > p. \end{cases}$$

Then $\mathrm{P}(X = 1) = \mathrm{P}(\omega \le p) = \mathrm{P}([0, p]) = p$ and $\mathrm{P}(X = 0) = 1 - p$. So $X \sim \text{Bernoulli}(p)$ (Figure 2.2). Every counting distribution can be built this way from uniform numbers. This is the basis of Monte Carlo simulation; Chapter 6 develops it in general (inverse-transform sampling). Here we only use it.

2.1.5 Worked examples

Example: the paddle, and a veto. A paddle has efficiency $\epsilon = 0.9$. Let X be the indicator "paddle fired". Then $\mathbb{E}\,X = 0.9$ and $\mathrm{Var}\,X = 0.9 \times 0.1 = 0.09$, so the standard deviation of a single trial is 0.3.

Now use the same paddle as a *veto*: we keep the event only if the paddle did *not* fire. The indicator of "kept" is $Y = 1 - X$. It is 1 exactly when $X = 0$, which happens with probability $1 - \epsilon$. So $Y \sim \text{Bernoulli}(1 - \epsilon)$, with $\mathbb{E}\,Y = 0.1$ and $\mathrm{Var}\,Y = 0.1 \times 0.9 = 0.09$. The same variance also follows directly from X:

$$\mathrm{Var}\,Y \overset{(1)}{=} \mathrm{Var}(1 - X) \overset{(2)}{=} (-1)^2\,\mathrm{Var}\,X = 0.09.$$

(1) Definition of Y.
(2) Adding a constant does not change a spread, and multiplying by a multiplies the variance by a^2 (Chapter 1).

The veto and the paddle have the same variance: $p(1 - p)$ is symmetric under $p \to 1 - p$, because relabelling success and failure cannot change how unpredictable the trial is.

Example: one trial read two ways. The formula $p^x(1 - p)^{1-x}$ can be read in two directions, exactly the two directions of probability, prediction and inference (Chapter 1).

Prediction. Fix $p = 0.9$. The formula is a distribution over x: $f(1) = 0.9$, $f(0) = 0.1$, and $\sum_x f(x) = 1$.

Inference. Suppose we observed one firing, $x = 1$. Read the same formula as a function of p: $\mathcal{L}(p) = p^1(1 - p)^0 = p$. This is the **likelihood** of p. It is *not* a probability distribution over p:

$$\int_0^1 \mathcal{L}(p)\,\mathrm{d}p \overset{(1)}{=} \int_0^1 p\,\mathrm{d}p \overset{(2)}{=} \left[\tfrac{p^2}{2}\right]_0^1 = \tfrac{1}{2} \neq 1.$$

(1) The likelihood for the observation $x = 1$.
(2) Elementary integral.

The likelihood tells us that values of p near 1 make the observed firing more plausible than values near 0, and nothing more [Kruschke §6.1, p. 125]. Turning it into a distribution over p needs a prior and Bayes' theorem (Chapter 5).

Example: many independent trials of the same paddle. Record N muons, with outcomes $x_1, \ldots, x_N$, each 0 or 1, and assume the trials are independent. The probability of this particular sequence is

$$\mathrm{P}(x_1, \ldots, x_N \mid p) \overset{(1)}{=} \prod_{i=1}^{N} p^{x_i}(1-p)^{1-x_i} \overset{(2)}{=} p^{\sum_i x_i}(1-p)^{\sum_i (1-x_i)} \overset{(3)}{=} p^{z}(1-p)^{N-z},$$

(1) Independence: the joint probability is the product of the single-trial probabilities (Chapter 1).
(2) Powers of the same base multiply by adding exponents.
(3) Write $z = \sum_i x_i$ for the number of firings. Then $\sum_i (1 - x_i) = N - z$.

Only the count z survives: the order in which the firings happened does not matter. This collapse is what the switch exponents in (2.1) were for, and it is the seed of the binomial distribution in the next section [Kruschke (6.2), p. 125].

2.1.6 Python: build the trial from a uniform number

The script `code/ch02/01_bernoulli.py` does exactly what Figure 2.2 shows. The core is three lines:

Listing 2.1: One Bernoulli trial per uniform number (`code/ch02/01_bernoulli.py`)

```
# --- one trial = one uniform number compared with p -----------------------
def bernoulli(p, size):
    u = rng.uniform(0.0, 1.0, size)      # omega, uniform on [0,1]
    return (u <= p).astype(int)          # X(omega) = 1 if omega <= p else 0
```

Line by line: `rng.uniform` throws the points ω on $[0, 1]$; the comparison `u <= p` is the map $X(\omega)$, returning `True` or `False`; `astype(int)` codes these as 1 and 0. The rest of the script runs 5000 trials for three values of p and plots the running fraction of successes, then draws 10^6 trials at $p = 0.9$ and compares the observed frequencies with (2.1).

For 10^6 trials at $p = 0.9$ the sample mean is 0.9002 and the sample variance is 0.0898, against the derived $p = 0.9$ and $p(1-p) = 0.09$ (Figure 2.3). On the left, the running fraction slowly settles onto p: this is the law of large numbers at work. Its typical distance from p after n trials shrinks like $\sqrt{p(1-p)/n}$, which follows from the binomial variance in the next section.

What this bought us. A yes/no trial is fully described by one number p. The pmf $p^x(1-p)^{1-x}$ is built so that products over many independent trials collapse onto the number of successes. The mean is p and the variance is $p(1-p)$, largest for a fair coin.

2.2 Many successive trials: the binomial distribution

A downward-going muon crosses all four scintillator layers of a cosmic-ray telescope. Given such a crossing, each layer fires with probability $\epsilon = 0.95$, independently of the others. To reject

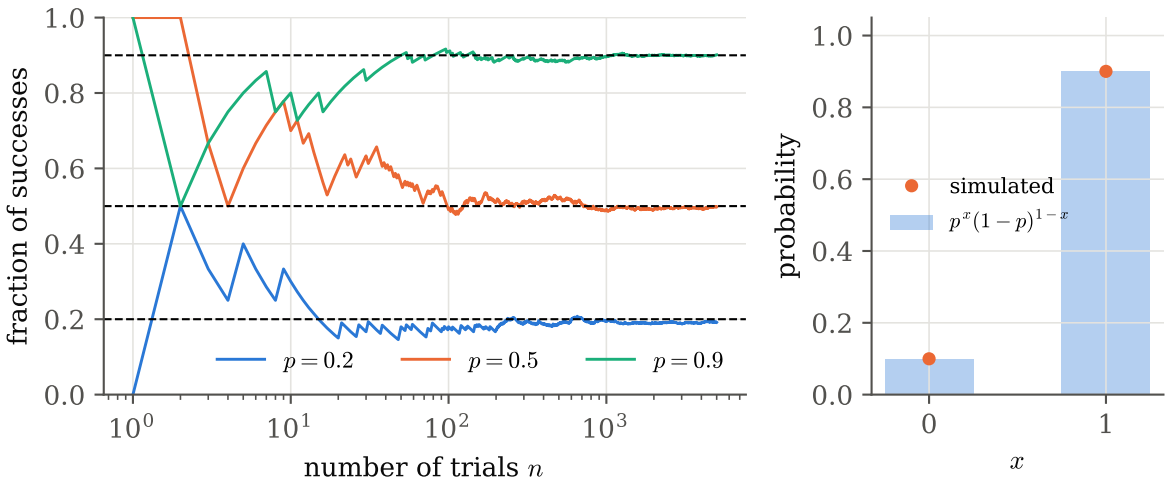


Figure 2.3: Left: the fraction of successes after n trials, for three paddles. Early on the fraction jumps around wildly; by a few thousand trials each curve has settled onto its dashed line at p. Right: for $p = 0.9$ the bars are the pmf and the dots the frequencies in a million simulated trials; they coincide.

noise, the trigger demands that *at least three* of the four layers fire. What fraction of real muons does the trigger keep? Requiring all four would be cleaner, but how many muons would it cost? Both questions need the same thing: the probability of exactly k successes in n independent trials. The binomial distribution supplies it. It turns "each trial succeeds with probability p" into the full distribution of the number of successes.

Physical picture: paths through a tree. Picture each trial as a fork in a road: up for success, down for failure. After n forks, every possible history is a path, and the number of successes is the number of "up" steps along that path. A random walk in one dimension has the same combinatorics. So do n non-interacting spin-$\frac{1}{2}$ particles: the number of spin-up particles is binomially distributed. The binomial coefficient $\binom{n}{k}$ that appears below is the *multiplicity* of the macrostate "k up", the same multiplicity that statistical mechanics uses to define entropy.

The question, step by step.

1. What is the probability of one *particular* sequence of outcomes, say success, failure, success? Multiply, because the trials are independent.
2. Does that probability depend on the order? No: only on how many successes there are.
3. How many sequences have exactly k successes? Count them.
4. So the probability of k successes is (number of sequences) × (probability of each one).

2.2.1 Deriving the pmf by counting paths

Take n independent trials, each a Bernoulli(p). A history is a string of n letters, S (success) or F (failure). By the product formula for independent Bernoulli trials (Section 2.1), a string with k S's has probability $p^k(1-p)^{n-k}$ whatever the order.

How many strings have exactly k S's? We must choose which k of the n positions carry an S. Count ordered choices first: there are n ways to pick the first position, $n-1$ for the second,

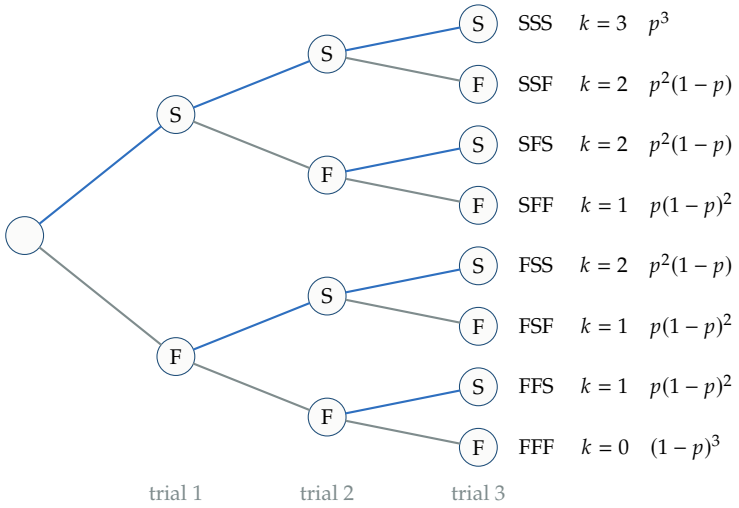


Figure 2.4: All $2^3 = 8$ histories of three trials. Blue branches are successes (probability p), grey ones failures (probability $1-p$). Each path's probability is the product along it and depends only on its number of successes k. Counting paths per k gives $1, 3, 3, 1$, the third row of Pascal's triangle.

and so on down to $n-k+1$ for the k-th, giving $n(n-1)\cdots(n-k+1) = \frac{n!}{(n-k)!}$ ordered choices. Each set of k positions has been counted $k!$ times, once for each order in which we could have picked the same positions. So the number of sets is

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}, \tag{2.3}$$

the **binomial coefficient**, read "n choose k". (By convention $0! = 1$, so $\binom{n}{0} = \binom{n}{n} = 1$.)

Working definition 2.4 (binomial distribution)

$X \sim \text{Binomial}(n, p)$ if X is the number of successes in n independent trials, each succeeding with the same probability p [n fixed in advance, trials do not influence each other, p the same for every trial]. Its pmf is

$$f(k \mid n, p) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, \ldots, n. \tag{2.4}$$

Recognise it by: multiple successive trials (where Bernoulli is for a single trial), a fixed number of them, each yes/no, independent, with constant p, and we record only the total number of successes.

It is normalised. The probabilities (2.4) add up to one. The tool is the **binomial theorem**: $(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$. Here is why it is true. Expanding the product of n factors $(a+b)(a+b)\cdots(a+b)$ means choosing a or b from each factor. A term $a^k b^{n-k}$ arises once for each way of choosing which k factors give a, and there are $\binom{n}{k}$ such ways. Then

$$\sum_{k=0}^{n} f(k \mid n, p) \overset{(1)}{=} \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \overset{(2)}{=} \big(p + (1-p)\big)^n = 1^n = 1.$$

(1) Definition (2.4).
(2) The binomial theorem with $a = p$ and $b = 1-p$.

With $p = \frac{1}{2}$ the same line says $\sum_k \binom{n}{k} = 2^n$: the total number of paths in the tree [CasellaBerger (3.2.4), p. 74].

2.2.2 Mean and variance, two ways

The slick way: a sum of indicators. Let X_i be the indicator of success on trial i: it is 1 if trial i succeeds and 0 otherwise. Counting successes is adding up these indicators, so $X = X_1 + \cdots + X_n$, and the X_i are independent Bernoulli(p).

$$\mathbb{E}\,X \overset{(1)}{=} \sum_{i=1}^{n} \mathbb{E}\,X_i \overset{(2)}{=} np,$$
$$\operatorname{Var} X \overset{(3)}{=} \sum_{i=1}^{n} \operatorname{Var} X_i \overset{(4)}{=} np(1-p). \tag{2.5}$$

(1) Expectation is linear: $\mathbb{E}(\sum X_i) = \sum \mathbb{E}\,X_i$, with or without independence (Chapter 1).
(2) $\mathbb{E}\,X_i = p$ from (2.2).
(3) For *independent* variables the covariances vanish, so the variance of a sum is the sum of variances.
(4) $\operatorname{Var} X_i = p(1-p)$ from (2.2).

The direct way: sum the pmf. This route uses a trick that recurs throughout the chapter: absorb the factor k into the binomial coefficient, then recognise the sum that is left as the total probability of a smaller distribution. The identity we need is

$$k\binom{n}{k} = k\frac{n!}{k!(n-k)!} = \frac{n!}{(k-1)!(n-k)!} = n\binom{n-1}{k-1} \qquad (k \geq 1).$$

Then

$$\mathbb{E}\,X \overset{(1)}{=} \sum_{k=1}^{n} k\binom{n}{k}p^k(1-p)^{n-k} \overset{(2)}{=} np\sum_{k=1}^{n}\binom{n-1}{k-1}p^{k-1}(1-p)^{(n-1)-(k-1)}$$
$$\overset{(3)}{=} np\sum_{j=0}^{n-1}\binom{n-1}{j}p^j(1-p)^{n-1-j} \overset{(4)}{=} np.$$

(1) Definition of the mean. The $k = 0$ term is zero, so we start at $k = 1$.
(2) The identity above, and one factor p taken out of p^k.
(3) Rename $j = k - 1$.
(4) The sum is the total probability of a Binomial($n-1, p$), which is 1.

For the variance, the same trick works on the *factorial moment* $\mathbb{E}[X(X-1)]$, using $k(k-1)\binom{n}{k} = n(n-1)\binom{n-2}{k-2}$:

$$\mathbb{E}[X(X-1)] \overset{(1)}{=} n(n-1)p^2\sum_{j=0}^{n-2}\binom{n-2}{j}p^j(1-p)^{n-2-j} \overset{(2)}{=} n(n-1)p^2,$$
$$\operatorname{Var} X \overset{(3)}{=} \mathbb{E}[X(X-1)] + \mathbb{E}\,X - (\mathbb{E}\,X)^2 = n(n-1)p^2 + np - n^2p^2 = np(1-p).$$

(1) The same three steps as for the mean, with two factors of k absorbed and $j = k - 2$.
(2) A Binomial($n-2, p$) sums to one.
(3) $\mathbb{E}\,X^2 = \mathbb{E}[X(X-1)] + \mathbb{E}\,X$, and then the variance shortcut.

Both ways agree with [Cowan (2.3)–(2.4), p. 27].[1]

[1] Cowan writes the upper limit of the sum in (2.3) as ∞. It should be n, the number of trials, since the count cannot exceed it.

What the numbers mean. The mean count np grows like n, but the standard deviation $\sqrt{np(1-p)}$ grows only like $\sqrt{n}$. Dividing both by n, the *fraction* of successes X/n has mean p and standard deviation $\sqrt{p(1-p)/n}$. This is the rate at which the running fractions in Figure 2.3 settle, and the origin of every "$1/\sqrt{N}$" error bar on an efficiency.

2.2.3 The moment generating function and sums of binomials

The **moment generating function** (mgf) $M_X(t) = \mathbb{E}\, e^{tX}$ (Chapter 1) packs all the moments into one function: its r-th derivative at $t = 0$ is $\mathbb{E}\, X^r$. For the binomial,

$$M_X(t) \overset{(1)}{=} \sum_{k=0}^{n} e^{tk}\binom{n}{k}p^k(1-p)^{n-k} \overset{(2)}{=} \sum_{k=0}^{n}\binom{n}{k}(pe^t)^k(1-p)^{n-k} \overset{(3)}{=} \left(1-p+pe^t\right)^n. \tag{2.6}$$

(1) Definition of the mgf.
(2) $e^{tk}p^k = (pe^t)^k$.
(3) The binomial theorem with $a = pe^t$ and $b = 1-p$.

Check: $M'(0) = n(1-p+p)^{n-1}p = np$, the mean again.

Key idea: binomials with the same p add

If $X_1 \sim \text{Binomial}(n_1, p)$ and $X_2 \sim \text{Binomial}(n_2, p)$ are independent, then $X_1 + X_2 \sim \text{Binomial}(n_1 + n_2, p)$.

Why: the counting story gives it at once. X_1 counts successes in n_1 trials, and X_2 counts them in another n_2 independent trials with the same p. So $X_1 + X_2$ counts successes in $n_1 + n_2$ such trials. The mgf says the same. The mgf of a sum of independent variables is the product of their mgfs, and $(1-p+pe^t)^{n_1}(1-p+pe^t)^{n_2} = (1-p+pe^t)^{n_1+n_2}$, the mgf (2.6) of Binomial(n_1+n_2, p). The same p is essential: with different p's the product is not of binomial form.

2.2.4 Worked examples

Example: the coincidence trigger. Return to the four-layer telescope from the start of the section. Each layer fires with $\epsilon = 0.95$, independently, so the number of layers that fire is $X \sim \text{Binomial}(4, 0.95)$. The trigger needs $X \geq 3$.

$$\begin{aligned} \mathrm{P}(X \geq 3) &\overset{(1)}{=} \binom{4}{3}\epsilon^3(1-\epsilon) + \binom{4}{4}\epsilon^4 \\ &\overset{(2)}{=} 4(0.857375)(0.05) + 0.81450625 \\ &= 0.171475 + 0.814506 = 0.98598. \end{aligned}$$

(1) Only $k = 3$ and $k = 4$ satisfy the condition. Add their pmf values.
(2) $0.95^3 = 0.857375$ and $0.95^4 = 0.81450625$.

The 3-out-of-4 trigger keeps 98.6% of muons. Demanding all four keeps only $\epsilon^4 = 0.8145$, so the cleaner trigger loses about one muon in five. The simulation in `02_binomial.py` (one million four-layer events) gives 0.9859 against the exact 0.986.

Example: the Chevalier de Méré's two bets. Is it more likely to get at least one six in 4 rolls of a die, or at least one double six in 24 rolls of two dice? The number of successes is binomial

in both cases, and "at least one" is the complement of "none":

$$\mathrm{P}(\text{at least one six in 4}) \overset{(1)}{=} 1 - \binom{4}{0}\left(\tfrac{1}{6}\right)^0\left(\tfrac{5}{6}\right)^4 = 1 - 0.4823 = 0.518,$$

$$\mathrm{P}(\text{at least one double six in 24}) \overset{(2)}{=} 1 - \left(\tfrac{35}{36}\right)^{24} = 1 - 0.5086 = 0.491.$$

(1) Binomial$(4, 1/6)$ at $k = 0$.
(2) Binomial$(24, 1/36)$ at $k = 0$: two dice show a double six with probability $1/36$.

The gambler's intuition ($4 \times \frac{1}{6} = 24 \times \frac{1}{36} = \frac{2}{3}$) is wrong because it adds probabilities of events that are not mutually exclusive [CasellaBerger Ex. 3.2.3, p. 75]. The same mistake appears in "the probability of at least one fake signal in N search bins is N times the single-bin probability": true only while that product is small (the look-elsewhere effect, Chapter 4).

Example: drawing without replacement: the hypergeometric. The binomial needs the same p on every trial. That fails when we sample *without replacement* from a small population, because each draw changes what is left. A batch of $N = 25$ front-end chips contains M faulty ones. We test $K = 10$ chips at random and find none faulty. How surprising is this if $M = 6$?

The number of faulty chips in the sample follows the **hypergeometric** distribution. Count equally likely samples: there are $\binom{N}{K}$ ways to choose K chips, and $\binom{M}{x}\binom{N-M}{K-x}$ of them contain exactly x faulty ones (choose x of the M faulty and $K - x$ of the $N - M$ good ones). So

$$\mathrm{P}(X = x \mid N, M, K) = \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}},$$

with mean KM/N and variance $\frac{KM}{N}\frac{(N-M)(N-K)}{N(N-1)}$. For our batch,

$$\mathrm{P}(X = 0) \overset{(1)}{=} \frac{\binom{6}{0}\binom{19}{10}}{\binom{25}{10}} \overset{(2)}{=} \frac{92378}{3268760} = 0.028.$$

(1) Hypergeometric with $N = 25$, $M = 6$, $K = 10$, $x = 0$.
(2) $\binom{19}{10} = 92378$ and $\binom{25}{10} = 3268760$.

If six chips were faulty, a clean sample is unlikely: under 3% [CasellaBerger Ex. 3.2.1, p. 73]. The binomial with $p = M/N = 0.24$ would give $(0.76)^{10} = 0.064$, more than twice as large. Why the difference? Without replacement, each good chip drawn leaves a remaining pool that is *more* faulty, so a clean run is rarer. The factor $(N - K)/(N - 1)$ in the variance is the "finite-population correction". When the batch is much larger than the sample, $N \gg K$, this factor goes to one and the hypergeometric becomes the binomial.

2.2.5 Python: repeat the n trials, histogram the count

The script `code/ch02/02_binomial.py` simulates the generating process itself, not the formula. It builds a table of uniform numbers, one row per experiment and one column per trial.

Listing 2.2: Binomial counts by running the trials (`code/ch02/02_binomial.py`)

```
21 def binomial_by_trials(n, p, reps):
22     """Run 'reps' experiments; each is n independent Bernoulli(p) trials."""
23     trials = rng.uniform(size=(reps, n)) <= p     # True = success
24     return trials.sum(axis=1)                      # number of successes per experiment
```

`rng.uniform(size=(reps, n)) <= p` is a `reps`×n table of Bernoulli trials, built exactly as in Section 2.1. Summing along a row (`axis=1`) counts the successes in one experiment. The script does this 20 000 times for six (n, p) pairs and overlays `scipy.stats.binom.pmf` as black ticks (Figure 2.5).

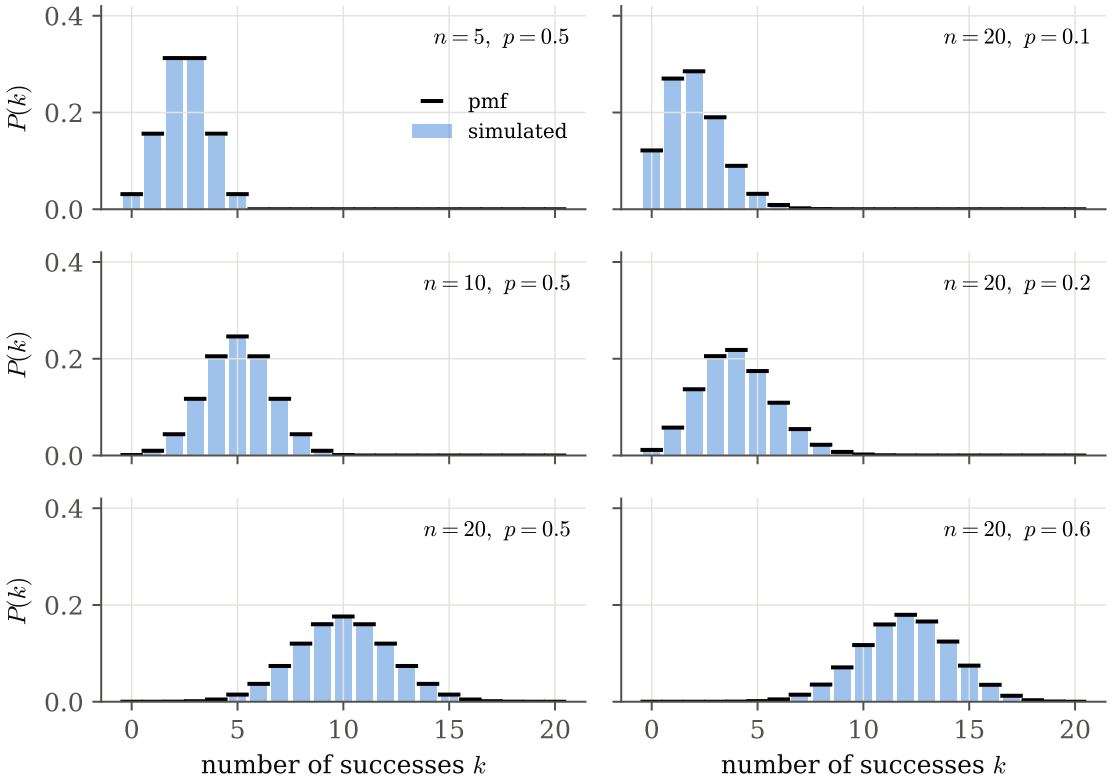


Figure 2.5: Histograms of simulated success counts (bars) against the pmf (black ticks). Left column: $p = \frac{1}{2}$ and more trials; the distribution stays symmetric, moves right and widens like $\sqrt{n}$. Right column: $n = 20$ and growing p; at small p the distribution is squashed against $k = 0$ and skewed, at $p = 0.6$ it leans the other way.

For $n = 20$, $p = 0.2$ and 200 000 experiments the sample mean is 4.001 and the sample variance 3.205, against $np = 4$ and $np(1-p) = 3.2$. Figure 2.5 shows two features that matter later. For $p = \frac{1}{2}$ and growing n the shape is already bell-like: this is the de Moivre–Laplace limit, derived in the next part. For small p the distribution piles up near zero with a tail to the right: this is the regime that becomes the Poisson distribution in Section 2.4.

Watch out: when the binomial is the wrong model

Three assumptions go into (2.4), and each can fail in a real detector.

- *Fixed n.* If the number of trials is itself random (the number of muons in a run), the count of successes is not binomial. We return to this in Section 2.5.
- *Independence.* Layers that share a high-voltage supply fail together, so the number of dead layers is spread wider than binomial.
- *Constant p.* Efficiency that varies across the paddle (edges versus centre), or between runs, also inflates the spread. The *beta-binomial* distribution models a p that itself varies [Lambert §8.4.5]. Sampling without replacement (the hypergeometric example above) is the opposite case: the spread is *narrower* than binomial.

A quick check is to compare the observed variance of a count with $np(1-p)$ computed

from its observed mean.

2.3 Histogram bins: the multinomial distribution

A calorimeter records the energy of $N = 50$ electrons from a beta source, and we fill a histogram with $m = 5$ energy bins. Each electron lands in bin j with probability p_j, set by the shape of the spectrum. The number of electrons in bin j, the bin count n_j, fluctuates from one run to the next. To fit the spectrum later (Chapter 3) we need the *joint* distribution of all the bin contents, including how the bins fluctuate together. That joint distribution is the multinomial: the binomial with more than two outcomes. Its main lesson is that, with N fixed, the bins of a histogram are *negatively correlated*.

Physical picture: a fixed number of balls in boxes. Fifty balls are thrown one at a time into five boxes. If box 1 happens to catch more than its share, there are fewer balls left for the other boxes. The total is fixed, so the counts cannot all go up together. Statistical mechanics has the same situation in the canonical ensemble, where the particle number is fixed. Letting the total fluctuate, as in the grand canonical ensemble, removes the constraint. We will see that it also removes the correlation.

The question, step by step.

1. What is the probability of one *particular* sequence of bin assignments (electron 1 in bin 3, electron 2 in bin 1, ...)? Multiply the p's.
2. Which feature of the sequence does that probability depend on? Only how many electrons went into each bin.
3. How many sequences give the counts $(n_1, \dots, n_m)$? Count them.
4. What does one bin look like on its own, and how do two bins co-vary?

2.3.1 Deriving the pmf

Each of N independent trials ends in one of m outcomes, with probabilities $p_1, \dots, p_m$ and $\sum_j p_j = 1$. A particular sequence with n_1 outcomes of type 1, n_2 of type 2, and so on, has probability $p_1^{n_1} p_2^{n_2} \cdots p_m^{n_m}$ by independence, whatever the order.

How many sequences have these counts? Start by treating the N trials as labelled objects in a row: there are $N!$ orderings. Many orderings give the same assignment of outcomes. Shuffling the n_1 trials of type 1 among themselves ($n_1!$ ways) does not change which trials are of type 1, and the same holds for every type. So each distinct assignment has been counted $n_1!\, n_2! \cdots n_m!$ times, and the number of distinct assignments is

$$\binom{N}{n_1, n_2, \dots, n_m} = \frac{N!}{n_1!\, n_2! \cdots n_m!}, \tag{2.7}$$

the **multinomial coefficient**. For $m = 2$ it is $N!/(n_1!(N - n_1)!)$, the binomial coefficient.

Working definition 2.5 (multinomial distribution)

The vector of counts $\boldsymbol{n} = (n_1, \dots, n_m)$ is Multinomial$(N, \boldsymbol{p})$ if it records how many of N independent trials fell into each of m categories, each trial landing in category j with the

same probability p_j [N fixed, categories exclusive and exhaustive, $\sum_j p_j = 1$]. Its pmf is

$$f(n_1, \ldots, n_m \,|\, N, \boldsymbol{p}) \;=\; \frac{N!}{n_1! \cdots n_m!}\, p_1^{n_1} \cdots p_m^{n_m}, \qquad n_j \geq 0, \quad \sum_j n_j = N. \tag{2.8}$$

Recognise it by: a histogram of a fixed number of independent entries, or any tally of categorical outcomes (particle species, decay channels, blood types).

The pmf (2.8) sums to one by the *multinomial theorem*, $(p_1 + \cdots + p_m)^N = \sum \frac{N!}{n_1! \cdots n_m!} p_1^{n_1} \cdots p_m^{n_m}$. It is proved exactly like the binomial theorem: expanding the product means picking one p_j from each of the N factors. The left side is $1^N = 1$.

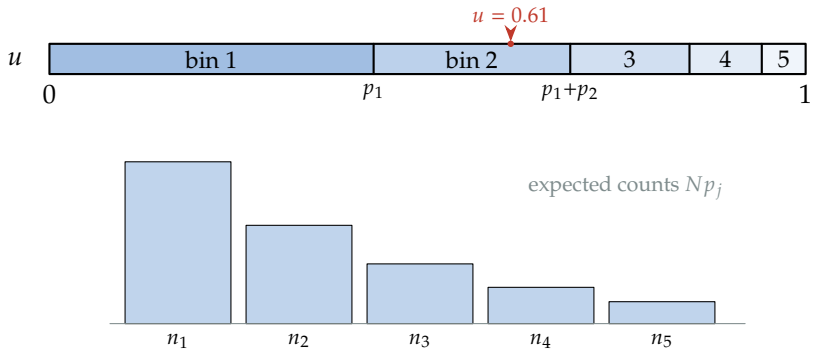


Figure 2.6: How one event chooses its bin. The unit interval is cut into pieces of lengths $p_1, \ldots, p_5$ (a falling spectrum). A uniform number u lands in one piece, here bin 2, and that bin's count goes up by one. Repeating this N times fills the histogram below, whose bar heights are the expected counts Np_j for $N = 50$.

2.3.2 One bin at a time: the marginal is binomial

Often we care about one bin only. Then the other bins are *nuisance* variables, and we **marginalise**: sum the joint pmf over all values of the variables we do not care about (Chapter 1).

Key idea: each bin is binomial

If $\boldsymbol{n} \sim \text{Multinomial}(N, \boldsymbol{p})$, then $n_j \sim \text{Binomial}(N, p_j)$ for every j.

Why: look at bin j alone. Each trial either lands in bin j (probability p_j) or does not (probability $1 - p_j$). Lump all the other bins into the single outcome "not in bin j", and we have N independent yes/no trials with the same probability: the binomial story. The algebra confirms it. Take three categories (bin i, bin j, "everything else"), where the pmf is the *trinomial*

$$f(n_i, n_j) = \frac{N!}{n_i!\, n_j!\, (N - n_i - n_j)!}\, p_i^{n_i} p_j^{n_j} (1 - p_i - p_j)^{N - n_i - n_j}.$$

Summing out n_j:

$$\sum_{n_j=0}^{N-n_i} f(n_i, n_j) \overset{(1)}{=} \frac{N!}{n_i!\,(N-n_i)!} p_i^{n_i} \sum_{n_j=0}^{N-n_i} \frac{(N-n_i)!}{n_j!\,(N-n_i-n_j)!}\, p_j^{n_j} (1 - p_i - p_j)^{N-n_i-n_j}$$

$$\overset{(2)}{=} \binom{N}{n_i} p_i^{n_i} \left(p_j + 1 - p_i - p_j\right)^{N-n_i} \overset{(3)}{=} \binom{N}{n_i} p_i^{n_i} (1 - p_i)^{N-n_i}.$$

(1) Multiply and divide by $(N-n_i)!$, and take out everything that does not depend on n_j.
(2) The remaining sum is the binomial theorem with $n = N-n_i$, $a = p_j$, $b = 1-p_i-p_j$.
(3) The two p_j's cancel.

The result is the Binomial(N, p_i) pmf, so $\mathbb{E}\, n_i = Np_i$ and $\operatorname{Var} n_i = Np_i(1-p_i)$ from (2.5), and the same holds for every bin. Summing over the nuisance bin removed it and left the distribution of the one bin we care about.

2.3.3 Two bins together: the covariance

The question is how two bin counts n_i and n_j fluctuate together. The tool is the same as for the binomial mean: write each count as a sum of indicators, one per event. Let I_{ej} be the indicator that event e landed in bin j (it is 1 if so, 0 if not). Then $n_j = \sum_{e=1}^{N} I_{ej}$. Two facts about these indicators do all the work:

- *same event, different bins:* one event lands in exactly one bin, so $I_{ei}I_{ej} = 0$ for $i \neq j$;
- *different events:* I_{ei} and $I_{e'j}$ are independent for $e \neq e'$, so their covariance is zero.

For $i \neq j$,

$$\begin{aligned}\operatorname{Cov}(n_i, n_j) &\overset{(1)}{=} \sum_{e=1}^{N}\sum_{e'=1}^{N} \operatorname{Cov}(I_{ei}, I_{e'j}) \overset{(2)}{=} \sum_{e=1}^{N} \operatorname{Cov}(I_{ei}, I_{ej}) \\ &\overset{(3)}{=} \sum_{e=1}^{N} \big(\mathbb{E}[I_{ei}I_{ej}] - \mathbb{E}\, I_{ei}\ \mathbb{E}\, I_{ej}\big) \overset{(4)}{=} \sum_{e=1}^{N} (0 - p_i p_j) = -Np_ip_j. \end{aligned} \tag{2.9}$$

(1) Covariance is linear in each argument, so the covariance of two sums is the sum of all pairwise covariances.
(2) Terms with $e \neq e'$ vanish: different events are independent.
(3) Definition $\operatorname{Cov}(A,B) = \mathbb{E}[AB] - \mathbb{E}\, A\ \mathbb{E}\, B$.
(4) $I_{ei}I_{ej} = 0$, and $\mathbb{E}\, I_{ej} = p_j$ because an indicator is Bernoulli.

Together with the variances, the full **covariance matrix** of the histogram is

$$\mathbf{C}_{ij} \;=\; \operatorname{Cov}(n_i, n_j) \;=\; N\big(p_i\delta_{ij} - p_ip_j\big), \qquad \mathbf{C} = N\big(\operatorname{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^{\mathsf{T}}\big), \tag{2.10}$$

and the correlation of two bins is

$$\operatorname{Corr}(n_i, n_j) = \frac{-Np_ip_j}{\sqrt{Np_i(1-p_i)\,Np_j(1-p_j)}} = -\sqrt{\frac{p_ip_j}{(1-p_i)(1-p_j)}}.$$

It does not depend on N: more events do not make the anti-correlation go away.

The matrix is singular. Every row of $\mathbf{C}$ sums to zero: $\sum_j N(p_i\delta_{ij} - p_ip_j) = N(p_i - p_i\sum_j p_j) = 0$. So the vector $\mathbf{1} = (1,\dots,1)$ is a zero mode, $\mathbf{C}\,\mathbf{1} = 0$, and $\operatorname{Var}(\sum_j n_j) = \mathbf{1}^{\mathsf{T}}\mathbf{C}\mathbf{1} = 0$. This is no surprise: the total $\sum_j n_j = N$ does not fluctuate at all. A χ^2 built from the full $\mathbf{C}$ must therefore drop one bin or use a pseudo-inverse, a point that returns when we fit histograms (Chapter 3).

2.3.4 Worked examples

Example: a die rolled twelve times. What is the probability that a fair die rolled 12 times shows each face exactly twice?

$$f(2,2,2,2,2,2) \overset{(1)}{=} \frac{12!}{(2!)^6}\left(\tfrac{1}{6}\right)^{12} \overset{(2)}{=} \frac{479001600}{64}\cdot\frac{1}{2176782336} = \frac{7484400}{2176782336} = 0.00344.$$

(1) Multinomial with $N = 12$, $m = 6$, all $p_j = 1/6$, all $n_j = 2$.
(2) $12! = 479001600$, $(2!)^6 = 64$ and $6^{12} = 2176782336$.

The "perfectly flat" histogram is rare: about one run in 290. A histogram that looks too flat should make us as suspicious as one that looks too bumpy.

Example: blood groups in a sample. In a population the ABO blood groups O, A, B, AB have frequencies roughly $p = (0.45, 0.40, 0.11, 0.04)$. In a sample of $N = 10$ donors, what is the probability of exactly 5 O, 4 A, 1 B and no AB?

$$f \overset{(1)}{=} \frac{10!}{5!\,4!\,1!\,0!}(0.45)^5(0.40)^4(0.11)^1(0.04)^0 \overset{(2)}{=} 1260 \times 0.018453 \times 0.0256 \times 0.11 = 0.0655.$$

(1) Multinomial pmf with $N = 10$, $m = 4$.
(2) $10!/(5!\,4!) = 3628800/(120 \times 24) = 1260$, $0.45^5 = 0.018453$, $0.40^4 = 0.0256$.

Each donor is one categorical trial. The multinomial aggregates the categorical outcomes into counts [Lambert §8.4.10], exactly as a histogram aggregates individual events into bins.

Example: the covariance of the first two bins of the spectrum. Return to the beta spectrum from the start of the section: $N = 50$ electrons in five bins. Take a falling spectrum, with the five bin probabilities proportional to $e^{-j/2}$ for $j = 0, \ldots, 4$. Normalising them to sum to one gives $p_1 = 0.4287$ and $p_2 = 0.26$ for the first two bins. Then

$$\begin{aligned} \operatorname{Var} n_1 &= N p_1(1-p_1) = 50 \times 0.4287 \times 0.5713 = 12.25, \\ \operatorname{Cov}(n_1, n_2) &= -N p_1 p_2 = -50 \times 0.4287 \times 0.2600 = -5.57, \\ \operatorname{Corr}(n_1, n_2) &= -\sqrt{\frac{0.4287 \times 0.2600}{0.5713 \times 0.7400}} = -0.513. \end{aligned}$$

A strong anti-correlation: a run with an excess in the first bin very probably has a deficit in the second.

2.3.5 Python: drop the events into the bins

Listing 2.3: Filling many histograms event by event (`code/ch02/03_multinomial.py`)

```
# a falling "energy spectrum": bin probabilities proportional to exp(-bin/2)
m, N, reps = 5, 50, 30_000
p = np.exp(-np.arange(m) / 2.0)
p /= p.sum()

def histogram_of_events(N):
    """Drop N events into bins one by one: each event picks bin j with prob p_j."""
    u = rng.uniform(size=N)
    bins = np.searchsorted(np.cumsum(p), u)      # the bin whose cumulative interval
    contains u
    return np.bincount(bins, minlength=m)
```

```
30
31 H = np.array([histogram_of_events(N) for _ in range(reps)])        # fixed N
32 Hp = np.array([histogram_of_events(rng.poisson(N)) for _ in range(reps)])  # N ~ Poisson
```

The function `histogram_of_events` is Figure 2.6 in code. `np.cumsum(p)` gives the cut points p_1, p_1+p_2, ... of the unit interval; `np.searchsorted` returns, for each uniform number, the index of the piece it fell into, that is, the bin; `np.bincount` counts the events per bin. Line 31 fills 30 000 histograms with exactly $N = 50$ events. Line 32 does the same but first draws the number of events from a Poisson distribution with mean 50 (the distribution of Section 2.4), to see what happens when the total is allowed to fluctuate.

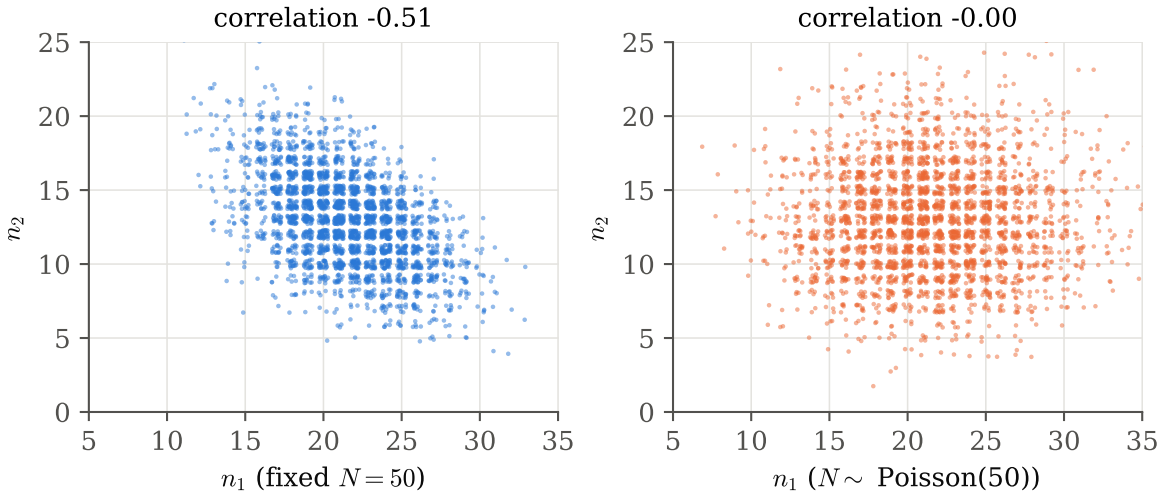


Figure 2.7: Contents of bin 2 against bin 1 for 3000 simulated histograms (a small random jitter separates the integer points). Left: with exactly 50 events per histogram the cloud slopes downwards, because an extra event in bin 1 is an event missing from somewhere else. Right: when the number of events is itself Poisson, the cloud is round and the two bins are uncorrelated.

The simulated variance of bin 1 is 12.27 (theory 12.25), the covariance of bins 1 and 2 is −5.577 (theory −5.572) and their correlation is −0.5119 (theory −0.5134). The variance of the row sums is exactly 0: the zero mode of $\mathbf{C}$. With a Poisson number of events the correlation drops to −0.002175, compatible with zero (Figure 2.7). We prove in Section 2.5 that Poisson-distributed totals make the bins *exactly* independent. This is why fits to histograms whose total is not fixed in advance (most collider analyses) may treat bins as independent Poisson counts.

Watch out: independent bins are an assumption

It is tempting to treat histogram bins as independent, each with error $\sqrt{n_j}$. With a fixed number of entries that is wrong: the bins are anti-correlated by (2.9). For many bins with small p_j the correlation $\approx -\sqrt{p_i p_j}$ is small and the independent approximation is usually fine; for a histogram with a few dominant bins it is not. Normalised shapes ("unit-area" histograms) always carry this constraint.

2.4 Rare events in many trials: the Poisson distribution

A microgram of a long-lived isotope contains of order 10^{15} nuclei. In one second each nucleus decays with a tiny probability p, and the decays are independent. A Geiger counter records the number of decays in each one-second window. Strictly, this count is Binomial(n, p) with

$n \approx 10^{15}$ nuclei. But nobody knows n or p separately, and nobody wants to evaluate $\binom{10^{15}}{k}$. What we *can* measure is the average number of clicks per window, $\lambda = np$. The Poisson distribution is what the binomial becomes when the trials are very many, each is very unlikely to succeed, and only their product λ matters. It is the distribution of *counts of rare events*.

Physical picture: rare, independent, many. Photons arriving at a CCD pixel, dark counts in a photomultiplier, cosmic rays through a detector, events in one bin of a collider histogram, galaxies in a small cell of a sparse survey: each count is the sum of a huge number of chances, each tiny, none influencing the others. In every one of these cases only the mean count λ enters the answer. The microscopic details, such as how many photons were emitted upstream or how many protons collided, drop out.

The question, step by step.

1. Start from the exact binomial for n trials with $p = \lambda/n$, so that the mean $np = \lambda$ stays fixed.
2. Let $n \to \infty$. Which factors change and which survive?
3. Is the limit a proper distribution? What are its mean and variance?
4. How large must n be for the approximation to be good?

2.4.1 The limit, line by line

The derivation needs one limit from calculus. We prove it first.

The compound-interest limit. For any fixed real λ, $\lim_{n\to\infty}\left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda}$. *Why:* take the logarithm and use the Taylor series $\ln(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \cdots$, valid for $|x| < 1$:

$$\ln\left(1-\frac{\lambda}{n}\right)^n \overset{(1)}{=} n\ln\left(1-\frac{\lambda}{n}\right) \overset{(2)}{=} n\left(-\frac{\lambda}{n} - \frac{\lambda^2}{2n^2} - \cdots\right) = -\lambda - \frac{\lambda^2}{2n} - \cdots \xrightarrow{n\to\infty} -\lambda.$$

(1) $\ln a^n = n \ln a$.
(2) Taylor series with $x = \lambda/n$, which is less than one in size once $n > |\lambda|$.

Exponentiate: the limit is $e^{-\lambda}$. Figure 2.8 shows how fast it is approached. The correction $-\lambda^2/(2n)$ in the exponent tells us that $(1-\lambda/n)^n$ is always a little *below* $e^{-\lambda}$, by a relative amount of about $\lambda^2/2n$.

With this limit in hand, we can take the binomial to the Poisson. Put $p = \lambda/n$ into the binomial pmf (2.4), so that the mean $np = \lambda$ stays fixed, and hold the count k fixed while n grows:

$$\begin{aligned}
\binom{n}{k}p^k(1-p)^{n-k} &\overset{(1)}{=} \frac{n!}{k!\,(n-k)!}\,\frac{\lambda^k}{n^k}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-k} \\
&\overset{(2)}{=} \frac{\lambda^k}{k!}\cdot\frac{n(n-1)\cdots(n-k+1)}{n^k}\cdot\left(1-\frac{\lambda}{n}\right)^n\cdot\left(1-\frac{\lambda}{n}\right)^{-k} \\
&\overset{(3)}{=} \frac{\lambda^k}{k!}\cdot\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)\cdot\left(1-\frac{\lambda}{n}\right)^n\cdot\left(1-\frac{\lambda}{n}\right)^{-k} \\
&\xrightarrow{n\to\infty} \frac{\lambda^k}{k!}\cdot 1\cdot e^{-\lambda}\cdot 1. \qquad (2.11)
\end{aligned}$$

(1) Substitute $p = \lambda/n$ and split $(1-p)^{n-k} = (1-p)^n(1-p)^{-k}$.

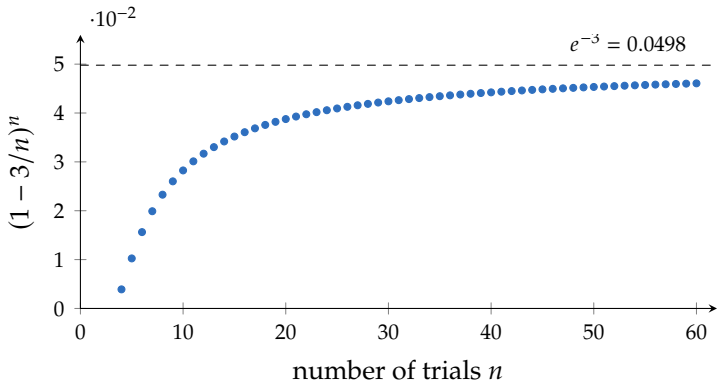


Figure 2.8: The probability that none of n nuclei decays, when each decays with probability $3/n$. With four nuclei it is tiny; as the nuclei become more numerous and individually less likely to decay, it climbs towards the dashed limit e^{-3}, getting within ten per cent of it by $n \approx 50$.

(2) Regroup. $n!/(n-k)! = n(n-1)\cdots(n-k+1)$ has exactly k factors.
(3) Divide each of the k factors by one of the k powers of n.

Now take the four factors in the third line of the chain one at a time as $n \to \infty$. The first, $\lambda^k/k!$, does not involve n. In the second, each of the $k-1$ brackets $(1-j/n)$ goes to 1; there is a *fixed* number of them because k is fixed, so their product goes to 1. The third, $(1-\lambda/n)^n$, goes to $e^{-\lambda}$ by the compound-interest limit. The fourth is a fixed power, $-k$, of something that goes to one, so it goes to one.

Working definition 2.6 (Poisson distribution)

$X \sim \text{Poisson}(\lambda)$ if

$$f(k \mid \lambda) \;=\; e^{-\lambda}\,\frac{\lambda^k}{k!}, \qquad k = 0, 1, 2, \ldots \tag{2.12}$$

for some $\lambda > 0$ [the mean number of events, a real number, not necessarily an integer]. There is no upper limit on k. *Recognise it by:* a count of events that are individually rare and mutually independent, out of a very large (or unknown) number of opportunities.

The number of trials n has disappeared. The only parameter left is the mean count λ, the one thing we can measure.

2.4.2 Normalisation, mean, variance, mgf

Every sum here is the exponential series $e^x = \sum_{k=0}^{\infty} x^k/k!$, valid for every real (or complex) x, in disguise.

$$\sum_{k=0}^{\infty} f(k) \overset{(1)}{=} e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \overset{(2)}{=} e^{-\lambda} e^{\lambda} = 1,$$

$$\mathbb{E}\,X \overset{(3)}{=} e^{-\lambda} \sum_{k=1}^{\infty} k \frac{\lambda^k}{k!} \overset{(4)}{=} \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \overset{(2)}{=} \lambda e^{-\lambda} e^{\lambda} = \lambda,$$

$$\mathbb{E}[X(X-1)] \overset{(5)}{=} e^{-\lambda} \sum_{k=2}^{\infty} k(k-1) \frac{\lambda^k}{k!} = \lambda^2 e^{-\lambda} \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} \overset{(2)}{=} \lambda^2,$$

$$\text{Var}\,X \overset{(6)}{=} \mathbb{E}[X(X-1)] + \mathbb{E}\,X - (\mathbb{E}\,X)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

(1) Take the constant $e^{-\lambda}$ out of the sum.

(2) Exponential series (after renaming $j = k-1$ or $j = k-2$).
(3) The $k = 0$ term vanishes.
(4) $k/k! = 1/(k-1)!$, and one factor λ taken out.
(5) The same trick with $k(k-1)/k! = 1/(k-2)!$; the $k = 0, 1$ terms vanish.
(6) As for the binomial.

Key idea: mean equals variance

For a Poisson count, $\mathbb{E}\,X = \operatorname{Var} X = \lambda$. The standard deviation is $\sqrt{\lambda}$, and the *relative* fluctuation is $1/\sqrt{\lambda}$.

So a Poisson count has mean and variance both equal to λ. Its standard deviation is $\sqrt{\lambda}$, and its *relative* fluctuation, standard deviation over mean, is $1/\sqrt{\lambda}$. This is the origin of the "$\sqrt{N}$" error bar on a count, and of photon shot noise: to measure a flux to 1% we need $1/\sqrt{\lambda} = 0.01$, that is, about 10^4 photons. The same results follow from the binomial ones by taking the limit: $np = \lambda$ and $np(1-p) = \lambda(1-\lambda/n) \to \lambda$.

The mgf is

$$M_X(t) \overset{(1)}{=} e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^t)^k}{k!} \overset{(2)}{=} e^{-\lambda} e^{\lambda e^t} = \exp\bigl[\lambda(e^t - 1)\bigr]. \tag{2.13}$$

(1) $e^{tk}\lambda^k = (\lambda e^t)^k$.
(2) Exponential series with $x = \lambda e^t$.

It is also the limit of the binomial mgf: $(1 + p(e^t - 1))^n = (1 + \lambda(e^t - 1)/n)^n \to e^{\lambda(e^t-1)}$ by the same compound-interest limit.

The shape: a recursion. Dividing consecutive terms of (2.12),

$$\frac{f(k)}{f(k-1)} = \frac{e^{-\lambda}\lambda^k/k!}{e^{-\lambda}\lambda^{k-1}/(k-1)!} = \frac{\lambda}{k}. \tag{2.14}$$

Each step from $k-1$ to k multiplies the probability by λ/k. That factor is larger than one while $k < \lambda$ and smaller than one once $k > \lambda$. So the pmf rises while $k < \lambda$ and falls once $k > \lambda$. The most probable count is $\lfloor \lambda \rfloor$ (and if λ is an integer, $\lambda - 1$ and λ are equally probable). The recursion is also the numerically stable way to compute the pmf: start at $f(0) = e^{-\lambda}$ and multiply by λ/k, never forming λ^k or $k!$ separately [CasellaBerger (3.2.6)–(3.2.7), p. 77].

2.4.3 How good is the approximation?

A convenient single number for the distance between two distributions f and g on the integers is the **total variation distance** $\mathrm{TV}(f, g) = \frac{1}{2}\sum_k |f(k) - g(k)|$ [the largest possible difference between the probabilities the two distributions give to the same event]. For the binomial with $p = \lambda/n$ and the Poisson with the same mean, a classical inequality due to Le Cam states

$$\mathrm{TV}\bigl(\text{Binomial}(n, \lambda/n),\ \text{Poisson}(\lambda)\bigr) \ \le\ np^2 \ =\ \frac{\lambda^2}{n}.$$

Why np^2? Each Bernoulli trial differs from a Poisson(p) variable only in the rare event of two or more "successes" in one trial, which has probability of order p^2. There are n trials, so the differences add up to at most about np^2. The script `code/ch02/04_poisson_limit.py` computes the actual distance for $\lambda = 3$: 0.2191 at $n = 5$, 0.0402 at $n = 20$, 0.007607 at $n = 100$ and 7.515×10^{-4} at $n = 1000$. It falls like $1/n$, as the bound says, but is about ten times smaller than the bound.

2.4.4 Worked examples

Example: a typesetter's errors, exact and Poisson. A typesetter makes on average one error per 500 words. Five pages of 300 words hold 1500 words. What is the probability of at most two errors? Exactly, $X \sim \text{Binomial}(1500, 1/500)$; approximately, $X \sim \text{Poisson}(3)$.

$$\mathrm{P}(X \le 2) \approx e^{-3}\left(1 + 3 + \frac{3^2}{2}\right) \overset{(1)}{=} 0.049787 \times 8.5 = 0.4232.$$

(1) Sum of (2.12) for $k = 0, 1, 2$ with $\lambda = np = 1500/500 = 3$.

The exact binomial gives 0.423; the Poisson approximation gives 0.4232, wrong in the fourth decimal [CasellaBerger Ex. 3.2.5, p. 77].

Example: photomultiplier dark counts. A photomultiplier produces on average 5 dark pulses every 3 ms. In the next 1 ms, what is the probability of no dark pulse, and of at least two? The count in 1 ms is Poisson with $\lambda = 5/3$.

$$\mathrm{P}(X = 0) = e^{-5/3} = 0.189,$$

$$\mathrm{P}(X \ge 2) \overset{(1)}{=} 1 - \mathrm{P}(X = 0) - \mathrm{P}(X = 1) \overset{(2)}{=} 1 - 0.189 - \tfrac{5}{3}(0.189) = 1 - 0.189 - 0.315 = 0.496.$$

(1) Complement of "zero or one".
(2) $\mathrm{P}(X = 1) = \lambda e^{-\lambda}$, which is λ times $\mathrm{P}(X = 0)$ by the recursion (2.14).

So a 1 ms coincidence window contains at least two dark pulses half the time: at this rate, a two-fold coincidence of dark noise is not rare. The numbers are those of a telephone switchboard in [CasellaBerger Ex. 3.2.4, p. 76], a reminder that the Poisson story does not care what is being counted.

Example: events in a collider signal region. A new process has cross-section $\sigma = 0.5$ fb, the data set has integrated luminosity $L = 20$ fb^{-1} and the selection efficiency is $\epsilon = 0.3$. Each of the enormous number of proton–proton collisions has a tiny, independent chance to produce a selected signal event, so the number of selected events is Poisson with mean [Cowan (2.12), p. 30]

$$\lambda = \sigma L \epsilon = 0.5 \times 20 \times 0.3 = 3.$$

The chance of seeing *nothing* is $f(0\,|\,3) = e^{-3} = 0.0498$, about 5%. Read backwards, this is the classic "three events" rule: if we see zero events, any model predicting $\lambda \ge 3$ is excluded at about 95% confidence (Chapter 4).

2.4.5 Python: literally many nuclei

Listing 2.4: Counting decays of 2000 nuclei per window (`code/ch02/04_poisson_limit.py`)

```
# (a) the generating process, literally: many nuclei, each with a tiny chance
n_nuclei, windows = 2000, 20_000
counts = np.empty(windows, dtype=int)
for w in range(windows):
    counts[w] = np.count_nonzero(rng.uniform(size=n_nuclei) <= lam / n_nuclei)
```

Each counting window gets 2000 fresh uniform numbers, one per nucleus, and a nucleus "decays" if its number is below $\lambda/n = 3/2000$. The count per window is the number of decays. Nothing in this code mentions $e^{-\lambda}$ or $k!$: the Poisson law has to emerge from the process. The script then computes the total variation distances quoted above and the typesetter example.

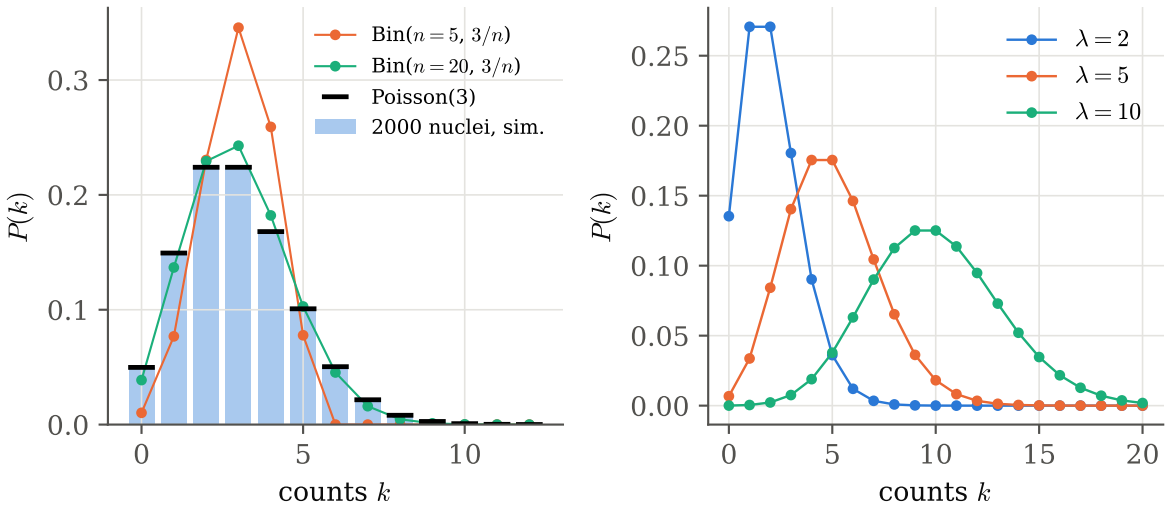


Figure 2.9: Left: the histogram of decay counts from 2000 simulated nuclei per window (bars) sits on the Poisson(3) pmf (black ticks). The binomials with only 5 or 20 trials and the same mean are visibly narrower, because each trial can produce at most one event. Right: Poisson shapes for $\lambda = 2, 5, 10$; skewed to the right at small λ, nearly symmetric by $\lambda = 10$.

Over 20 000 windows the sample mean is 3 and the sample variance is 2.986, equal within the statistical error as *mean = variance* demands. The fraction of empty windows is 0.04965, against $e^{-3} = 0.04979$ (Figure 2.9). The right panel reproduces the shapes of [Cowan Fig. 2.3, p. 29]. The growing symmetry with λ is the start of the Gaussian limit, previewed in Section 2.8.

What this bought us. When many independent chances each have a small probability, the count is Poisson with the single parameter $\lambda = np$. Its variance equals its mean, so the relative scatter of a count is $1/\sqrt{\lambda}$. The approximation error is of order λ^2/n, negligible for any macroscopic number of trials.

Watch out: "rare" refers to each trial, not to the count

The count itself need not be small: a CCD pixel may collect 10^5 photons, and that count is still Poisson, because each of the 10^{20} photons emitted by the star had a tiny chance of landing in this pixel. What must be small is the probability per opportunity, and what must hold is independence. Conversely, a small count is not automatically Poisson: the number of dead channels in a 4-channel module is binomial with $n = 4$, and Figure 2.9 shows how different that can be.

2.5 The Poisson process: constant rate, no memory

In Section 2.4 the Poisson distribution came from a binomial over n nuclei, each with a tiny chance to decay. Often there is no such n to imagine. Cosmic rays arrive from the sky, photons from a distant quasar, dark pulses from a photocathode. All we know is that events come *at random in time*, with some average rate λ (events per second), and that the process has no memory: what happened in the last second does not affect the next. These two physical statements are enough. Through a differential equation, they force the count in any window of length t to be Poisson(λt). The same machinery then shows that three common operations give Poisson counts again: thinning a Poisson stream (losing events at random), adding two streams, and splitting one into histogram bins.

Physical picture: a master equation with one-way hops. Think of the count $N(t)$, the number of events up to time t, as the state of a system that can only climb a ladder: $0 \to 1 \to 2 \to \cdots$, one rung at a time, with hopping rate λ from every rung. The probability of being on rung n is fed from rung $n-1$ and drains into rung $n+1$. The rate equations for a radioactive decay chain have exactly this form, and so does the master equation of a birth process in statistical mechanics.

The question, step by step.

1. What exactly do "random in time at rate λ" and "no memory" assume? State the inputs, then derive what they imply for a very short interval h.
2. How does the probability of having n events by time t change between t and $t+h$?
3. Let $h \to 0$ to get differential equations, and solve them.
4. What happens to such a stream when we lose some events, add another stream, or sort the events into bins?

2.5.1 From three physical inputs to the short-interval rules

Think of photons from a star whose brightness does not change. Even a perfect detector does not count the same number in every exposure, because the photons arrive one at a time at random instants. That scatter is **shot noise**. It is not interference added to the signal. It is what counting discrete, independent arrivals means. We now find its distribution from first principles: we say exactly what we assume about the arrivals, and derive everything else.

The inputs. Three are physical, and one is a mathematical regularity condition.

(P1) *No memory.* The numbers of events in non-overlapping time windows are independent. An arrival tells us nothing about when the next one comes.

(P2) *Constant rate.* The distribution of the count in a window depends only on its length, not on when it starts. The source does not brighten or fade during the exposure.

(P3) *One at a time.* Events arrive singly, never in bunches at the same instant.

(R) *Regularity.* The probability $P_0(h)$ that a window of length h is empty changes continuously with h. This is a property we demand of the mathematics, not a statement about photons.

The empty window. Split a window of length $h_1 + h_2$ into two pieces. It is empty exactly when both pieces are empty, so

$$P_0(h_1+h_2) \overset{(1)}{=} \text{P(first piece empty)}\ \text{P(second piece empty)} \overset{(2)}{=} P_0(h_1)\,P_0(h_2). \tag{2.15}$$

(1) (P1): the two pieces do not overlap, so their counts are independent and the probabilities multiply.

(2) (P2): each piece behaves like any window of the same length, wherever it sits.

Write $g(h) = -\ln P_0(h)$. Then (2.15) says $g(h_1+h_2) = g(h_1) + g(h_2)$: g turns sums of lengths into sums of values. Applying it n times gives $g(n\,h) = n\,g(h)$, so $g(m/n) = (m/n)\,g(1)$ for every rational length, and by (R) the same holds for every length. Hence, with $\lambda \equiv g(1) \geq 0$,

$$P_0(h) = e^{-\lambda h}. \tag{2.16}$$

Nothing was assumed about the size of λ; it is simply the constant that the inputs leave free, and below it turns out to be the mean number of events per unit time, the **rate**.

At least one event in a short window. Expanding (2.16) for small h,

$$\mathrm{P}(\text{at least one event in } h) = 1 - e^{-\lambda h} \overset{(1)}{=} \lambda h - \tfrac{1}{2}(\lambda h)^2 + \cdots = \lambda h + O(h^2).$$

(1) The Taylor series $e^{-x} = 1 - x + \frac{1}{2}x^2 - \cdots$.

So "probability proportional to the length" is not an assumption. It is the first term of the exact answer, and it holds for any short window.

Two or more in a short window. By (P3) two events never share an instant, so two arrivals in one window are two separate events. Cut the window into two halves: two events in the window means at least one event in each half, or two in one half. The first has probability $\left(\lambda\frac{h}{2} + O(h^2)\right)^2 = O(h^2)$ by (P1), and repeating the halving shows the second is also $O(h^2)$. So

$$\mathrm{P}(\text{two or more in } h) = O(h^2), \qquad \mathrm{P}(\text{exactly one in } h) = \lambda h + O(h^2).$$

Two arrivals need two rare events to coincide, a product of two small probabilities. That is why, in a short enough window, the only outcomes that matter are "nothing" and "exactly one".

We need one piece of notation for what follows. A function $g(h)$ is $o(h)$ ("little-o of h") if $g(h)/h \to 0$ as $h \to 0$ [it vanishes faster than h itself]. For example $h^2 = o(h)$, but $\sqrt{h}$ and $3h$ are not. Every $O(h^2)$ above is $o(h)$.

Working definition 2.7 (Poisson process)

A **counting process** $N(t)$ is the number of events in $[0, t]$. Inputs (P1)–(P3) and (R) give it the five properties of a **homogeneous Poisson process** with rate λ:

(i) $N(0) = 0$ [we start counting at $t = 0$];

(ii) the counts in non-overlapping intervals are independent [input (P1)];

(iii) the distribution of the count in $(t, t + h]$ depends only on the length h [input (P2)];

(iv) $\mathrm{P}(\text{exactly one event in } (t, t + h]) = \lambda h + o(h)$ [derived above, not assumed];

(v) $\mathrm{P}(\text{two or more events in } (t, t + h]) = o(h)$ [derived above from (P3)].

Recognise it by: events that arrive one at a time, independently, at a steady average rate.

Watch out: When the inputs fail

Each input can fail, and then the counts are not Poisson. A source whose brightness changes during the exposure breaks (P2) (the count is then Poisson with mean $\int \lambda(t)\,\mathrm{d}t$, see below). Dead time after each detection breaks (P1): a detector that is blind for a while after a hit makes the next arrival depend on the last. Events that come in bursts break (P3): an air shower fires many detectors at once, and thermal light shows photon bunching. In each case the variance of the count differs from its mean, which is the quickest test.

Figure 2.10 already suggests the answer. Cut $[0, t]$ into $n = t/h$ slices. Each slice is a Bernoulli trial with success probability $p = \lambda h = \lambda t/n$, and the trials are independent by (ii). So $N(t)$ is binomial with $np = \lambda t$, and as $h \to 0$ we are in the limit of Section 2.4: $N(t) \sim \mathrm{Poisson}(\lambda t)$. Now we derive the same result a second way, which does not rely on the binomial at all and which generalises to rates that vary in time.

2.5.2 The differential equations

Let $P_n(t) = \mathrm{P}(N(t) = n)$. At $t = 0$, $P_0(0) = 1$ and $P_n(0) = 0$ for $n \geq 1$, by property (i).

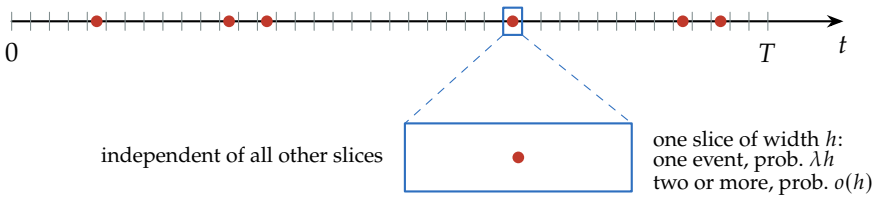


Figure 2.10: The short-interval rules as a picture. The time axis is cut into short slices; each slice independently contains an event (red dots) with probability λh, and practically never two. Counting the dots in $[0, T]$ means counting successes in T/h Bernoulli trials with $p = \lambda h$.

No events. To have no events by $t + h$ we need no events by t *and* none in $(t, t + h]$:

$$P_0(t+h) \overset{(1)}{=} P_0(t)\ \mathrm{P}(\text{no event in } (t, t+h]) \overset{(2)}{=} P_0(t)\big(1 - \lambda h - o(h)\big),$$
$$\frac{P_0(t+h) - P_0(t)}{h} \overset{(3)}{=} -\lambda P_0(t) - P_0(t)\frac{o(h)}{h} \xrightarrow{h\to 0} \frac{\mathrm{d}P_0}{\mathrm{d}t} = -\lambda P_0.$$

(1) Property (ii): the count in $(t, t + h]$ is independent of the count in $[0, t]$, so the probabilities multiply.
(2) Properties (iv) and (v): the probability of zero events is 1 minus the probability of one minus the probability of two or more, and both $o(h)$ terms combine into one $o(h)$.
(3) Subtract $P_0(t)$ and divide by h. The last term vanishes as $h \to 0$ by the definition of $o(h)$.

With $P_0(0) = 1$ the solution is $P_0(t) = e^{-\lambda t}$: the probability of an empty window decays exponentially with its length, just as the survival probability of an unstable nucleus does.

$n \geq 1$ events. To have n events by $t + h$, either we had n by t and none arrive, or we had $n - 1$ and one arrives, or we had $n - j$ for $j \geq 2$ and j arrive:

$$P_n(t+h) \overset{(1)}{=} P_n(t)\big(1 - \lambda h\big) + P_{n-1}(t)\,\lambda h + o(h),$$
$$\frac{\mathrm{d}P_n}{\mathrm{d}t} \overset{(2)}{=} -\lambda P_n(t) + \lambda P_{n-1}(t).$$

(1) The three routes are mutually exclusive, so their probabilities add, and each is a product by property (ii). All routes with two or more arrivals, and the corrections to the first two, are $o(h)$ by (iv) and (v).
(2) Subtract $P_n(t)$, divide by h and let $h \to 0$.

This is the **master equation** of the process (Figure 2.11): rung n gains probability from rung $n - 1$ at rate λP_{n-1} and loses it to rung $n + 1$ at rate λP_n.

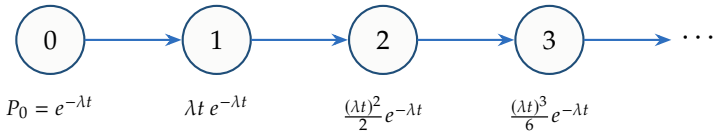


Figure 2.11: The count as a system climbing a ladder. Every hop goes one rung up and happens at the same rate λ, whatever the rung; no hop ever goes down. Below each rung is its occupation probability at time t, the solution derived in the text.

Solving the ladder. The master equation has the form $y' + \lambda y = g(t)$, a linear first-order equation, with $y = P_n$ and $g = \lambda P_{n-1}$. Such an equation is solved with the *integrating factor* $e^{\lambda t}$.

Multiplying by it turns the left side into a single derivative: $(e^{\lambda t}y)' = e^{\lambda t}(y' + \lambda y) = e^{\lambda t}g$. So define $Q_n(t) = e^{\lambda t}P_n(t)$. Then

$$\frac{\mathrm{d}Q_n}{\mathrm{d}t} \overset{(1)}{=} e^{\lambda t}\left(\frac{\mathrm{d}P_n}{\mathrm{d}t} + \lambda P_n\right) \overset{(2)}{=} e^{\lambda t}\lambda P_{n-1} = \lambda Q_{n-1}, \qquad Q_0(t) = 1, \quad Q_n(0) = 0\ (n \geq 1).$$

(1) Product rule.
(2) The master equation: $\mathrm{d}P_n/\mathrm{d}t + \lambda P_n = \lambda P_{n-1}$.

Now climb the ladder by induction. Suppose $Q_{n-1}(t) = (\lambda t)^{n-1}/(n-1)!$. This is true for $n - 1 = 0$, since $Q_0 = e^{\lambda t}P_0 = e^{\lambda t}e^{-\lambda t} = 1$. Then

$$Q_n(t) \overset{(1)}{=} \int_0^t \lambda\, Q_{n-1}(s)\,\mathrm{d}s \overset{(2)}{=} \int_0^t \frac{\lambda^n s^{n-1}}{(n-1)!}\,\mathrm{d}s \overset{(3)}{=} \frac{\lambda^n t^n}{n!}.$$

(1) Integrate $\mathrm{d}Q_n/\mathrm{d}s = \lambda Q_{n-1}$ from 0 to t, using $Q_n(0) = 0$.
(2) The induction hypothesis.
(3) $\int_0^t s^{n-1}\mathrm{d}s = t^n/n$, and $(n-1)!\,n = n!$.

Key idea: the count of a Poisson process

For a homogeneous Poisson process with rate λ,

$$\mathrm{P}\big(N(t) = n\big) = e^{-\lambda t}\frac{(\lambda t)^n}{n!}, \qquad N(t) \sim \mathrm{Poisson}(\lambda t), \tag{2.17}$$

and by (iii) the count in *any* window of length t has the same distribution.

Undoing the substitution, $P_n(t) = e^{-\lambda t}Q_n(t) = e^{-\lambda t}(\lambda t)^n/n!$: the count $N(t)$ is Poisson with mean λt, as in (2.17). The two routes, the binomial limit and the master equation, give the same answer. So whenever events are rare, independent and at a steady rate, the Poisson distribution is forced on us.

A rate that varies in time. Suppose the rate is a function $\lambda(t)$: a source that fades, or a beam whose intensity falls during a fill. Then input (P2), and with it property (iii), is dropped, and (iv) becomes $\lambda(t)h + o(h)$. The master equation keeps its form with λ replaced by $\lambda(t)$. The integrating factor becomes $e^{m(t)}$, where $m(t) = \int_0^t \lambda(s)\,\mathrm{d}s$ is the expected number of events up to t. Because $\mathrm{d}m/\mathrm{d}t = \lambda(t)$, the same induction gives $Q_n = m(t)^n/n!$. So

$$N(t) \sim \mathrm{Poisson}\big(m(t)\big), \qquad N(s+t) - N(s) \sim \mathrm{Poisson}\big(m(s+t) - m(s)\big).$$

Only the *integrated* rate over the window matters [Wasserman Thm. 23.33, p. 395]. For a collider, m is $\sigma\epsilon$ times the integrated luminosity, whatever the luminosity did during the run.

2.5.3 Three closure properties

Real counting chains lose events, add sources and sort events into bins. Each of these operations keeps us inside the Poisson family.

Adding independent sources: sums of independent Poissons are Poisson. If $X_1 \sim \mathrm{Poisson}(\lambda_1)$ and $X_2 \sim \mathrm{Poisson}(\lambda_2)$ are independent, then $X_1 + X_2 \sim \mathrm{Poisson}(\lambda_1 + \lambda_2)$. The direct proof sums

over all the ways to split k events between the two sources (a *convolution*):

$$\mathrm{P}(X_1 + X_2 = k) \overset{(1)}{=} \sum_{j=0}^{k} \mathrm{P}(X_1 = j)\ \mathrm{P}(X_2 = k - j) \overset{(2)}{=} \sum_{j=0}^{k} e^{-\lambda_1} \frac{\lambda_1^j}{j!}\, e^{-\lambda_2} \frac{\lambda_2^{k-j}}{(k-j)!}$$
$$\overset{(3)}{=} \frac{e^{-(\lambda_1+\lambda_2)}}{k!} \sum_{j=0}^{k} \binom{k}{j} \lambda_1^j \lambda_2^{k-j} \overset{(4)}{=} e^{-(\lambda_1+\lambda_2)} \frac{(\lambda_1+\lambda_2)^k}{k!}.$$

(1) The event $\{X_1 + X_2 = k\}$ is the disjoint union of $\{X_1 = j, X_2 = k - j\}$, and independence factorises each term.
(2) Poisson pmfs.
(3) Multiply and divide by $k!$ to form $\binom{k}{j} = k!/(j!(k-j)!)$.
(4) Binomial theorem.

With mgfs it is one line: $e^{\lambda_1(e^t-1)} e^{\lambda_2(e^t-1)} = e^{(\lambda_1+\lambda_2)(e^t-1)}$ [Wasserman Ex. 3.35, p. 58][2]. In the process picture the result is obvious. Interleave two independent streams, and each short slice holds one event with probability $(\lambda_1 + \lambda_2)h + o(h)$: the merged stream is a Poisson process with the summed rate. So signal plus background, with expected counts s and b, is Poisson with mean $s + b$.

Losing events: thinning. Each event of a Poisson(λ) stream is detected independently with efficiency ϵ. Given $N = n$ true events, the detected number is $Y \mid N = n \sim \mathrm{Binomial}(n, \epsilon)$. Marginalise over the unobserved N:

$$\mathrm{P}(Y = y) \overset{(1)}{=} \sum_{n=y}^{\infty} \mathrm{P}(Y = y \mid N = n)\, \mathrm{P}(N = n) \overset{(2)}{=} \sum_{n=y}^{\infty} \frac{n!}{y!\,(n-y)!} \epsilon^y (1-\epsilon)^{n-y}\, e^{-\lambda} \frac{\lambda^n}{n!}$$
$$\overset{(3)}{=} \frac{(\epsilon\lambda)^y}{y!} e^{-\lambda} \sum_{j=0}^{\infty} \frac{\big((1-\epsilon)\lambda\big)^j}{j!} \overset{(4)}{=} \frac{(\epsilon\lambda)^y}{y!} e^{-\lambda} e^{(1-\epsilon)\lambda} = e^{-\epsilon\lambda} \frac{(\epsilon\lambda)^y}{y!}.$$

(1) Law of total probability over the nuisance variable N (Chapter 1); $n < y$ is impossible.
(2) Binomial and Poisson pmfs.
(3) Cancel $n!$, write $\lambda^n = \lambda^y \lambda^{n-y}$, and rename $j = n - y$.
(4) Exponential series.

So $Y \sim \mathrm{Poisson}(\epsilon\lambda)$ [Wasserman p. 395]. A detector with efficiency ϵ looking at a Poisson source records a Poisson count with mean $\epsilon\lambda$. This is why $\lambda = \sigma L \epsilon$ in the collider example of Section 2.4 could simply include the efficiency as a factor.

Sorting into bins: Poissonisation. Now let a Poisson(ν) number of events be sorted into m histogram bins with probabilities p_j (Section 2.3). Given the total $N = \sum n_j$, the bins are multinomial. The joint pmf of the bins is

$$\mathrm{P}(n_1, \ldots, n_m) \overset{(1)}{=} \mathrm{P}(N = \textstyle\sum_j n_j)\ \frac{N!}{n_1! \cdots n_m!} p_1^{n_1} \cdots p_m^{n_m} \overset{(2)}{=} e^{-\nu} \frac{\nu^N}{N!} \frac{N!}{\prod_j n_j!} \prod_j p_j^{n_j}$$
$$\overset{(3)}{=} \prod_{j=1}^{m} e^{-\nu p_j} \frac{(\nu p_j)^{n_j}}{n_j!}.$$

(1) The counts determine N, so the joint probability is $\mathrm{P}(N)$ times the multinomial given N.

[2] The line "$Y = Y_1 + Y + 2$" in the source is a misprint for $Y = Y_1 + Y_2$.

(2) Poisson pmf of N.
(3) Cancel $N!$; write $\nu^N = \prod_j \nu^{n_j}$ and $e^{-\nu} = \prod_j e^{-\nu p_j}$ using $\sum_j p_j = 1$.

The joint pmf *factorises* into a product of Poisson(νp_j) pmfs, one per bin. So the bins are *independent* Poisson counts. This is the proof promised in Section 2.3: the anti-correlation came entirely from fixing the total. It is also why a binned likelihood with a Poisson term per bin, $\prod_j e^{-\nu_j} \nu_j^{n_j}/n_j!$, is the standard model of a collider histogram (Chapter 3).

2.5.4 Worked examples

Example: an empty window. A counter sees background at $\lambda = 2$ counts/s. What is the probability that a 5 s window is empty, and that a 1 s window is? By (2.17), $\mathrm{P}(N(5) = 0) = e^{-10} = 4.540 \times 10^{-5}$ and $\mathrm{P}(N(1) = 0) = e^{-2} = 0.135$. The five-second window is five one-second windows in a row, and by independence (input P1) $e^{-10} = (e^{-2})^5$: the empty-window probabilities multiply, which is precisely why P_0 had to be an exponential.

Example: signal on top of background. A search window expects $b = 2.4$ background and $s = 1.2$ signal events, both Poisson and independent. The total count is Poisson(3.6). The probability of observing at least 6 events is

$$\mathrm{P}(N \geq 6) \overset{(1)}{=} 1 - e^{-3.6} \sum_{k=0}^{5} \frac{3.6^k}{k!} \overset{(2)}{=} 1 - 0.02732\,(1 + 3.6 + 6.48 + 7.776 + 6.998 + 5.039)$$
$$= 1 - 0.02732 \times 30.893 = 1 - 0.8441 = 0.156,$$

(1) Sum of Poissons is Poisson; complement of $k \leq 5$.
(2) $e^{-3.6} = 0.02732$, and the terms $3.6^k/k!$ built with the recursion (2.14): multiply successively by $3.6/1, 3.6/2, \ldots, 3.6/5$.

while with background alone ($\lambda = 2.4$) the same calculation gives $1 - 0.9643 = 0.036$. Six events would be four times more likely with the signal than without: the *ratio* of these two probabilities is the germ of the likelihood ratio of Chapter 4.

Example: a detector that sees 30% of the events. A source emits into the detector at $\lambda = 2\ \mathrm{s}^{-1}$ and each photon is detected with $\epsilon = 0.3$. In $T = 5$ s the true count is Poisson(10) and the detected count is Poisson(3). The detected count is *not* "0.3 times a Poisson(10) count". That scaled count could only take the values $0, 0.3, 0.6, \ldots$, and its variance would be $0.3^2 \times 10 = 0.9$. Thinning keeps the counts integer, and the thinned count has variance equal to its mean, 3. The extra variance comes from the losses: each event is lost at random, and the random losses add binomial noise.

2.5.5 Python: the short-interval rules, slice by slice

The script `05_poisson_process.py` implements Figure 2.10 literally: $T = 5$ s is cut into 5000 slices of $h = 1$ ms, and each slice holds an event with probability $\lambda h = 0.002$.

Listing 2.5: Events slice by slice, thinned, and a second source (`code/ch02/05_poisson_process.py`)

```
N = np.empty(windows, int); Ndet = np.empty(windows, int); Nsum = np.empty(windows, int)
for s in range(0, windows, chunk):
    ev = rng.uniform(size=(chunk, slices)) < lam * h          # the postulate, slice by
    slice
```

```
31      det = ev & (rng.uniform(size=(chunk, slices)) < eps)      # each event survives with
        prob eps
32      ev2 = rng.uniform(size=(chunk, slices)) < lam2 * h        # an independent second source
33      N[s:s + chunk] = ev.sum(1)
34      Ndet[s:s + chunk] = det.sum(1)
35      Nsum[s:s + chunk] = ev.sum(1) + ev2.sum(1)
```

Line 30 is property (iv): one uniform number per slice, an event if it is below λh. Line 31 is thinning: each event survives only if a second, independent uniform number is below ϵ. Line 32 is an independent second source at 1 s^{-1}. The chunks of 1000 windows keep the memory use small. A second block integrates the master equation with Euler steps of 1 ms:

Listing 2.6: Euler integration of the master equation (`code/ch02/05_poisson_process.py`)

```
37  # Euler integration of the master equation, P_n(0) = delta_{n0}
38  dt, tmax, nmax = 1e-3, 4.0, 5
39  t = np.arange(0, tmax + dt / 2, dt)
40  P = np.zeros((t.size, nmax)); P[0, 0] = 1.0
41  for i in range(1, t.size):
42      prev = P[i - 1]
43      gain = np.concatenate([[0.0], prev[:-1]])          # P_{n-1}, with P_{-1} = 0
44      P[i] = prev + dt * lam * (gain - prev)
```

`gain` is the vector of P_{n-1} (with $P_{-1} = 0$), and line 44 is $P_n(t + \mathrm{d}t) = P_n(t) + \mathrm{d}t\,\lambda(P_{n-1} - P_n)$.

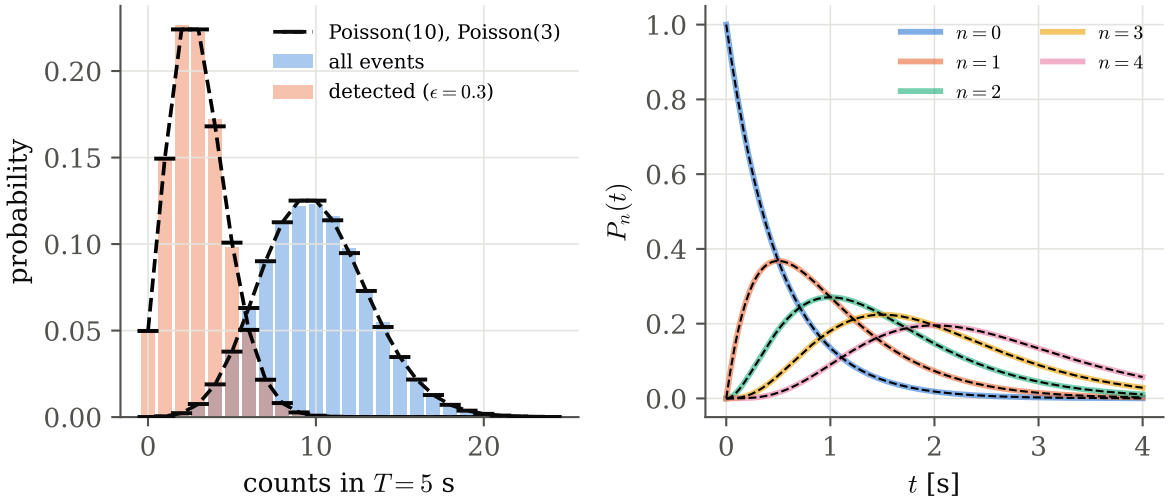


Figure 2.12: Left: counts in 5 s windows built slice by slice, for all events (blue) and for the 30% that are detected (orange); the black ticks are Poisson(10) and Poisson(3). Right: the ladder probabilities $P_0, \ldots, P_4$ from integrating the master equation (thick coloured curves) lie on the analytic Poisson curves (thin dashed). P_0 decays from one; each higher rung fills up later and peaks at $t = n/\lambda$.

Over 20 000 windows the total count has mean 10.01 and variance 9.866 (theory 10 and 10), the thinned count 3.004 and 3.008 (theory 3 and 3), and the sum of the two sources 15 and 14.76 (theory 15 and 15). Empty windows are rare: with $e^{-10} = 4.540 \times 10^{-5}$ we expect $20\,000\,e^{-10} \approx 0.9$ of them, and this run has 0. The Euler solution of the master equation differs from $e^{-\lambda t}(\lambda t)^n/n!$ by at most 4.619×10^{-4} over $0 \le t \le 4$ s (Figure 2.12), an error that shrinks in proportion to the step size.

How this is used in real research: verification versus research

Everything above is *verification*: the simulation checks an analytic result we already had. The same slice-by-slice machinery becomes a research instrument as soon as the inputs

(P1)–(P3) are only approximately true and no formula exists: a rate that depends on the recent history (afterpulsing in a PMT, pile-up in a calorimeter), events that trigger other events (cosmic-ray air showers, aftershocks), or a detector whose efficiency depends on the previous hit (Section 2.8). Then we simulate the process and histogram the counts, and that histogram *is* the sampling distribution $P(S \mid \theta)$ fed into the likelihood.

2.6 Waiting times: the exponential and gamma distributions

The same Geiger counter can be read in two ways. We can count the clicks in fixed windows (Section 2.5), or we can record the *time* of each click and look at the gaps between them. A timing detector with a photomultiplier does the second: it waits for the first photon, or for the tenth, and its time resolution is set by how much that waiting time fluctuates. A muon stopped in a scintillator waits a random time before it decays. For a Poisson process these waiting times have simple laws. The time to the first event is exponential. The time to the n-th event is gamma (also called Erlang). And the "no memory" input becomes the statement that a waiting time never "ages".

Physical picture: the survival law of an unstable particle. An unstable nucleus has no clock inside it. If it has survived a thousand years, its chance of decaying in the next second is the same as on the day it was made. This is why the survival fraction is $e^{-t/\tau}$, and why we can speak of *the* lifetime τ of a species without saying how old the sample is. The time between Poisson events is the same law: the process never "remembers" when the last event happened.

The question, step by step.

1. The first event comes after time t exactly when the window $[0, t]$ is empty. We already know the probability of that.
2. The n-th event comes before time w exactly when the window $[0, w]$ holds at least n events. We know that too.
3. Differentiate these probabilities to get densities. Compute means and variances.
4. Check the "no memory" property directly, and simulate a stream from its gaps.

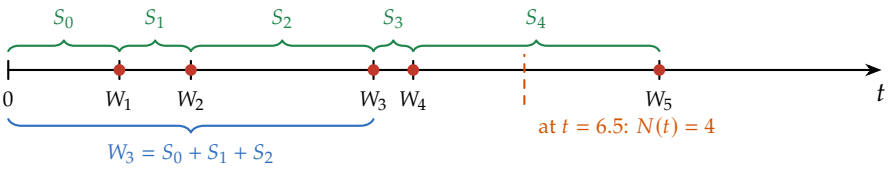


Figure 2.13: One stretch of a Poisson stream. The green braces are the gaps (sojourn times) $S_0, S_1, \ldots$ between successive events; the arrival times W_n are their running sums. At the dashed line four events have happened, so $W_4 \le t < W_5$: counts and waiting times are two descriptions of the same stream.

2.6.1 The first event: exponential

Let $T = W_1$ be the time of the first event, and recall that $N(t)$ is the number of events in $[0, t]$. Figure 2.13 shows the link between the two descriptions: the first event comes after t if and

only if there is no event in $[0, t]$, that is, $T > t$ exactly when $N(t) = 0$. This duality between waiting times and counts turns a question about times into one about counts, which we have already answered. So

$$\mathrm{P}(T > t) \overset{(1)}{=} \mathrm{P}(N(t) = 0) \overset{(2)}{=} e^{-\lambda t}, \qquad F(t) = \mathrm{P}(T \le t) = 1 - e^{-\lambda t}, \qquad f(t) \overset{(3)}{=} \frac{\mathrm{d}F}{\mathrm{d}t} = \lambda e^{-\lambda t}. \quad (2.18)$$

(1) The duality between waiting times and counts.
(2) The empty-window probability $P_0(t)$ from Section 2.5.
(3) A density is the derivative of the CDF (Chapter 1).

Working definition 2.8 (exponential distribution)

$T \sim \mathrm{Exp}(\lambda)$ in the *rate* parameterisation if $f(t) = \lambda e^{-\lambda t}$ for $t \ge 0$ [λ is events per unit time]; equivalently, in the *scale* parameterisation used by AoS, $f(t) = \frac{1}{\beta} e^{-t/\beta}$ with $\beta = 1/\lambda$ [β is a time, the mean wait]. For a decaying particle $\beta = \tau$, the mean lifetime. *Recognise it by:* the waiting time until something that happens at a constant rate and does not age.

The moments of the exponential, and those of the gamma distribution below, need one special function. The **gamma function** is, for $\alpha > 0$, $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y}\,\mathrm{d}y$ [a continuous extension of the factorial, with $\Gamma(n) = (n-1)!$]. Its three essential properties:

$$\Gamma(1) = \int_0^\infty e^{-y}\mathrm{d}y = 1,$$
$$\Gamma(\alpha+1) \overset{(1)}{=} \Big[-y^\alpha e^{-y}\Big]_0^\infty + \alpha \int_0^\infty y^{\alpha-1} e^{-y}\mathrm{d}y \overset{(2)}{=} \alpha\,\Gamma(\alpha),$$
$$\Gamma(\tfrac{1}{2}) \overset{(3)}{=} \int_0^\infty y^{-1/2} e^{-y}\mathrm{d}y \overset{(4)}{=} 2\int_0^\infty e^{-x^2}\mathrm{d}x = \sqrt{\pi}.$$

(1) Integration by parts, with $u = y^\alpha$ and $\mathrm{d}v = e^{-y}\mathrm{d}y$.
(2) The boundary term vanishes at both ends for $\alpha > 0$.
(3) Definition with $\alpha = \frac{1}{2}$.
(4) Substitute $y = x^2$, $\mathrm{d}y = 2x\,\mathrm{d}x$, and use the Gaussian integral $\int_{-\infty}^\infty e^{-x^2}\mathrm{d}x = \sqrt{\pi}$.

From the first two, $\Gamma(n) = (n-1)(n-2)\cdots 1\cdot\Gamma(1) = (n-1)!$ for integer n [CasellaBerger (3.3.2)–(3.3.4), p. 81].

Now the moments of the exponential, substituting $y = \lambda t$:

$$\mathbb{E}\,T^r \overset{(1)}{=} \int_0^\infty t^r \lambda e^{-\lambda t}\mathrm{d}t \overset{(2)}{=} \frac{1}{\lambda^r}\int_0^\infty y^r e^{-y}\mathrm{d}y \overset{(3)}{=} \frac{\Gamma(r+1)}{\lambda^r} = \frac{r!}{\lambda^r}.$$

(1) Definition of the r-th moment.
(2) $t = y/\lambda$, $\mathrm{d}t = \mathrm{d}y/\lambda$.
(3) Definition of $\Gamma(r+1)$.

So $\mathbb{E}\,T = 1/\lambda$, $\mathbb{E}\,T^2 = 2/\lambda^2$ and $\mathrm{Var}\,T = 2/\lambda^2 - 1/\lambda^2 = 1/\lambda^2$. The mean waiting time is the inverse rate, and the standard deviation equals the mean: an exponential wait is as uncertain as it is long.

2.6.2 No memory

Key idea: the exponential is memoryless

For all $t, s \geq 0$, $\mathrm{P}(T > t + s \mid T > t) = \mathrm{P}(T > s)$.

$$\mathrm{P}(T > t+s \mid T > t) \overset{(1)}{=} \frac{\mathrm{P}(T > t+s \text{ and } T > t)}{\mathrm{P}(T > t)} \overset{(2)}{=} \frac{\mathrm{P}(T > t+s)}{\mathrm{P}(T > t)} \overset{(3)}{=} \frac{e^{-\lambda(t+s)}}{e^{-\lambda t}} = e^{-\lambda s} = \mathrm{P}(T > s).$$

(1) Definition of conditional probability (Chapter 1).
(2) $T > t + s$ already implies $T > t$.
(3) Survival function (2.18).

In words: having waited a time t without an event, the remaining wait has the same distribution as a fresh one (Figure 2.14). The converse is also true: the exponential is the *only* continuous distribution with this property. To see why, write $G(t) = \mathrm{P}(T > t)$ for the survival function. Memorylessness says $G(t + s)/G(t) = G(s)$, that is, $G(t + s) = G(t)G(s)$. Taking logarithms, $\ln G(t + s) = \ln G(t) + \ln G(s)$: the function $\ln G$ is additive. The only continuous additive functions are linear, so $\ln G(t) = -\lambda t$ for some constant $\lambda > 0$. So memorylessness *forces* $G(t) = e^{-\lambda t}$ [CasellaBerger (3.3.11), p. 83].

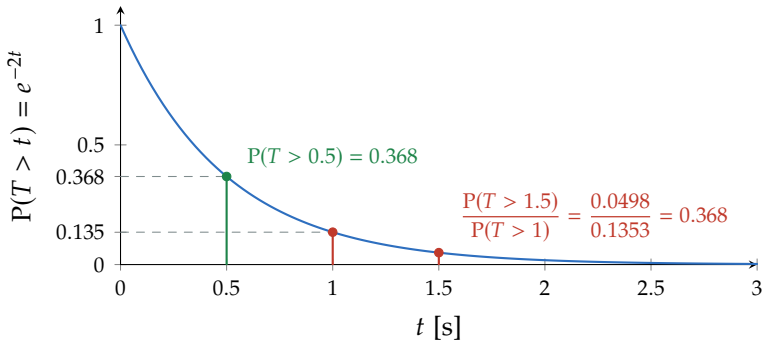


Figure 2.14: The survival curve of an exponential wait with $\lambda = 2\ \mathrm{s}^{-1}$. The chance of waiting a further half second is the same whether we start the clock at zero (green: height at 0.5) or after an empty first second (red: ratio of the heights at 1.5 and 1). Every stretch of the curve is a scaled copy of the whole.

The gaps are independent exponentials. Recall from Figure 2.13 that $S_0, S_1, \ldots$ are the gaps between successive events and W_n is the time of the n-th event. The argument for the first wait restarts at every event. Given that the first event is at $W_1 = s$, the second gap S_1 exceeds t exactly when $(s, s + t]$ is empty:

$$\mathrm{P}(S_1 > t \mid W_1 = s) \overset{(1)}{=} \mathrm{P}(\text{no events in } (s, s+t] \mid W_1 = s) \overset{(2)}{=} \mathrm{P}(\text{no events in } (s, s+t]) \overset{(3)}{=} e^{-\lambda t}.$$

(1) Duality between the gap and an empty interval.
(2) Independent increments: what happens in $(s, s + t]$ does not care about $[0, s]$.
(3) Stationarity: an interval of length t is empty with probability $e^{-\lambda t}$ wherever it sits.

The answer does not depend on s, so S_1 is independent of $W_1 = S_0$ and is again $\mathrm{Exp}(\lambda)$. Repeating the argument, the gaps $S_0, S_1, S_2, \ldots$ are independent and identically distributed (iid) $\mathrm{Exp}(\lambda)$ [Wasserman Thm. 23.35, p. 396]. This gives a second, equivalent definition of the Poisson process: *a stream of events whose gaps are iid exponential*. It is the one we use to simulate.

2.6.3 The n-th event: gamma (Erlang)

The time of the n-th event is $W_n = S_0 + S_1 + \cdots + S_{n-1}$, a sum of n independent exponentials. Use the duality again: $W_n \le w$ exactly when at least n events happened by w.

$$\begin{aligned} \mathrm{P}(W_n \le w) &\overset{(1)}{=} \mathrm{P}(N(w) \ge n) \overset{(2)}{=} 1 - \sum_{k=0}^{n-1} e^{-\lambda w} \frac{(\lambda w)^k}{k!}, \\ f_{W_n}(w) &\overset{(3)}{=} \lambda e^{-\lambda w} \sum_{k=0}^{n-1} \frac{(\lambda w)^k}{k!} - e^{-\lambda w} \sum_{k=1}^{n-1} \frac{\lambda^k w^{k-1}}{(k-1)!} \\ &\overset{(4)}{=} \lambda e^{-\lambda w} \Big[\sum_{k=0}^{n-1} \frac{(\lambda w)^k}{k!} - \sum_{j=0}^{n-2} \frac{(\lambda w)^j}{j!} \Big] \overset{(5)}{=} \frac{\lambda^n w^{n-1}}{(n-1)!} e^{-\lambda w}. \end{aligned} \tag{2.19}$$

(1) The n-th arrival is before w if and only if the count at w is at least n.
(2) $N(w) \sim \text{Poisson}(\lambda w)$; complement of $k \le n-1$.
(3) Differentiate with the product rule: the derivative of $e^{-\lambda w}$ gives the first sum (with the sign flipped by the leading minus), the derivative of $(\lambda w)^k/k!$ is $\lambda^k w^{k-1}/(k-1)!$ (zero for $k = 0$).
(4) In the second sum take out λ and rename $j = k - 1$.
(5) The two sums *telescope*: every term cancels except $k = n - 1$ in the first.

This is the **Erlang** density.[3] With $\Gamma(n) = (n-1)!$ and the scale $\beta = 1/\lambda$ it is a special case of the general gamma family.

Working definition 2.9 (gamma distribution)

$X \sim \text{Gamma}(\alpha, \beta)$ (shape $\alpha > 0$, scale $\beta > 0$) if

$$f(x \mid \alpha, \beta) = \frac{1}{\Gamma(\alpha)\,\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, \qquad x > 0. \tag{2.20}$$

[for integer $\alpha = n$ and $\beta = 1/\lambda$, the waiting time for the n-th event of a rate-λ Poisson process]. Exp is Gamma$(1, \beta)$. *Recognise it by:* a sum of independent exponential waits, or more generally a positive quantity whose density rises like a power and falls like an exponential.

It is normalised because substituting $y = x/\beta$ turns $\int_0^\infty x^{\alpha-1} e^{-x/\beta} \mathrm{d}x$ into $\beta^\alpha \Gamma(\alpha)$. Its moments follow from the *kernel trick*: recognise the integrand as another gamma density, up to its constant.

$$\mathbb{E}\, X \overset{(1)}{=} \int_0^\infty \frac{x^\alpha e^{-x/\beta}}{\Gamma(\alpha)\beta^\alpha} \mathrm{d}x \overset{(2)}{=} \frac{\Gamma(\alpha+1)\beta^{\alpha+1}}{\Gamma(\alpha)\beta^\alpha} \overset{(3)}{=} \alpha\beta, \qquad \mathbb{E}\, X^2 \overset{(2)}{=} \frac{\Gamma(\alpha+2)\beta^{\alpha+2}}{\Gamma(\alpha)\beta^\alpha} = \alpha(\alpha+1)\beta^2.$$

(1) One extra factor of x raises the power to x^α.
(2) $\int_0^\infty x^{(\alpha+1)-1} e^{-x/\beta} \mathrm{d}x = \Gamma(\alpha+1)\beta^{\alpha+1}$, the normalisation of Gamma$(\alpha+1, \beta)$; likewise with $\alpha + 2$.
(3) $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$.

So $\mathrm{Var}\, X = \alpha(\alpha+1)\beta^2 - \alpha^2\beta^2 = \alpha\beta^2$ [CasellaBerger (3.3.7)–(3.3.8), p. 82]. For the Erlang wait, $\mathbb{E}\, W_n = n/\lambda$ and $\mathrm{Var}\, W_n = n/\lambda^2$, exactly n times the exponential values, as a sum of n independent gaps must give.

[3] The density printed in [Wasserman Thm. 23.35, p. 396] has $e^{-\lambda t}$ where $e^{-\lambda w}$ is meant. The proof there also writes S_1 for the first gap after defining the gaps from S_0.

The mgf, by the same substitution, is $M_X(t) = (1-\beta t)^{-\alpha}$ for $t < 1/\beta$. Multiplying mgfs shows that independent Gamma(α_i, β) variables with a *common scale* add to Gamma$(\sum \alpha_i, \beta)$: n gaps of Gamma$(1, 1/\lambda)$ make Gamma$(n, 1/\lambda)$, and the first n gaps followed by the next m make W_{n+m}.

The gamma–Poisson identity. The first line of (2.19) is a useful identity in its own right. For integer α and $X \sim \text{Gamma}(\alpha, \beta)$,

$$\mathrm{P}(X \le x) = \mathrm{P}(Y \ge \alpha), \qquad Y \sim \text{Poisson}(x/\beta).$$

A tail probability of a continuous waiting time equals a tail probability of a discrete count [CasellaBerger (3.3.9), p. 82]. Casella and Berger prove it by repeated integration by parts; the duality "n-th event before x $\Leftrightarrow$ at least n events by x" makes it obvious.

2.6.4 Worked examples

Example: muon survival. Stopped muons decay with mean lifetime $\tau = 2.197\ \mu$s. What fraction survives 5 μs? By (2.18) with $\lambda = 1/\tau$, $\mathrm{P}(T > 5\ \mu\text{s}) = e^{-5/2.197} = 0.1027$. By memorylessness, a muon that has already survived 5 μs again has about a 10% chance to survive the next 5 μs. So measuring τ does not require knowing when each muon was produced. Any start time after the muon stops gives the same exponential shape with the same τ. This is why a muon-lifetime experiment can start its clock at the stop signal and ignore everything earlier.

Example: waiting for the third photon. Photons reach a photocathode at $\lambda = 2$ per ns (after thinning by the quantum efficiency). The discriminator fires on the third photoelectron. What is the probability that it has fired within 1 ns? By the gamma–Poisson identity,

$$\mathrm{P}(W_3 \le 1) \overset{(1)}{=} \mathrm{P}(N(1) \ge 3) \overset{(2)}{=} 1 - e^{-2}\left(1 + 2 + \frac{2^2}{2}\right) = 1 - 5e^{-2} = 1 - 0.677 = 0.323.$$

(1) Duality; $N(1) \sim \text{Poisson}(2)$.
(2) Complement of $k = 0, 1, 2$.

The timing jitter of this trigger is the spread of W_3. With $\beta = 1/\lambda = 0.5$ ns, $W_3 \sim \text{Gamma}(3, 0.5\text{ ns})$ has mean $\alpha\beta = 1.5$ ns and standard deviation $\sqrt{\alpha}\,\beta = \sqrt{3} \times 0.5 = 0.87$ ns. In general, firing on the n-th photon gives mean n/λ and standard deviation $\sqrt{n}/\lambda$, so the relative jitter is $1/\sqrt{n}$. The more photons we wait for, the later the trigger, but the more *predictable*.

Example: counts or gaps: the same information about λ. We watch a Poisson stream for a time T and see N events at times $w_1 < \cdots < w_N$. From the counts alone, the probability of the data is $e^{-\lambda T}(\lambda T)^N/N!$. From the gaps: N gaps of the observed lengths (density $\lambda e^{-\lambda s_i}$ each) and then no event in the last stretch $T - w_N$ (probability $e^{-\lambda(T-w_N)}$). The product is

$$\prod_{i=1}^{N} \lambda e^{-\lambda s_i} \times e^{-\lambda(T-w_N)} \overset{(1)}{=} \lambda^N e^{-\lambda w_N} e^{-\lambda(T-w_N)} = \lambda^N e^{-\lambda T}.$$

(1) The gaps add up to the time of the last event, $\sum_i s_i = w_N$.

As functions of λ, the count probability and the gap product are both proportional to $\lambda^N e^{-\lambda T}$ [Wasserman Ex. 23.36, p. 396]. So the detailed arrival times carry no information about a constant rate beyond N and T. Both likelihoods peak at the same estimate, $\hat{\lambda} = N/T$ (Chapter 3). The exact times *do* matter as soon as we want to test whether the rate is constant.

2.6.5 Python: a stream built from its gaps

To simulate the gaps we need exponential numbers, made from uniform ones. One fact does it [Cowan (2.16)–(2.18), p. 30]: if X has a continuous CDF F, then $U = F(X)$ is uniform on $[0, 1]$. Reading it backwards, $X = F^{-1}(U)$ has CDF F. For the exponential, $U = 1 - e^{-\lambda T}$ gives $T = -\ln(1-U)/\lambda$. Since $1 - U$ is also uniform, $T = -\ln U/\lambda$ does the same job. This is *inverse-transform sampling*, developed in general in Chapter 6.

Listing 2.7: Exponential gaps, arrival times, counts per window (code/ch02/06_waiting_times.py)

```
23 # (a) gaps by inverse transform, then the event times are the running sums
24 gaps = -np.log(rng.uniform(size=200_000)) / lam
25 times = np.cumsum(gaps)
26 Tmax = np.floor(times[-1])
27 counts = np.bincount(np.floor(times[times < Tmax]).astype(int), minlength=int(Tmax))  #
       events per 1 s window
28
29 # (b) arrival time of the n-th event: sum of n independent gaps
30 def erlang_pdf(w, n):
31     return lam**n * w**(n - 1) * np.exp(-lam * w) / factorial(n - 1)
```

Line 24 draws 200 000 exponential gaps with $\lambda = 2\ \mathrm{s}^{-1}$; line 25 turns them into arrival times by a running sum, exactly as in Figure 2.13. Line 27 counts the arrivals in each 1 s window: `np.floor(times)` is the window index of each event and `np.bincount` counts events per index. The function `erlang_pdf` is (2.19). Further down, the gaps are grouped in blocks of n and summed to give samples of W_n.

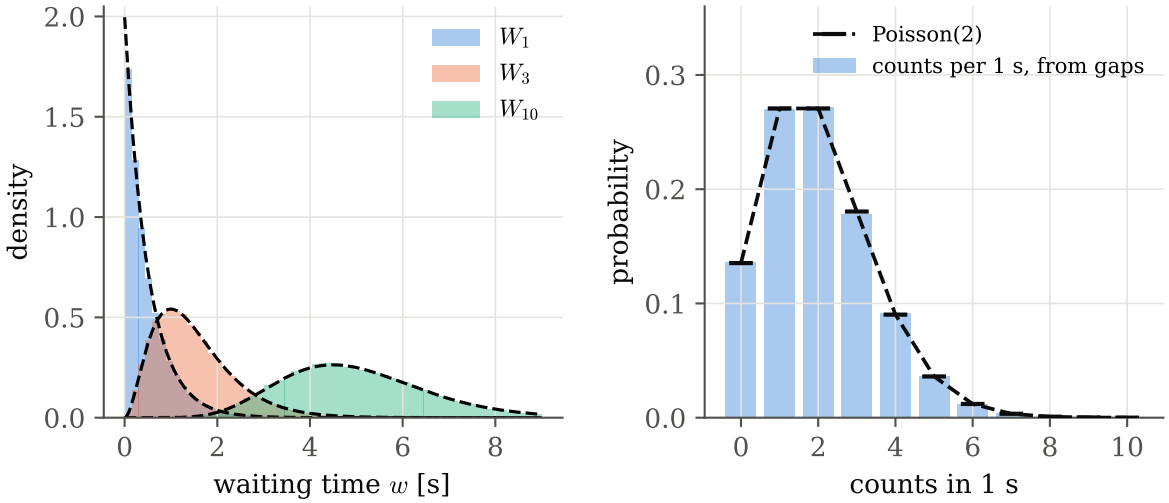


Figure 2.15: Left: histograms of the arrival time of the first, third and tenth event, built by summing 1, 3 or 10 simulated gaps, with the Erlang densities as dashed curves. The first arrival is most likely immediately; later arrivals have a peak that moves right and a width that grows only like $\sqrt{n}$. Right: counting the same stream in 1 s windows gives the Poisson(2) distribution, closing the loop from gaps back to counts.

The gaps have sample mean 0.4993 s and variance 0.2501 s^2 (theory $1/\lambda = 0.5$ and $1/\lambda^2 = 0.25$). The counts per second have mean 2.003 and variance 2.017 (theory 2 and 2). The tenth arrival has mean 4.993 s and variance 2.539 s^2 (theory 5 and 2.5). For memorylessness, among the gaps that exceeded 1 s the fraction that exceeded 1.5 s is 0.3694, against 0.3675 for all gaps exceeding 0.5 s and the theoretical $e^{-1} = 0.3679$ (Figure 2.15).

Watch out: rate or scale? three conventions for one distribution

The same gamma distribution is written with a *scale* β (AoS, Casella–Berger, `scipy.stats.gamma(a, scale=beta)`, `numpy`'s `rng.gamma(shape, scale)`), or with a *rate* $1/\beta$ (Lambert, most Bayesian texts and Stan). "Exp(2)" means mean 2 in one book and mean 1/2 in another. Always check by computing the mean. Lambert's lottery example is a warning: three prize-claiming times, each exponential with rate 1 per day, sum to a gamma with *shape* 3 and rate 1, which the text describes with the words "scale" and "shape" swapped [Lambert §8.4.9].

Not every lifetime is exponential. Memorylessness means a constant *hazard* $h(t) = f(t)/\mathrm{P}(T > t)$ [the probability per unit time of failing now, given survival so far]. For the exponential $h(t) = \lambda$. Components that wear out have a rising hazard and are better described by the Weibull distribution, $\mathrm{P}(T > t) = e^{-(t/\beta)^{\gamma}}$ with $\gamma > 1$ [CasellaBerger (3.3.12), p. 83]. Particles and nuclei have no wear, which is why the exponential law of decay is exact (to the extent quantum mechanics allows).

2.7 Discrete waiting times and extra spread: geometric and negative binomial

At a collider, bunches cross every 25 ns and each crossing fires a rare trigger with probability p. How many crossings do we wait until the first trigger? Until the third? This is the waiting-time question of Section 2.6, but with time coming in discrete ticks. A second puzzle has the same answer. A galaxy survey counts galaxies in cells of equal volume. The mean is 10 per cell, but the variance is 30, not the 10 a Poisson count would have. The counts are *overdispersed*: they scatter more than Poisson, because galaxies cluster and the underlying density varies from cell to cell. Two distributions answer both questions. The geometric and negative binomial distributions count trials until a given number of successes. The negative binomial then reappears as a Poisson count whose mean itself fluctuates, and in that role it is the standard model for counts that scatter more than Poisson.

Physical picture: ticks of a clock, and a Poisson with a wobbling rate. The geometric wait is the exponential wait with time chopped into ticks; shrink the ticks and it becomes the exponential. The overdispersed count is a Poisson count seen through a fluctuating medium, in two stages. First nature picks the local rate. Then a Poisson count is drawn with that rate. The total scatter is the sum of the scatter from the two stages: the Poisson "shot noise" plus the variance of the rate. Photon counts from a twinkling star behave this way.

The question, step by step.

1. The first success is on trial k exactly when trials $1, \ldots, k-1$ fail and trial k succeeds. Multiply.
2. The r-th success is on trial x exactly when trial x succeeds and the first $x-1$ trials contain $r-1$ successes. That is a binomial count times one more trial.
3. If a Poisson mean is itself random, average the Poisson pmf over its distribution (marginalise).

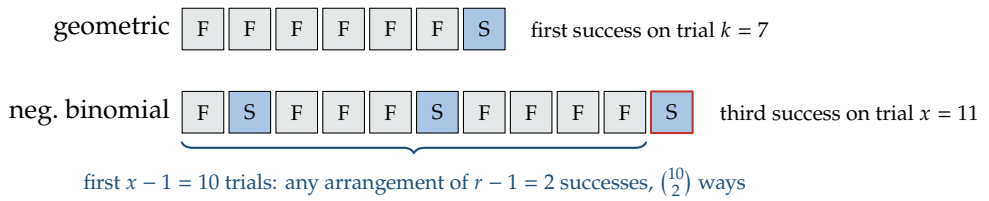


Figure 2.16: Two waits in a sequence of Bernoulli trials. Top: the first success after six failures. Bottom: the third success (outlined in red) is the last trial; before it, the two earlier successes can sit in any of $\binom{10}{2} = 45$ arrangements among the first ten trials, each equally probable.

2.7.1 The geometric distribution

Working definition 2.10 (geometric distribution)

$X \sim \text{Geom}(p)$ if X is the number of the trial on which the first success occurs, in a sequence of independent Bernoulli(p) trials [X counts the successful trial too, so $X \geq 1$]:

$$\mathrm{P}(X = k) = (1-p)^{k-1}p, \qquad k = 1, 2, 3, \ldots \tag{2.21}$$

Recognise it by: the number of tries until the first success, with a constant chance per try.

The formula follows from the first step listed above. The first success is on trial k when $k-1$ failures, each with probability $1-p$, are followed by a success, all independent; multiplying gives $(1-p)^{k-1}p$. Some books (and `scipy.stats.nbinom`) count instead the number of *failures* before the first success, $Y = X - 1 \in \{0, 1, \ldots\}$, with $\mathrm{P}(Y = y) = (1-p)^y p$.[4]

Geometric series. Everything follows from $\sum_{j=0}^{\infty} r^j = \frac{1}{1-r}$ for $|r| < 1$. Proof: the partial sum $S_n = 1 + r + \cdots + r^n$ satisfies $S_n - rS_n = 1 - r^{n+1}$, so $S_n = \frac{1-r^{n+1}}{1-r} \to \frac{1}{1-r}$. Differentiating term by term (allowed inside the radius of convergence) gives $\sum_{j=1}^{\infty} j r^{j-1} = \frac{1}{(1-r)^2}$ and $\sum_{j=2}^{\infty} j(j-1) r^{j-2} = \frac{2}{(1-r)^3}$. With $r = 1 - p$:

$$\sum_{k=1}^{\infty} \mathrm{P}(X = k) \overset{(1)}{=} p \sum_{k=1}^{\infty} (1-p)^{k-1} \overset{(2)}{=} \frac{p}{1-(1-p)} = 1,$$

$$\mathbb{E}\, X \overset{(3)}{=} p \sum_{k=1}^{\infty} k(1-p)^{k-1} \overset{(4)}{=} \frac{p}{p^2} = \frac{1}{p},$$

$$\mathbb{E}[X(X-1)] \overset{(3)}{=} p(1-p) \sum_{k=2}^{\infty} k(k-1)(1-p)^{k-2} \overset{(4)}{=} \frac{2p(1-p)}{p^3} = \frac{2(1-p)}{p^2},$$

$$\mathrm{Var}\, X \overset{(5)}{=} \frac{2(1-p)}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{1-p}{p^2}.$$

(1) Take out p.
(2) Geometric series with $j = k - 1$ running from 0.
(3) Definition of the (factorial) moment; for the second line one factor $(1-p)$ is taken out so that the power is $k-2$.
(4) The differentiated series with $1 - r = p$.
(5) $\mathrm{Var}\, X = \mathbb{E}[X(X-1)] + \mathbb{E}\, X - (\mathbb{E}\, X)^2$; put over p^2: $2 - 2p + p - 1 = 1 - p$.

[4] The normalisation check for the geometric distribution in [Wasserman pp. 26–27] has a misprint in its middle sum, which is printed as $p \sum_{k \geq 1} (1-p)^k$. That sum equals $1 - p$, not 1. The exponent must be $k - 1$, as below.

A trigger with $p = 0.1$ fires on average every 10 crossings, with a standard deviation $\sqrt{0.9}/0.1 = 9.5$ crossings: nearly as large as the mean, just like the exponential.

The tail and no memory. The first success comes after trial k exactly when the first k trials all fail, so $\mathrm{P}(X > k) = (1-p)^k$. Then

$$\mathrm{P}(X > s+t \mid X > t) \overset{(1)}{=} \frac{\mathrm{P}(X > s+t)}{\mathrm{P}(X > t)} = \frac{(1-p)^{s+t}}{(1-p)^t} = (1-p)^s = \mathrm{P}(X > s)$$

(1) $X > s+t$ implies $X > t$, and the definition of conditional probability.

for integers $s, t \geq 0$. The geometric is memoryless: a trigger that has not fired for 1000 crossings is not "due" [CasellaBerger (3.2.11), p. 80].

From ticks to continuous time. Let the trials be ticks of length h with $p = \lambda h$. The waiting time is $T = hX$, and

$$\mathrm{P}(T > t) = \mathrm{P}(X > t/h) = (1-\lambda h)^{t/h} \xrightarrow{h \to 0} e^{-\lambda t},$$

by the compound-interest limit of Section 2.4 with $n = t/h$. The geometric becomes the exponential, just as the binomial became the Poisson. The slices of Figure 2.10 are exactly these ticks.

2.7.2 The negative binomial

Working definition 2.11 (negative binomial distribution)

In a sequence of independent Bernoulli(p) trials, let X be the number of the trial on which the r-th success occurs. Then $X \sim \mathrm{NegBin}(r, p)$ with

$$\mathrm{P}(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \qquad x = r, r+1, \ldots \tag{2.22}$$

Equivalently, the number of failures before the r-th success, $Y = X - r$, has

$$\mathrm{P}(Y = y) = \binom{r+y-1}{y} p^r (1-p)^y, \qquad y = 0, 1, 2, \ldots \tag{2.23}$$

[for $r = 1$ these are the two forms of the geometric]. *Recognise it by:* the number of tries until a fixed number r of successes, or (see below) a count that is more spread out than Poisson.

Derivation of (2.22). This is the second step listed above (Figure 2.16). The r-th success is on trial x if and only if two things happen. First, the first $x-1$ trials contain exactly $r-1$ successes; this has the binomial probability $\binom{x-1}{r-1} p^{r-1}(1-p)^{x-r}$. Second, trial x succeeds, with probability p, independently of the earlier trials. Multiplying the two gives (2.22). Replacing $x = y + r$ and using $\binom{n}{k} = \binom{n}{n-k}$ gives (2.23) [CasellaBerger (3.2.9)–(3.2.10), p. 78].

Mean and variance: a sum of geometric waits. After each success the process starts afresh, because the trials are independent. So Y, the number of failures before the r-th success, is the sum of r independent copies of $Y_1 = X_1 - 1$, the number of failures before the first

success; here $X_1 \sim \mathrm{Geom}(p)$. From the geometric moments, Y_1 has $\mathbb{E}\,Y_1 = \frac{1}{p} - 1 = \frac{1-p}{p}$ and $\mathrm{Var}\,Y_1 = \mathrm{Var}\,X_1 = \frac{1-p}{p^2}$. Hence

$$\mathbb{E}\,Y = \frac{r(1-p)}{p}, \qquad \mathrm{Var}\,Y = \frac{r(1-p)}{p^2} = \mathbb{E}\,Y + \frac{(\mathbb{E}\,Y)^2}{r}. \tag{2.24}$$

The last form follows by writing $\mu = r(1-p)/p$: then $\mu + \mu^2/r = \frac{r(1-p)}{p}\big(1 + \frac{1-p}{p}\big) = \frac{r(1-p)}{p^2}$. *The variance always exceeds the mean*, by μ^2/r.

Why "negative" binomial. Write out the coefficient with $r + y - 1$ over y and pull out a sign from each of the y factors:

$$\binom{r+y-1}{y} \overset{(1)}{=} \frac{(r+y-1)(r+y-2)\cdots r}{y!} \overset{(2)}{=} (-1)^y \frac{(-r)(-r-1)\cdots(-r-y+1)}{y!} \overset{(3)}{=} (-1)^y \binom{-r}{y}.$$

(1) Cancel $(r-1)!$ from $(r+y-1)!/(y!(r-1)!)$, leaving y factors on top.
(2) Each factor $(r+j) = -(-r-j)$, for $j = 0, \ldots, y-1$, read in reverse order.
(3) The *generalised* binomial coefficient $\binom{a}{y} = a(a-1)\cdots(a-y+1)/y!$ makes sense for any real a, including $a = -r$.

Newton's binomial series $(1+z)^a = \sum_y \binom{a}{y} z^y$ for $|z| < 1$ then normalises (2.23): $\sum_y \binom{-r}{y}(-(1-p))^y p^r = (1-(1-p))^{-r} p^r = 1$. The pmf is a term of a binomial expansion with a negative exponent.

Its mgf and the Poisson limit. For one geometric failure count, $\mathbb{E}\,e^{tY_1} = \sum_y p(1-p)^y e^{ty} = p/(1-(1-p)e^t)$ (geometric series, valid for $(1-p)e^t < 1$). A sum of r independent copies has the r-th power. Now let $r \to \infty$ while holding the mean $r(1-p)/p$ close to a fixed λ; to leading order $1 - p = \lambda/r$:

$$M_Y(t) = \left(\frac{p}{1-(1-p)e^t}\right)^r \overset{(1)}{=} \frac{(1-\lambda/r)^r}{(1-\lambda e^t/r)^r} \xrightarrow{r\to\infty} \frac{e^{-\lambda}}{e^{-\lambda e^t}} = e^{\lambda(e^t-1)}.$$

(1) Substitute $p = 1 - \lambda/r$.

The limit is the Poisson mgf (2.13): with many required successes and rare failures, the failure count becomes Poisson, and the excess variance μ^2/r disappears [CasellaBerger p. 79].

2.7.3 A Poisson count with a fluctuating mean

The most important appearance of the negative binomial in physics is not about waiting at all. Suppose the count in a cell is Poisson with mean Λ, but Λ itself varies: from cell to cell (clustering), from run to run (beam intensity) or from pixel to pixel (a twinkling source). Take Λ to follow a gamma distribution with shape κ and scale μ/κ. By the gamma moments of Section 2.6, $\mathbb{E}\,\Lambda = \kappa \cdot \mu/\kappa = \mu$ and $\mathrm{Var}\,\Lambda = \kappa(\mu/\kappa)^2 = \mu^2/\kappa$. The distribution of the count is then the **mixture**: the marginal over the nuisance variable Λ. Write $\theta = \mu/\kappa$ for the gamma scale:

$$\begin{aligned} \mathrm{P}(N=n) &\overset{(1)}{=} \int_0^\infty e^{-\lambda}\frac{\lambda^n}{n!}\,\frac{\lambda^{\kappa-1}e^{-\lambda/\theta}}{\Gamma(\kappa)\theta^\kappa}\,\mathrm{d}\lambda \overset{(2)}{=} \frac{1}{n!\,\Gamma(\kappa)\,\theta^\kappa}\int_0^\infty \lambda^{n+\kappa-1}e^{-\lambda(1+\theta)/\theta}\,\mathrm{d}\lambda \\ &\overset{(3)}{=} \frac{\Gamma(n+\kappa)}{n!\,\Gamma(\kappa)\,\theta^\kappa}\Big(\frac{\theta}{1+\theta}\Big)^{n+\kappa} \overset{(4)}{=} \frac{\Gamma(n+\kappa)}{n!\,\Gamma(\kappa)}\Big(\frac{\kappa}{\kappa+\mu}\Big)^\kappa\Big(\frac{\mu}{\kappa+\mu}\Big)^n. \end{aligned}$$

(1) Law of total probability for a continuous nuisance variable: Poisson pmf given $\Lambda = \lambda$, times the gamma density of Λ, integrated.
(2) Collect the powers of λ and the exponentials: $\lambda + \lambda/\theta = \lambda(1+\theta)/\theta$.
(3) The integral is the gamma normalisation with shape $n + \kappa$ and scale $\theta/(1+\theta)$.
(4) $\theta^{-\kappa}(\theta/(1+\theta))^{\kappa} = (1+\theta)^{-\kappa}$, and $\theta/(1+\theta) = \mu/(\kappa+\mu)$, $1/(1+\theta) = \kappa/(\kappa+\mu)$.

For integer κ, $\Gamma(n+\kappa)/(n!\,\Gamma(\kappa)) = \binom{\kappa+n-1}{n}$, and this is (2.23) with $r = \kappa$ and $p = \kappa/(\kappa+\mu)$. The formula makes sense for any real $\kappa > 0$, and in this role κ is called the **inverse dispersion** [Lambert §8.4.4].

The variance is cleanest by the law of total variance (Chapter 1). It splits the scatter into the two stages of the physical picture above: nature picks the rate, then the Poisson count is drawn.

$$\operatorname{Var} N \overset{(1)}{=} \mathbb{E}\big[\operatorname{Var}(N \mid \Lambda)\big] + \operatorname{Var}\big(\mathbb{E}[N \mid \Lambda]\big) \overset{(2)}{=} \mathbb{E}\,\Lambda + \operatorname{Var}\Lambda \overset{(3)}{=} \mu + \frac{\mu^2}{\kappa}. \tag{2.25}$$

(1) Law of total variance: average of the conditional variance plus variance of the conditional mean.
(2) Given Λ, N is Poisson, so its conditional mean and variance both equal Λ.
(3) Gamma moments.

The variance is the Poisson shot noise μ plus the variance of the rate, μ^2/κ. Nothing in the second step used the gamma form: *any* fluctuating mean gives $\operatorname{Var} N = \mu + \operatorname{Var}\Lambda > \mu$. The gamma only makes the full distribution a closed form.

2.7.4 Worked examples

Example: how long until the trigger fires? With $p = 0.1$ per crossing, what is the probability of waiting more than 30 crossings for the first trigger? $\mathrm{P}(X > 30) = (0.9)^{30} = 0.04239$. And for a light bulb that fails on any given day with probability 0.001, $\mathrm{P}(X > 30) = 0.999^{30} = 0.970$: a 97% chance to last a month [CasellaBerger Ex. 3.2.7, p. 80]. By memorylessness, the same 97% applies to the next month of a bulb that has already lasted a year. Real bulbs age, so the geometric is only a rough model for them; a trigger has no memory, so for it the model is exact.

Example: inverse binomial sampling. To measure a rare branching fraction p we can fix the number of trials and count successes (binomial), or keep collecting until we have r successes (negative binomial) [CasellaBerger Ex. 3.2.6, p. 79]. Suppose $p = 0.02$ and we stop at $r = 100$ signal events. The number of events examined is $X = Y + r$ with

$$\mathbb{E}\,X \overset{(1)}{=} \frac{r}{p} = 5000, \qquad \operatorname{sd}(X) \overset{(2)}{=} \frac{\sqrt{r(1-p)}}{p} = \frac{\sqrt{98}}{0.02} = 495.$$

(1) $\mathbb{E}\,X = \mathbb{E}\,Y + r = r(1-p)/p + r = r/p$.
(2) $\operatorname{Var} X = \operatorname{Var} Y = r(1-p)/p^2$ from (2.24).

The relative uncertainty of the stopping point is $\sqrt{1-p}/\sqrt{r} \approx 10\%$, set by r alone. Stopping rules like this matter for inference: Chapter 3 and Chapter 5 show that the binomial and negative binomial give proportional likelihoods for p, so a Bayesian analysis does not care which rule was used, while a p-value does.

Example: counts in cells of a clustered galaxy field. Cells of a survey contain on average $\mu = 10$ galaxies. If galaxies were placed independently (a Poisson process in space, Section 2.5), the variance of the cell counts would be 10. But galaxies trace a fluctuating matter density. The

local mean in a cell is $\Lambda = \mu(1+\delta)$, where δ is the density contrast smoothed over the cell, with variance σ^2_{cell}. So $\operatorname{Var}\Lambda = \mu^2\sigma^2_{\text{cell}}$, and by (2.25),

$$\operatorname{Var} N = \mu + \mu^2\sigma^2_{\text{cell}}.$$

If $\sigma^2_{\text{cell}} = 0.2$, then $\operatorname{Var} N = 10 + 20 = 30$, three times the Poisson value. This is the gamma–Poisson model with $1/\kappa = \sigma^2_{\text{cell}}$, that is, $\kappa = 5$. Read backwards, the formula is a measurement. The *excess* variance of counts in cells over the Poisson value measures the clustering amplitude σ^2_{cell}. That excess is the one-point statistic behind the two-point function of Chapter 7. The Poisson term μ is the *shot noise* that every galaxy power spectrum must subtract.

2.7.5 Python: trials one after another, and a wobbling rate

Listing 2.8: Geometric and negative binomial by running trials; the gamma–Poisson mixture (`code/ch02/07_geometric_negbin.py`)

```
# (a),(b): run Bernoulli trials one after another and record where successes fall
p, r, reps = 0.1, 3, 50_000
first, fails_before_r = np.empty(reps, int), np.empty(reps, int)
for i in range(reps):
    trials = rng.uniform(size=400) <= p            # 400 trials is (much) more than enough here
    s = np.flatnonzero(trials) + 1                 # trial numbers (1-based) of the successes
    first[i] = s[0]                                # geometric: trial of the first success
    fails_before_r[i] = s[r - 1] - r               # negative binomial: failures before r-th success

# (c): gamma-distributed mean, then a Poisson count with that mean
mu, kappa, runs = 10.0, 5.0, 200_000
lam_run = rng.gamma(shape=kappa, scale=mu / kappa, size=runs)   # mean mu, variance mu^2/kappa
N = rng.poisson(lam_run)
```

For each of 50 000 experiments the script runs 400 Bernoulli(0.1) trials (the chance that fewer than three successes occur in 400 trials is negligible) and finds the trial numbers of the successes with `np.flatnonzero`. The first one is the geometric wait; the third one minus $r = 3$ is the number of failures before the third success. For the mixture, line 34 draws a rate for each run from a gamma distribution with shape $\kappa = 5$ and scale μ/κ (numpy's convention is scale), and line 35 draws one Poisson count with that rate.

The geometric wait has sample mean 10.03 and variance 91.61 (theory $1/p = 10$ and $(1-p)/p^2 = 90$); 0.04384 of the experiments waited more than 30 crossings, against $0.9^{30} = 0.04239$. The failures before the third success have mean 26.97 and variance 271 (theory 27 and 270). The mixture counts have mean 9.997 and variance 29.87, against $\mu + \mu^2/\kappa = 30$ (Figure 2.17).

What this bought us. Discrete waits (geometric, negative binomial) are the tick-by-tick versions of the exponential and gamma. The negative binomial also describes a Poisson count with a gamma-distributed mean, with variance $\mu + \mu^2/\kappa$. Whenever counts scatter more than their mean, look for a fluctuating rate.

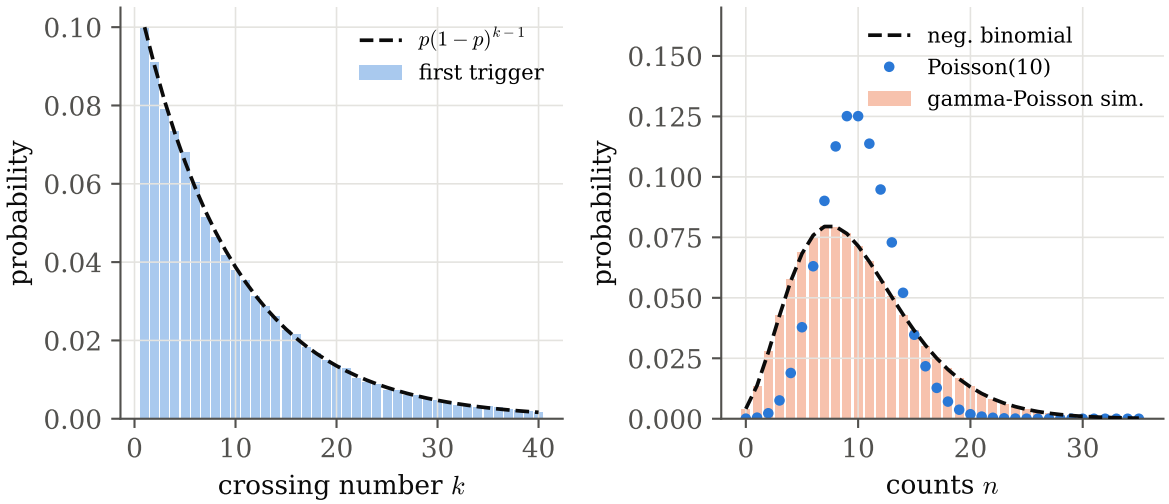


Figure 2.17: Left: the crossing number of the first trigger (bars) against $p(1-p)^{k-1}$ (dashed): a geometric staircase, highest at $k = 1$, falling by a factor 0.9 per crossing. Right: counts whose Poisson mean wobbles from run to run (bars) follow the negative binomial (dashed), far wider than the Poisson with the same mean 10 (blue dots).

2.8 When the Poisson assumptions fail, and where the Poisson goes for large λ

Point a Geiger counter at a radioactive source and move it closer. At first the count rate rises like $1/r^2$, as the solid angle says it should. Closer in, it rises more slowly than $1/r^2$, and with some counters, very close to a strong source, it even *falls*. Now look at the counts in successive short windows at a fixed distance. A second oddity appears: at high rate they scatter *less* than the $\sqrt{\text{mean}}$ that the Poisson distribution promises. A clustered galaxy survey shows the opposite oddity, as we saw in the counts-in-cells example of Section 2.7: the counts in cells scatter *more* than Poisson.

Both oddities are the Poisson inputs of Section 2.5 failing, in opposite directions. Each physical effect breaks a particular input (P1)–(P3) of Section 2.5.1. One number computed from the data detects the failure. And the commonest failure in counting detectors, dead time, can be corrected. A separate question concerns large means: there the Poisson distribution becomes close to a Gaussian, and the replacement is safe in some places and badly wrong in others.

The questions.

1. Which Poisson input does each physical effect break, and does it push the variance of the counts up or down?
2. Given a set of counts, how do we test whether they are Poisson?
3. For a counter that is blind for a time τ after each hit, what is the recorded rate, how do we get the true rate back, and how much do the recorded counts scatter?
4. For large λ, how close is Poisson(λ) to the Gaussian $\mathcal{N}(\lambda, \lambda)$, and where does the approximation fail?

2.8.1 Three inputs, two directions of failure

Recall the three physical inputs from which the Poisson process was derived (Section 2.5.1):

(P1) what happens in one time slice is independent of what happened in all the others [no memory, no triggering, no suppression];

(P2) the rate λ is the *same* at all times [the source does not drift];

(P3) events come one at a time [never two at the same instant].

From these came the pmf $e^{-\lambda t}(\lambda t)^n/n!$ for the count N in a window of length t, and with it the defining fingerprint $\operatorname{Var} N = \mathbb{E}\, N$. The Poisson is the distribution of "counts of rare events" [Wasserman p. 27]. Here "rare" means precisely that the events do not interact with each other and are not modulated by anything. The same point holds from the user's side [Lambert §8.4.3]: the mean and the variance are the same parameter, so the Poisson is useful only for data with $\operatorname{Var} N \approx \mathbb{E}\, N$.

Real effects break the inputs in one of two directions.

More spread than Poisson (overdispersion). Two kinds of effect make the counts clump.

- *The rate itself fluctuates* (input P2 fails). The beam intensity drifts from run to run, the sky background varies across the field, the density of dark matter modulates where galaxies form. In Section 2.7 we proved that a Poisson count with a random mean Λ has $\operatorname{Var} N = \mathbb{E}\,\Lambda + \operatorname{Var}\Lambda$, eq. (2.25). The extra term can only add.
- *Events trigger further events* (input P1 fails in the positive direction). A photomultiplier pulse is followed by an afterpulse from an ion drifting back to the photocathode. A cosmic-ray primary makes an air shower of secondaries. Events arrive in bunches, and a window either catches a bunch or misses it.

The binomial has the same failure mode. When the success probability p varies from group to group, for example between detector modules or between clinical trials, the number of successes is spread wider than $np(1-p)$. Its standard model is the beta-binomial, the binomial analogue of the negative binomial [Lambert §8.4.5].

Less spread than Poisson (underdispersion). Other effects make the events avoid each other, so that the stream becomes more regular than random.

- *Events suppress further events* (input P1 fails in the negative direction). After recording a hit the counter is blind for a **dead time** τ. A neuron cannot fire again during its refractory period. Pile-up rejection throws away events that come too close together.
- *The pool of possible events is finite.* With n nuclei and decay probability p per window the count is binomial, with $\operatorname{Var} N = np(1-p) < np = \mathbb{E}\, N$. This is negligible for a macroscopic source ($n \sim 10^{20}$), but not for a handful of trapped ions, or for a short-lived isotope that decays appreciably during the measurement (then input P2 fails as well: the rate falls with time).

Input (P3) fails when two events in one slice are recorded as one. This **pile-up** (two photons in one X-ray CCD pixel within one readout frame, two particles in one calorimeter cell in one bunch crossing) loses events and distorts the recorded energies. As far as the counts are concerned, it acts like a dead time of one frame.

Figure 2.18 shows the three kinds of stream. The eye reads clumping and regularity immediately. What we need is a number that does the same for a data set of thousands of windows.

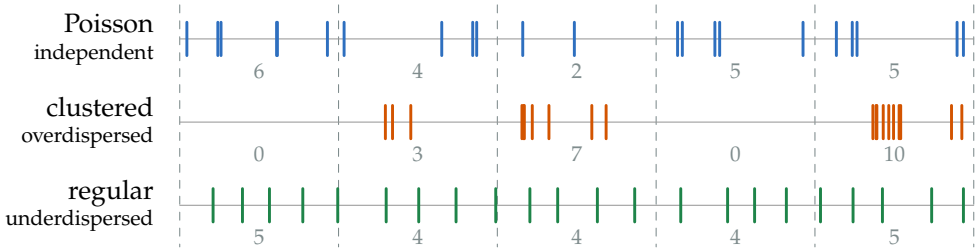


Figure 2.18: Three streams of about twenty events, each cut into five equal windows (dashed lines); the small numbers are the counts per window. Top: a Poisson stream, with the irregular gaps and chance near-coincidences of genuine randomness. Middle: the same kind of stream in clumps (a few independent "parents", each with a small bunch of offspring): whole windows are empty and others are crowded. Bottom: a dense Poisson stream passed through a counter that is blind for a fixed time after each recorded hit: no two events are close, the gaps are almost equal, and the counts barely change from window to window.

2.8.2 The Fano factor: a one-number test

Since a Poisson count has variance equal to its mean, the natural diagnostic is their ratio.

Working definition 2.12 (Fano factor (index of dispersion))

For a count N with mean $\mu > 0$, the **Fano factor** is

$$F = \frac{\operatorname{Var} N}{\mathbb{E}\, N}.$$

$F = 1$ for a Poisson count, $F > 1$ is **overdispersion** [clumped events or a fluctuating rate], and $F < 1$ is **underdispersion** [events that avoid each other, or a finite pool]. From k independent counts $n_1, \dots, n_k$ of the same window we estimate it by

$$\hat{F} = \frac{s^2}{\bar{n}}, \qquad \bar{n} = \frac{1}{k}\sum_{i=1}^{k} n_i, \quad s^2 = \frac{1}{k-1}\sum_{i=1}^{k}(n_i - \bar{n})^2.$$

For Poisson counts with $\bar{n} \gg 1$, $\hat{F}$ scatters about 1 with standard deviation $\sqrt{2/(k-1)}$.

Values of F we already know: Poisson $F = 1$; binomial $F = 1 - p$; negative binomial $F = 1 + \mu/\kappa$, from (2.25). The name comes from the physics of ionisation detectors, where U. Fano showed that the number of electron–ion pairs made by a particle of fixed energy fluctuates less than Poisson (the energy budget is fixed, so the pairs are not independent). The same ratio minus one is Mandel's Q parameter in quantum optics (see the example on sub-Poissonian light below).

How far from 1 is "significantly" far? A sample Fano factor is itself a random number, so a measured $\hat{F} = 0.97$ need not mean anything. To judge it we need the spread of $\hat{F}$ when the counts really are Poisson. Two facts give it. The first: for any independent sample of size k, the sample variance s^2 has

$$\operatorname{Var}(s^2) = \frac{2\sigma^4}{k-1} + \frac{\kappa_4}{k},$$

where σ^2 is the variance and κ_4 the fourth **cumulant** [the excess of the fourth central moment over its Gaussian value, $\kappa_4 = \mathbb{E}(N-\mu)^4 - 3\sigma^4$] of one count. The first term is the Gaussian result, derived with the χ^2 distribution in the next part. The second is the correction for a

non-Gaussian parent. The second fact: for a Poisson count every cumulant equals λ (proved in Section 2.8.5 below), so $\sigma^2 = \kappa_4 = \lambda$. The mean $\bar{n}$ fluctuates by only a relative $1/\sqrt{k\lambda}$, so in the ratio we may treat it as the constant λ. Then

$$\mathrm{sd}(\hat{F}) \overset{(1)}{=} \frac{\mathrm{sd}(s^2)}{\lambda} \overset{(2)}{=} \frac{1}{\lambda}\sqrt{\frac{2\lambda^2}{k-1} + \frac{\lambda}{k}} \overset{(3)}{=} \sqrt{\frac{2}{k-1} + \frac{1}{k\lambda}} \overset{(4)}{\approx} \sqrt{\frac{2}{k-1}}. \tag{2.26}$$

(1) $\hat{F} = s^2/\bar{n}$ with $\bar{n}$ replaced by its expected value λ; its relative fluctuation $1/\sqrt{k\lambda}$ is tiny next to that of s^2.
(2) The variance of s^2 with $\sigma^2 = \kappa_4 = \lambda$.
(3) Take λ^2 out of the square root.
(4) For $\lambda \gg 1$ the second term is negligible: $1/(k\lambda) \ll 2/(k-1)$.

The spread depends only on the number of windows k, not on the rate. The combination $D = (k-1)\hat{F} = \sum_i (n_i - \bar{n})^2/\bar{n}$ is the classical **index-of-dispersion statistic**. For Poisson counts it is close to a χ^2 variable with $k-1$ degrees of freedom, and using it to reject "the counts are Poisson" is a hypothesis test in the sense of Chapter 4.

Example: reading a Fano factor. We record $k = 4000$ windows. By (2.26), $\mathrm{sd}(\hat{F}) = \sqrt{2/3999} = 0.0224$.

- $\hat{F} = 0.96$ is $(0.96-1)/0.0224 = -1.8$ standard deviations from 1. That happens by chance in about 7% of Poisson data sets (two-sided), so it is no evidence against Poisson.
- $\hat{F} = 0.62$ is -17 standard deviations. The counts are certainly not Poisson, and they are underdispersed: look for dead time.
- $\hat{F} = 3.0$ is $+89$ standard deviations. Look for a fluctuating rate or clustering, as in the galaxy counts of Section 2.7, where $F = 1 + \mu\sigma^2_{\mathrm{cell}} = 3$.

With only $k = 20$ windows, $\mathrm{sd}(\hat{F}) = 0.32$, and even $\hat{F} = 0.62$ is barely more than one standard deviation from 1. A few windows cannot distinguish these streams. That is why the five windows of Figure 2.18 are only an illustration.

2.8.3 Dead time

The commonest failure in a counting detector is dead time. After a hit, the electronics need a time τ to shape the pulse, digitise it and reset. Anything arriving in that interval is lost. Typical values are $\tau \sim 100$–300 μs for a Geiger–Müller tube, ~ 10–100 ns for a photomultiplier with fast electronics, and one readout frame (seconds) for an X-ray CCD. Two idealised models bracket the real behaviour of most systems.

- **Non-paralysable**: the counter is blind for τ after each *recorded* hit. Hits that arrive during the dead period are lost and have no further effect.
- **Paralysable**: *every* hit, recorded or not, restarts the dead period. A hit is recorded only if the previous true hit came at least τ earlier. A burst of closely spaced hits keeps the counter blind for the whole burst.

Let n be the true rate (hits per second) and m the recorded rate. Figure 2.19 runs one stream of true hits through both models.

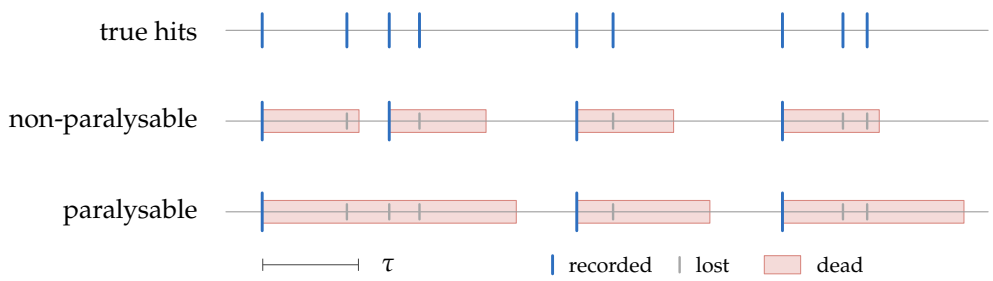


Figure 2.19: Nine true hits (top) seen by two counters with the same dead time τ (the bar at the bottom left). The non-paralysable counter (middle) is blind for exactly τ after each recorded hit, so it records four hits: the third true hit arrives just after the first dead period ends. The paralysable counter (bottom) is blind for τ after every hit, so the rapid run of four early hits keeps extending one long dead period and only three hits are recorded.

The recorded rate, non-paralysable. Count for a long time T. The counter records mT hits and is blind for τ after each, so it is dead for a total time $mT\tau$ and live for $T(1 - m\tau)$. During live time the true stream is untouched (input P1 for the true hits), so it delivers $nT(1 - m\tau)$ hits, and each of them is recorded:

$$mT \overset{(1)}{=} nT(1 - m\tau) \quad \Longrightarrow \quad m \overset{(2)}{=} \frac{n}{1 + n\tau}, \qquad n \overset{(3)}{=} \frac{m}{1 - m\tau}. \tag{2.27}$$

(1) Recorded hits = true hits that arrive while the counter is live.
(2) Solve for m: $m + nm\tau = n$.
(3) Solve instead for n. This is the dead-time correction: from the recorded rate and τ we recover the true rate.

At low rate, $n\tau \ll 1$, we get $m \approx n(1 - n\tau)$: the fractional loss is $n\tau$, the expected number of true hits in one dead period. At high rate, $m \to 1/\tau$: the counter records one hit per dead period and saturates.

The recorded rate, paralysable. A hit is recorded if and only if the gap since the previous *true* hit is at least τ. For a Poisson stream every gap is exponential with rate n (Section 2.6), so

$$m \overset{(1)}{=} n \ \mathrm{P}(\text{gap} \geq \tau) \overset{(2)}{=} n\, e^{-n\tau}. \tag{2.28}$$

(1) Of the nT true hits in time T, the recorded ones are those whose preceding gap is at least τ; by the law of large numbers their fraction is the probability of that.
(2) Survival function of the exponential gap, $\mathrm{P}(\text{gap} > t) = e^{-nt}$ from (2.18).

Now m is not monotonic in n. It rises, reaches its maximum $1/(e\tau)$ at $n\tau = 1$ (set $\mathrm{d}m/\mathrm{d}n = e^{-n\tau}(1 - n\tau) = 0$), and then falls back towards zero as the counter spends almost all its time paralysed. That is the Geiger counter whose rate fell as it approached the source. At low rate both models agree, $m \approx n(1 - n\tau)$, as the left panel of Figure 2.20 shows.

Watch out: a recorded rate can have two true rates

For a paralysable counter, every recorded rate below $1/(e\tau)$ is produced by *two* true rates, one below $1/\tau$ and one above. The measurement alone cannot tell which. Change the rate in a known way (move the source, add a second source) and see whether the recorded rate goes up or down. And a recorded rate above $1/(e\tau)$ is impossible for a paralysable counter: if one is seen, the model or the value of τ is wrong.

The scatter of the recorded counts. Dead time removes precisely the hits that come soon after another hit, so the recorded stream is more regular than the true one (Figure 2.18, bottom). How much more regular?

The recorded stream of either model is a **renewal process** [the gaps between successive recorded hits are independent and all have the same distribution]. After a recorded hit, the waiting time to the next recorded hit depends only on subsequent arrivals. By input P1 these are independent of the past, so each recorded hit starts the same waiting-time experiment afresh.

For a renewal process the count's scatter follows from the gaps. Let the gap I between recorded hits have mean m_I and variance σ_I^2. Count in a window $T \gg m_I$, and let S_k be the time of the k-th recorded hit, the sum of k independent gaps. The count reaches k when S_k reaches T. Write $S_k = km_I + \delta S$, with $\operatorname{Var}\delta S = k\sigma_I^2$. Suppose the gaps happen to add up to a little more time than average, $\delta S > 0$. Then fewer hits fit into T. Each missing hit is worth one mean gap, so the count changes by $\delta N \approx -\delta S/m_I$. Hence

$$\mathbb{E}\,N \overset{(1)}{=} \frac{T}{m_I}, \qquad \operatorname{Var} N \overset{(2)}{=} \frac{\operatorname{Var}\delta S}{m_I^2} \overset{(3)}{=} \frac{k\,\sigma_I^2}{m_I^2} \overset{(4)}{=} \frac{T\sigma_I^2}{m_I^3}, \qquad F = \frac{\operatorname{Var} N}{\mathbb{E}\,N} \overset{(5)}{=} \frac{\sigma_I^2}{m_I^2}. \tag{2.29}$$

(1) On average one hit per mean gap.
(2) $\delta N \approx -\delta S/m_I$, and the sign does not affect the variance.
(3) The k gaps are independent, so their variances add.
(4) $k \approx \mathbb{E}\,N = T/m_I$.
(5) Divide the two results: T cancels.

For a long window, the Fano factor of the counts equals the square of the relative spread of the gaps, σ_I/m_I. Check on the Poisson stream itself: exponential gaps have $\sigma_I = m_I = 1/\lambda$, so $F = 1$.

For the *non-paralysable* counter, a gap between recorded hits is the dead period τ followed by a fresh exponential wait for the next true hit (memorylessness, Section 2.6): $I = \tau + G$ with $G \sim$ Exp of rate n. So $m_I = \tau + 1/n$ and $\sigma_I^2 = 1/n^2$, and

$$m \overset{(1)}{=} \frac{1}{m_I} = \frac{n}{1+n\tau}, \qquad F \overset{(2)}{=} \frac{1/n^2}{(\tau+1/n)^2} = \frac{1}{(1+n\tau)^2}. \tag{2.30}$$

(1) The recorded rate is one over the mean gap; this re-derives (2.27) by a second route.
(2) Eq. (2.29) with the mean and variance of $\tau + G$ (a constant shift does not change a variance).

For the *paralysable* counter the gap is harder. After a recorded hit, true hits keep arriving, separated by true gaps $g_1, g_2, \ldots$ (independent, exponential with rate n). The next recorded hit is the first true hit whose own preceding gap is at least τ. So $I = g_1 + \cdots + g_J$, where J is the index of the first gap with $g_j \geq \tau$. The cleanest tool is the moment generating function $M_I(s) = \mathbb{E}\,e^{sI}$ (Chapter 1). Split one exponential gap by whether it is short or long:

$$a(s) = \mathbb{E}\big[e^{sg}\mathbb{1}\{g < \tau\}\big] = \frac{n}{n-s}\big(1 - e^{-(n-s)\tau}\big), \qquad b(s) = \mathbb{E}\big[e^{sg}\mathbb{1}\{g \geq \tau\}\big] = \frac{n}{n-s}e^{-(n-s)\tau},$$

both from $\int ne^{-(n-s)g}\,\mathrm{d}g$ over $[0,\tau)$ and $[\tau,\infty)$ (for $s < n$). Then

$$M_I(s) \overset{(1)}{=} \sum_{j=1}^{\infty} a(s)^{j-1}b(s) \overset{(2)}{=} \frac{b(s)}{1-a(s)} \overset{(3)}{=} \frac{n\,e^{-(n-s)\tau}}{(n-s)-n+n\,e^{-(n-s)\tau}} \overset{(4)}{=} \frac{1}{1-\frac{s}{n}\,e^{(n-s)\tau}}.$$

(1) $J = j$ means $j-1$ short gaps followed by one long gap. The gaps are independent, so $\mathbb{E}\,e^{sI}$ on that event factorises into $a^{j-1}b$, and we sum over j.

(2) Geometric series, valid for s near 0 where $a(s) < 1$.
(3) Insert a and b and multiply top and bottom by $n - s$.
(4) The denominator is $ne^{-(n-s)\tau} - s$; divide top and bottom by $ne^{-(n-s)\tau}$.

The mean and variance come from the **cumulant generating function** $K_I(s) = \ln M_I(s)$, whose first two Taylor coefficients are $K_I(s) = \mathbb{E}\,I\,s + \operatorname{Var} I\,s^2/2 + O(s^3)$. With $Q = e^{n\tau}$ and $u = \frac{s}{n}e^{(n-s)\tau} = \frac{sQ}{n}(1 - s\tau + O(s^2))$,

$$K_I(s) \overset{(1)}{=} -\ln(1-u) \overset{(2)}{=} u + \frac{u^2}{2} + O(s^3) \overset{(3)}{=} \frac{Q}{n}\,s + \Big(\frac{Q^2}{2n^2} - \frac{\tau Q}{n}\Big)s^2 + O(s^3). \tag{2.31}$$

(1) The logarithm of the result above.
(2) Taylor series $-\ln(1-u) = u + u^2/2 + \ldots$; u is of order s, so $u^3 = O(s^3)$.
(3) Insert $u = \frac{sQ}{n} - \frac{s^2\tau Q}{n} + O(s^3)$ and $u^2 = \frac{s^2Q^2}{n^2} + O(s^3)$.

Reading off, $\mathbb{E}\,I = e^{n\tau}/n$, so $m = 1/\mathbb{E}\,I = ne^{-n\tau}$, which re-derives (2.28). And $\operatorname{Var} I = Q^2/n^2 - 2\tau Q/n$, so by (2.29)

$$F = \frac{\operatorname{Var} I}{(\mathbb{E}\,I)^2} = 1 - 2n\tau\,e^{-n\tau}. \tag{2.32}$$

Both counters are underdispersed. At low rate both give $F \approx 1 - 2n\tau$. At $n\tau = 1$ the paralysable counter is at its most regular, $F = 1 - 2/e = 0.26$. At very high rate it becomes nearly Poisson again. The reason: the few recorded hits are then the rare long gaps of the true stream, and those are almost independent of each other.

Example: correcting a Geiger counter. A Geiger–Müller tube has $\tau = 200\ \mu$s and records $m = 1500$ counts per second, so $m\tau = 0.3$. If it is non-paralysable, (2.27) gives

$$n \overset{(1)}{=} \frac{m}{1 - m\tau} = \frac{1500}{0.7} = 2143\ \mathrm{s}^{-1}.$$

(1) Dead-time correction (2.27).

So 30% of the true hits were lost ($m\tau$ is exactly the dead fraction of the time). If the tube is paralysable, we must solve $n\tau e^{-n\tau} = m\tau = 0.3$. Numerically the two roots are $n\tau = 0.489$ and $n\tau = 1.781$, that is $n = 2447\ \mathrm{s}^{-1}$ or $n = 8907\ \mathrm{s}^{-1}$. This is the two-root ambiguity of a paralysable counter: one recorded rate, two possible true rates, and only a change of the true rate in a known way can tell them apart. The largest rate this counter could ever record is $1/(e\tau) = 1839\ \mathrm{s}^{-1}$, so a reading of $2000\ \mathrm{s}^{-1}$ would already be impossible for it.

Example: the error bar of a corrected rate. After the correction the rate is right on average, but the corrected count is *not* Poisson, and its error bar is not $\sqrt{\text{corrected counts}}$. For the non-paralysable counter, let M be the number recorded in a long time T, with $\mathbb{E}\,M = mT$ and $\operatorname{Var} M = F\,mT$. The estimate is $\hat{n} = g(\hat{m})$, with $\hat{m} = M/T$ and $g(m) = m/(1 - m\tau)$. For small fluctuations a smooth function of an estimate varies by the derivative times the fluctuation, $\delta\hat{n} = g'(m)\,\delta\hat{m}$ (the **delta method**, Chapter 3):

$$\operatorname{Var}\hat{n} \overset{(1)}{=} g'(m)^2 \operatorname{Var}\hat{m} \overset{(2)}{=} \frac{1}{(1-m\tau)^4}\,\frac{F\,m}{T} \overset{(3)}{=} (1+n\tau)^4\,\frac{1}{(1+n\tau)^2}\,\frac{n}{(1+n\tau)\,T} \overset{(4)}{=} \frac{n\,(1+n\tau)}{T}. \tag{2.33}$$

(1) Delta method.
(2) $g'(m) = \frac{(1-m\tau)+m\tau}{(1-m\tau)^2} = \frac{1}{(1-m\tau)^2}$, and $\operatorname{Var}\hat{m} = \operatorname{Var} M/T^2 = Fm/T$.
(3) $1 - m\tau = 1/(1+n\tau)$ from (2.27), together with F from (2.30) and $m = n/(1+n\tau)$.
(4) Collect the powers of $(1+n\tau)$: $4 - 2 - 1 = 1$.

Had we counted the true hits with a perfect counter, $\mathrm{Var}(\hat{n}) = n/T$ (Poisson). So dead time inflates the variance of the corrected rate by the factor $1 + n\tau$: the corrected count $\hat{n}T$ has variance $(1 + n\tau)$ times its mean. The correction removes the bias, but the information lost in the dead periods does not come back. For the Geiger counter above, $1 + n\tau = 1.43$: the corrected rate is $\sqrt{1.43} = 1.2$ times more uncertain than naive Poisson error bars claim.

Example: sub-Poissonian light, or dead time? In quantum optics the photocount statistics of a light beam are summarised by Mandel's parameter $Q = F - 1 = (\mathrm{Var}\, N - \mathbb{E}\, N)/\mathbb{E}\, N$. Coherent laser light gives Poisson photocounts, $Q = 0$. Thermal light is bunched, $Q > 0$: for a single mode the photocount is geometric (the Bose–Einstein distribution), and for many modes it is negative binomial, the gamma–Poisson mixture of Section 2.7 with the fluctuating intensity playing the role of Λ. A measured $Q < 0$, called *sub-Poissonian* light, has no classical explanation. A classical intensity can only add variance: $\mathrm{Var}\,\Lambda \geq 0$ in (2.25). So $Q < 0$ is a signature of non-classical light, such as that from single-photon sources. But a non-paralysable photon counter with dead time τ records $F = 1/(1 + n\tau)^2$ even for a perfect laser, by (2.30), and that is $Q \approx -2n\tau$. At $n = 10^6$ s^{-1} and $\tau = 50$ ns this is $Q \approx -0.1$, a fake non-classical signal. Any claim of $Q < 0$ must first remove the detector's own dead time. This is why such experiments use two detectors behind a beam splitter and look at coincidences, so that neither detector has to record two photons in quick succession.

2.8.4 Python: a counter that goes blind

Listing 2.9: A Poisson stream of true hits, and the two dead-time models (`code/ch02/08_deadtime_gauss.py`)

```
28 def true_hits(n, duration):
29     """Arrival times of a Poisson stream of rate n on [0, duration): exponential gaps."""
30     gaps = -np.log(rng.uniform(size=int(n * duration + 10 * np.sqrt(n * duration) + 10))) /
       n
31     t = np.cumsum(gaps)
32     return t[t < duration]
33
34
35 def record(t, paralysable):
36     """Times the counter records. Non-paralysable: blind for tau after each RECORDED hit.
37     Paralysable: blind for tau after EVERY hit, so a hit is recorded only if the
38     previous true hit is at least tau earlier."""
39     if paralysable:
40         gap_before = np.diff(t, prepend=-np.inf)
41         return t[gap_before >= tau]
42     kept, last = [], -np.inf
43     for ti in t:                        # a plain loop: the state is 'when did the last
       recorded hit start?'
44         if ti - last >= tau:
45             kept.append(ti)
46             last = ti
47     return np.array(kept)
```

The function `true_hits` builds the true stream exactly as in Section 2.6: exponential gaps $-\ln U/n$ from uniform numbers U, summed into arrival times. The `record` function implements the two models of Figure 2.19 literally. The paralysable rule needs no memory beyond the previous true hit, so it is one vectorised comparison of each gap with τ (lines 40–41). The non-paralysable rule needs to remember when the last *recorded* hit happened, so it is a plain loop that keeps that one number, `last` (lines 42–47). The script then cuts a 40 s stream at $n = 5 \times 10^4$ s^{-1} into 4000 windows of 10 ms (so $n\tau = 0.25$ with $\tau = 5$ μs), counts the recorded

hits per window, and repeats the rate measurement for a range of true rates.

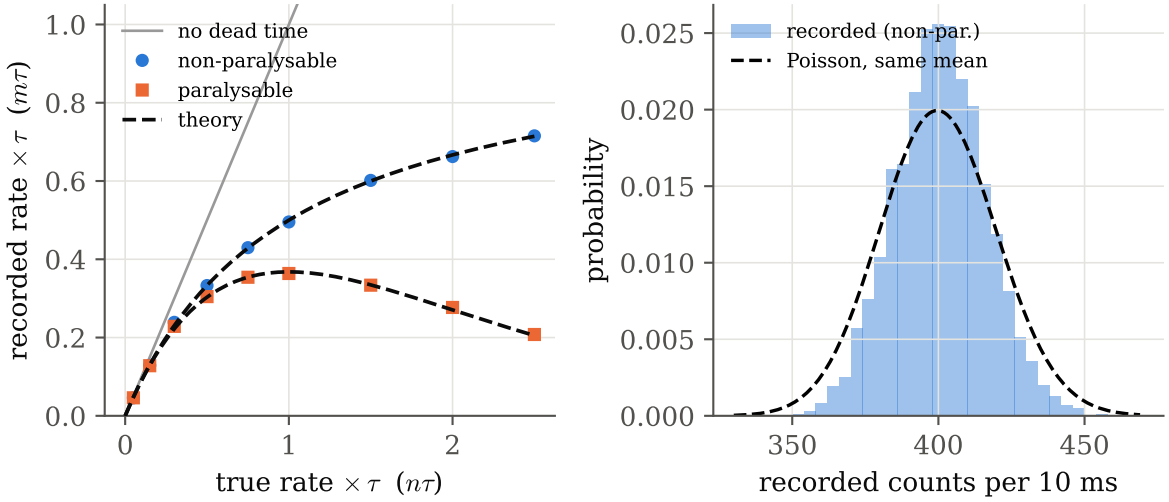


Figure 2.20: Left: recorded rate against true rate, both in units of $1/\tau$. The simulated counters (dots and squares) sit on the derived curves $n\tau/(1+n\tau)$ and $n\tau e^{-n\tau}$ (dashed). The grey line is a perfect counter. The non-paralysable counter saturates at one hit per dead time; the paralysable one peaks at $n\tau = 1$ and then falls. Right: the recorded counts per 10 ms window at $n\tau = 0.25$ (bars) form a visibly narrower histogram than a Poisson distribution with the same mean (dashed).

The numbers it produced are these. The true hits give 500.3 counts per window on average (expected 500) with $\hat{F} = 0.9597$; by (2.26) the Poisson scatter of $\hat{F}$ is 0.02236, so $\hat{F} - 1$ is -1.8 times this scatter. The non-paralysable counter records a mean of 400.1 per window, against 400 from (2.27), and its Fano factor is 0.6156, against $1/(1+n\tau)^2 = 0.64$. The paralysable counter records 389.5 (theory 389.4) with $\hat{F} = 0.5861$ (theory $1 - 2n\tau e^{-n\tau} = 0.6106$). In units of their own scatter (for $F \neq 1$ it is about $F\sqrt{2/(k-1)}$), the two dead-time Fano factors minus their predictions are -1.7 and -1.8. The three numbers are not independent: all three come from the *same* stream of true hits, so a chance fluctuation of that stream shows up in all three. Correcting the non-paralysable mean with (2.27) gives $\hat{n} = 5.002 \times 10^4\ \mathrm{s}^{-1}$ for the true rate of $5 \times 10^4\ \mathrm{s}^{-1}$, off by 0.5 times its statistical error $\sqrt{n(1+n\tau)/T} \approx 40\ \mathrm{s}^{-1}$ from (2.33). The right panel of Figure 2.20 is the underdispersion made visible.

How research uses it. Here the simulation *verified* formulas we had derived. Real detectors rarely match either idealised model: the dead time can depend on the pulse height, several stages of electronics can each have their own τ, and afterpulsing adds clustering on top. For them the analytic route stops, and the same few lines of code become the instrument: simulate the true stream, pass it through a model of the electronics, and histogram the recorded counts. That histogram is the sampling distribution $P(S \mid \theta)$ of the recorded count, and it goes into the likelihood (Chapter 3). In a collider analysis, the Fano factor of event counts in stable running periods is a routine check that the luminosity and trigger were steady; in a survey, the excess over $F = 1$ is the clustering signal itself.

2.8.5 Large λ: the Poisson count becomes Gaussian (a preview)

In Figure 2.9 the Poisson pmf became more symmetric as λ grew. For a large mean, a Poisson variable can be treated as a continuous Gaussian variable [Cowan §2.2, p. 29]. Cumulants show why most quickly. The Poisson mgf is $M_X(t) = \exp[\lambda(e^t - 1)]$, eq. (2.13), so its cumulant

generating function is

$$K_X(t) = \ln M_X(t) \overset{(1)}{=} \lambda(e^t - 1) \overset{(2)}{=} \lambda \sum_{r=1}^{\infty} \frac{t^r}{r!} \quad \Longrightarrow \quad \kappa_r(X) \overset{(3)}{=} \lambda \quad \text{for every } r \geq 1. \tag{2.34}$$

(1) Logarithm of the Poisson mgf.
(2) Taylor series of e^t with the $r = 0$ term cancelled by the -1.
(3) By definition the r-th cumulant κ_r is $r!$ times the coefficient of t^r in K_X. The first three cumulants are the mean, the variance and the third central moment.

Now standardise, $Z = (X - \lambda)/\sqrt{\lambda}$. Since $K_{aX+b}(t) = bt + K_X(at)$, shifting by b changes only κ_1 and scaling by a multiplies κ_r by a^r. With $a = 1/\sqrt{\lambda}$,

$$\kappa_1(Z) = 0, \qquad \kappa_2(Z) = 1, \qquad \kappa_3(Z) = \frac{\lambda}{\lambda^{3/2}} = \frac{1}{\sqrt{\lambda}}, \qquad \kappa_r(Z) = \lambda^{1-r/2} \xrightarrow{\lambda\to\infty} 0 \quad (r \geq 3).$$

So $K_Z(t) \to t^2/2$, which is the cumulant generating function of $\mathcal{N}(0, 1)$: *Poisson*(λ) *is approximately* $\mathcal{N}(\lambda, \lambda)$ *for large* λ. The leading correction is the skewness $\kappa_3(Z) = 1/\sqrt{\lambda}$, which shrinks only slowly: at $\lambda = 25$ it is still 0.2. The next part derives this limit properly (de Moivre–Laplace for the binomial, and the central limit theorem), with the corrections. It also explains why the limit holds. For integer λ, a Poisson(λ) count is the sum of λ independent Poisson(1) counts, by the sum rule of Section 2.5, and sums of many independent terms tend to a Gaussian.

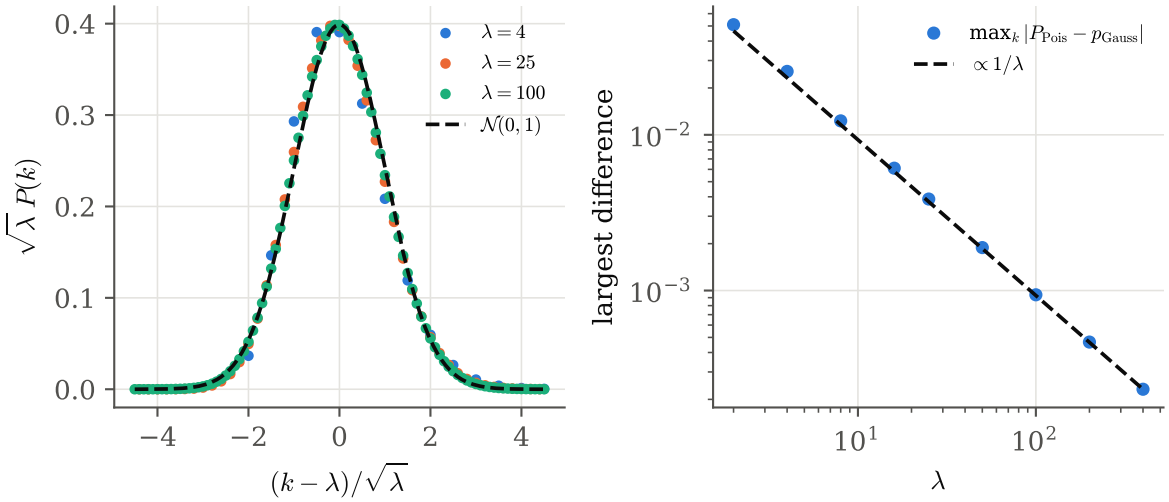


Figure 2.21: Left: Poisson pmfs for $\lambda = 4$, 25 and 100, standardised so that the horizontal axis is the distance from the mean in standard deviations, and rescaled by $\sqrt{\lambda}$ so that they are comparable with a density. They crowd onto the standard Gaussian (dashed); at $\lambda = 4$ the points are still lopsided, higher on the left of the peak and with a longer right tail. Right: the largest pointwise difference between the Poisson pmf and the Gaussian density falls like $1/\lambda$ (dashed line).

Figure 2.21 shows the convergence. The largest difference $\max_k |P(k) - p_{\text{Gauss}}(k)|$ is 0.02554 at $\lambda = 4$, 0.003868 at $\lambda = 25$, 9.388×10^{-4} at $\lambda = 100$ and 2.323×10^{-4} at $\lambda = 400$. The $1/\lambda$ law has a simple reason: the peak height is about $1/\sqrt{2\pi\lambda}$, and the skewness correction to it is a fraction of order $\kappa_3(Z) = 1/\sqrt{\lambda}$ of that height, so the absolute error is of order $1/\lambda$.

Example: a 4σ excess on a background of 25. A search region expects $b = 25$ background events, with Poisson standard deviation 5, and we observe 45, which is $(45 - 25)/5 = 4$ standard deviations above the background. How often does background alone give 45 or more? The

Gaussian answer is $\mathrm{P}(Z \geq 4) = 3.167 \times 10^{-5}$. The exact Poisson answer, $\sum_{k \geq 45} e^{-25} 25^k / k!$, is 2.001×10^{-4}: 6.319 times larger. Even the **continuity correction** [treating the integer 45 as the interval from 44.5 up], which gives 4.810×10^{-5}, is still wrong by a factor of four. Why so wrong? The Gaussian is good near the peak, where the $1/\lambda$ errors of Figure 2.21 are small in absolute terms. In the tail both probabilities are tiny, and a small absolute error is a large relative one. The Poisson right tail is heavier than the Gaussian one, because the Poisson is positively skewed. So the Gaussian *understates* the chance of an upward fluctuation and overstates the significance of an excess. Tail probabilities, p-values and discovery claims (Chapter 4) must use the exact Poisson distribution.

2.8.6 Choosing a counting distribution

Every counting distribution came from a mechanism, so recognising the mechanism is how we choose the distribution. The table collects them, with the Fano factor $F = \mathrm{Var}\, N / \mathbb{E}\, N$ as the quick check against data.

Taxonomy: which counting distribution, when

mechanism	distribution	mean	Fano F
one yes/no trial	Bernoulli(p)	p	$1 - p$
successes in n independent trials, same p	Binomial(n, p)	np	$1 - p$
N events sorted into bins, total fixed	Multinomial($N, \boldsymbol{p}$)	Np_j per bin	$1 - p_j$ per bin
many trials, rare successes; or a stream at constant rate with no memory	Poisson(λ)	λ	1
Poisson with a gamma-distributed rate, or clumped events	NegBin	μ	$1 + \mu/\kappa$
counter blind for τ after each recorded hit	renewal count	$nT/(1 + n\tau)$	$(1 + n\tau)^{-2}$
trials until the first / r-th success	Geometric / NegBin	$1/p, r/p$	—
time to the first / n-th event of a stream	Exponential / Gamma	$1/\lambda, n/\lambda$	—

The Fano column is the one to remember when data arrive. A value near 1, judged against $\sqrt{2/(k-1)}$, supports a Poisson model. A value clearly above 1 points to a fluctuating rate or clumping, and the negative binomial is the standard next model. A value clearly below 1 points to dead time, pile-up rejection or a finite pool, and calls for a model of the detector. For large means the Poisson itself may be replaced by a Gaussian with variance equal to the mean, but only near the peak, never in the tails.

Recap: Part summary: counting distributions

- **Bernoulli** (one trial): $p^x(1-p)^{1-x}$, mean p, variance $p(1-p)$. Many independent trials give the product $p^z(1-p)^{N-z}$, which depends only on the number of successes.
- **Binomial** (successes in n successive trials): count the $\binom{n}{k}$ paths, each of probability $p^k(1-p)^{n-k}$; mean np, variance $np(1-p)$; sums with the same p stay binomial.

- **Multinomial** (histogram with fixed total): marginals binomial, covariance $-Np_ip_j$ between bins; with a Poisson total the bins become independent Poisson counts.
- **Poisson**: the $n \to \infty$, $np = \lambda$ limit of the binomial, and independently the count of a constant-rate, memoryless stream (from the ladder of differential equations $\mathrm{d}P_n/\mathrm{d}t = \lambda(P_{n-1} - P_n)$). Mean = variance = λ. Closed under sums, thinning and binning.
- **Waiting times**: first event exponential (memoryless), n-th event Gamma$(n, 1/\lambda)$ (Erlang); discrete versions geometric and negative binomial.
- **Failures**: a fluctuating rate or clumping gives $F > 1$ (negative binomial); dead time gives $F < 1$, with recorded rate $n/(1 + n\tau)$ or $ne^{-n\tau}$ and $F = \sigma_I^2/m_I^2$ from the gaps; the dead-time correction removes the bias but inflates the variance by $1 + n\tau$.
- **Large** λ: all Poisson cumulants equal λ, so the standardised count tends to $\mathcal{N}(0, 1)$ with skewness $1/\sqrt{\lambda}$; use the exact Poisson in the tails.

2.9 The Gaussian, first way: the shape of a binomial with many trials

The same bell-shaped curve describes a calorimeter energy, a timing residual, a CMB temperature and a strain sample. None of these is a count: the counting distributions of the first half of the chapter live on the integers, and these quantities do not. There are three separate routes to the bell curve, and each route tells us *when* to trust it. The oldest route starts from a count after all. Take a binomial distribution with a very large number of trials and ask what shape it takes.

Ten thousand photo-electrons. A scintillator bar is hit by a cosmic muon. It produces $n = 10\,000$ photons, and each photon independently reaches the photomultiplier and makes a photo-electron with probability $p = 0.3$. The number of photo-electrons k is Binomial(n, p) (see the counting part of this chapter). To compute $\mathrm{P}(k \le 2950)$ exactly we would add 2951 terms, each containing factorials of numbers near 10^4. Nobody does that. Instead everyone replaces the binomial by a smooth bell curve with mean $np = 3000$ and standard deviation $\sqrt{np(1-p)} \approx 45.8$.

That replacement raises four questions. Which bell curve, exactly? Why a bell and not some other shape? How large must n be? And how do we correct for the fact that k is an integer while the curve is smooth? The large-n limit of the binomial answers all four.

Zooming in on the peak of a mountain. The Gaussian is what *every* smooth, single-peaked, sharply concentrated distribution looks like near its maximum. Here is why. For large n the binomial pmf is a tall, narrow mountain centred at np, and its width grows only like $\sqrt{n}$. Plot the *logarithm* of any smooth mountain: near its summit it looks like a downward parabola. The exponential of a parabola is a Gaussian. This is the same move as the harmonic approximation in mechanics: expand the potential to second order about its minimum and the motion becomes simple harmonic. Here we expand $\ln \mathrm{P}(k)$ to second order about its maximum.

The plan.

1. Which quantity do we want? The probability of k successes in n trials when n is large.

2. What makes it hard? Factorials of large numbers. So first find a good formula for $n!$ (Stirling).
3. Where is the probability concentrated? Near $k = np$, in a window of width $\sqrt{np(1-p)}$.
4. What does $\ln \mathrm{P}(k)$ look like in that window? Expand it in powers of the distance from the peak and keep what survives as $n \to \infty$.
5. How do we turn a statement about integers into one about a continuous variable, and what does that cost? (The continuity correction.)

2.9.1 The Gaussian density and what its two parameters do

Working definition 2.13 (Gaussian (normal) distribution)

A continuous random variable X has a **Gaussian** or **normal** distribution with parameters μ and σ^2, written $X \sim \mathcal{N}(\mu, \sigma^2)$, if its density [probability per unit x, so it has units of $1/x$] is

$$p(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \qquad -\infty < x < \infty, \tag{2.35}$$

with $\mu \in \mathbb{R}$ and $\sigma > 0$. The case $\mu = 0$, $\sigma = 1$ is the **standard normal** $Z \sim \mathcal{N}(0,1)$, with density $\varphi(z) = e^{-z^2/2}/\sqrt{2\pi}$ and cumulative distribution function [the probability of landing at or below z] $\Phi(z) = \int_{-\infty}^{z} \varphi(z')\,\mathrm{d}z'$. *Operational test:* standardise, $z = (x-\mu)/\sigma$, and compare the histogram of z with φ (or the ordered z values with the quantiles Φ^{-1}).

Three facts make the Gaussian usable. We check each one.

It is normalised. The function $e^{-z^2/2}$ has no elementary antiderivative, so the normalisation needs a trick: square the integral and go to polar coordinates [CasellaBerger p. 84–85]. Let $I = \int_{-\infty}^{\infty} e^{-z^2/2}\mathrm{d}z$. Then

$$\begin{aligned} I^2 &\overset{(1)}{=} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(u^2+v^2)/2}\,\mathrm{d}u\,\mathrm{d}v \\ &\overset{(2)}{=} \int_{0}^{2\pi}\int_{0}^{\infty} e^{-r^2/2}\, r\,\mathrm{d}r\,\mathrm{d}\phi \\ &\overset{(3)}{=} 2\pi\left[-e^{-r^2/2}\right]_0^{\infty} \\ &= 2\pi. \end{aligned} \tag{2.36}$$

(1) The two copies of the integral use independent dummy variables u and v, and a product of two integrals is a double integral over the plane.
(2) Polar coordinates $u = r\cos\phi$, $v = r\sin\phi$: then $u^2 + v^2 = r^2$ and the area element is $\mathrm{d}u\,\mathrm{d}v = r\,\mathrm{d}r\,\mathrm{d}\phi$.
(3) The ϕ integral gives 2π. The extra factor r makes the r integral elementary, since $\frac{\mathrm{d}}{\mathrm{d}r}\left(-e^{-r^2/2}\right) = r\,e^{-r^2/2}$.

So $I = \sqrt{2\pi}$, which is the constant in φ. Substituting $z = (x-\mu)/\sigma$, $\mathrm{d}x = \sigma\,\mathrm{d}z$, gives the constant $1/\sqrt{2\pi\sigma^2}$ in (2.35). A by-product: with $w = z^2/2$ the half-line integral becomes $\int_0^{\infty} w^{-1/2}e^{-w}\mathrm{d}w = \Gamma(\tfrac{1}{2})$, so

$$\Gamma(\tfrac{1}{2}) = \sqrt{\pi}, \tag{2.37}$$

where $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}\mathrm{d}t$ is the Gamma function, with $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$ (integrate by parts) and $\Gamma(n+1) = n!$ for integers. We need all three facts for χ^2 later.

μ is the mean and σ^2 the variance. With $x = \mu + \sigma z$,

$$\mathbb{E}(X) \overset{(1)}{=} \mu + \sigma \int_{-\infty}^{\infty} z\,\varphi(z)\,\mathrm{d}z \overset{(2)}{=} \mu,$$
$$\mathrm{Var}(X) \overset{(3)}{=} \sigma^2 \int_{-\infty}^{\infty} z^2\varphi(z)\,\mathrm{d}z \overset{(4)}{=} \sigma^2\Big\{\big[-z\varphi(z)\big]_{-\infty}^{\infty} + \int_{-\infty}^{\infty}\varphi(z)\,\mathrm{d}z\Big\} \overset{(5)}{=} \sigma^2. \tag{2.38}$$

(1) Change variables and use $\int \varphi = 1$ for the constant term.
(2) The integrand $z\varphi(z)$ is odd, so the integral over a symmetric range vanishes.
(3) $\mathrm{Var}(X) = \mathbb{E}(X-\mu)^2$ and $(x-\mu)^2 = \sigma^2 z^2$.
(4) Integrate by parts with $u = z$ and $\mathrm{d}v = z\varphi(z)\mathrm{d}z$, whose integral is $v = -\varphi(z)$ because $\varphi'(z) = -z\varphi(z)$.
(5) The boundary term vanishes because φ decays faster than any power, and the remaining integral is 1.

So the two parameters are exactly the mean and the variance [Cowan (2.23)–(2.24)]. This is special: for most distributions the parameters are *not* the mean and variance (the log-normal of Section 2.15 is a warning example).

Standardising reduces everything to one table. If $X \sim \mathcal{N}(\mu, \sigma^2)$ then $Z = (X-\mu)/\sigma \sim \mathcal{N}(0,1)$ and, conversely, $X = \mu + \sigma Z$ [Wasserman p. 28]. Hence

$$\mathrm{P}(a < X < b) = \Phi\!\left(\frac{b-\mu}{\sigma}\right) - \Phi\!\left(\frac{a-\mu}{\sigma}\right). \tag{2.39}$$

Two numbers are worth knowing by heart. First, the probability within 1, 2, 3 standard deviations of the mean is 0.6827, 0.9545, 0.9973 (Figure 2.22). Second, "one sigma" can be read off a plot by eye: it is where the curve stops bending down and starts bending up. To see this, note that the density has its maximum at $x = \mu$, and its second derivative $\varphi''(z) = (z^2-1)\varphi(z)$ vanishes at $z = \pm 1$. So the inflection points sit at $x = \mu \pm \sigma$.

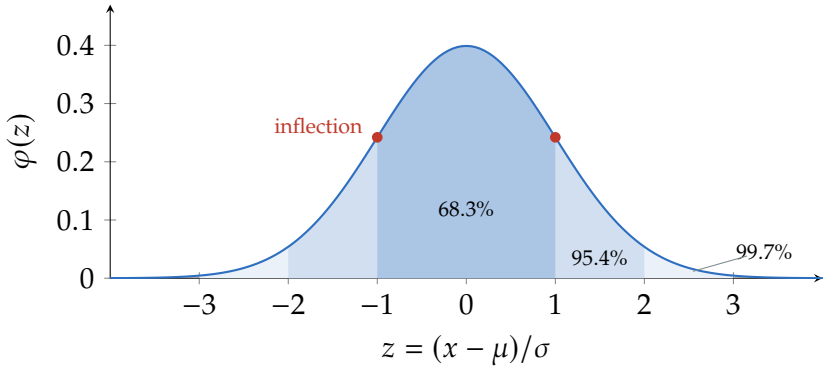


Figure 2.22: The standard normal density. The darkest band ($|z| < 1$) holds 68.3% of the probability, the band out to $|z| < 2$ holds 95.4%, and out to $|z| < 3$ holds 99.7%. The red dots are the inflection points at $z = \pm 1$: one standard deviation is where the curve changes from bending down to bending up.

Example: Reading the table both ways. Let $X \sim \mathcal{N}(3, 5)$. The second argument is the *variance*, so $\sigma = \sqrt{5}$. A table of Φ answers two kinds of question. The forward question gives a cut and asks for a probability: what is $\mathrm{P}(X > 1)$?

$$\mathrm{P}(X > 1) \overset{(1)}{=} 1 - \mathrm{P}(X < 1) \overset{(2)}{=} 1 - \Phi\left(\frac{1-3}{\sqrt{5}}\right) = 1 - \Phi(-0.8944) \overset{(3)}{=} 0.81.$$

(1) Complement rule. For a continuous variable $\mathrm{P}(X = 1) = 0$, so $<$ and $\leq$ do not matter.
(2) Standardise with (2.39).
(3) From a table or `scipy.stats.norm.cdf`: $\Phi(-0.8944) = 0.1855$.

The quantile question goes the other way. It gives a probability and asks for the cut: which q has $\mathrm{P}(X < q) = 0.2$?

$$0.2 = \mathrm{P}(X < q) \overset{(1)}{=} \Phi\left(\frac{q-3}{\sqrt{5}}\right) \implies \frac{q-3}{\sqrt{5}} \overset{(2)}{=} \Phi^{-1}(0.2) = -0.8416$$
$$\implies q \overset{(3)}{=} 3 - 0.8416\sqrt{5} = 1.1181.$$

(1) Standardise.
(2) Φ is strictly increasing, so it can be inverted. Φ^{-1} is `scipy.stats.norm.ppf`.
(3) Solve the linear equation for q.

These are the two numbers of [Wasserman p. 29, Example 2.17].

Watch out: $\mathcal{N}(\mu, \sigma^2)$ or $\mathcal{N}(\mu, \sigma)$?

AoS, Cowan and CasellaBerger write the *variance* as the second argument. Lambert and most software (`scipy.stats.norm(loc, scale)`, `numpy.random.normal(loc, scale)`) take the *standard deviation*. $\mathcal{N}(3, 5)$ in AoS is `norm(3, sqrt(5))` in scipy. Always check which one a formula or a function expects.

2.9.2 Stirling's formula from Laplace's method

The binomial is built from factorials, so we need a good formula for $n!$ when n is large. That formula is Stirling's:

$$n! \simeq \sqrt{2\pi n}\, n^n e^{-n} \qquad (n \to \infty), \qquad \text{i.e.} \qquad \ln n! = n \ln n - n + \tfrac{1}{2}\ln(2\pi n) + O(1/n). \tag{2.40}$$

It comes from the same zoom-in on a peak, now applied to an integral. *The starting point.* The Gamma function gives $n!$ as an integral, $n! = \Gamma(n+1) = \int_0^\infty t^n e^{-t} \mathrm{d}t$, and for large n its integrand is sharply peaked. *Where is the peak, and how sharp is it?* Write the integrand as $e^{f(t)}$ with $f(t) = n \ln t - t$. The slope $f'(t) = n/t - 1$ vanishes at $t = n$, so the peak is at $t = n$. The curvature $f''(t) = -n/t^2$ equals $-1/n$ there. *Only then the integral.* Expanding f to second order about its maximum,

$$\begin{aligned} n! &\overset{(1)}{=} \int_0^\infty \exp\Big[f(n) + f'(n)(t-n) + \tfrac{1}{2}f''(n)(t-n)^2 + \dots\Big]\mathrm{d}t \\ &\overset{(2)}{=} n^n e^{-n} \int_0^\infty \exp\Big[-\frac{(t-n)^2}{2n} + \dots\Big]\mathrm{d}t \\ &\overset{(3)}{=} n^n e^{-n} \int_{-\infty}^\infty e^{-s^2/(2n)}\mathrm{d}s\ \big(1 + O(1/n)\big) \\ &\overset{(4)}{=} \sqrt{2\pi n}\, n^n e^{-n}\big(1 + O(1/n)\big). \end{aligned} \tag{2.41}$$

(1) Taylor expansion of f about $t = n$ inside the exponential.
(2) $f(n) = n \ln n - n$ gives $e^{f(n)} = n^n e^{-n}$, the linear term vanishes because $t = n$ is the maximum, and $\frac{1}{2} f''(n) = -1/(2n)$.
(3) Substitute $s = t - n$. The integrand is negligible beyond a few multiples of $\sqrt{n}$ from the peak, so extending the lower limit from $s = -n$ to $-\infty$ adds only a Gaussian tail of order $e^{-n/2}$. The cubic and higher terms of the expansion give the relative correction of order $1/n$ (it is $1/(12n)$).
(4) The Gaussian integral (2.36) with $\sigma^2 = n$.

This is **Laplace's method**: a sharply peaked integrand $e^{f(t)}$ is replaced by a Gaussian with the same peak height and the same curvature. We will meet it again as the Gaussian (Laplace) approximation to a posterior in Chapter 5. Figure 2.23 shows how good the approximation already is at $n = 10$.

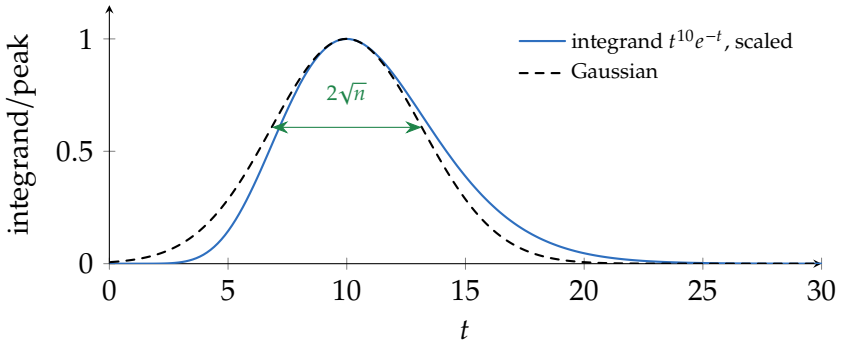


Figure 2.23: Laplace's method for $10! = \int_0^\infty t^{10} e^{-t} \mathrm{d}t$. The integrand (blue) is scaled to peak height one at $t = n = 10$. The dashed Gaussian has the same peak and the same curvature there, so its width is $\sqrt{n}$. The two areas differ by less than one percent, and the mismatch is the slight skew of the blue curve, which the $1/(12n)$ correction describes.

Stirling's formula is accurate even for small n. The script below finds $n!/\text{Stirling} = 1.0844$ at $n = 1$ and 1.00837 at $n = 10$, in agreement with $1 + 1/(12n)$ (Figure 2.25a). Even 1! is off by only eight percent.

2.9.3 The de Moivre–Laplace limit

Key idea: de Moivre–Laplace

Let $k \sim \text{Binomial}(n, p)$ and $q = 1 - p$. For large n and k within a few $\sqrt{npq}$ of np,

$$\mathrm{P}(k) = \binom{n}{k} p^k q^{n-k} \simeq \frac{1}{\sqrt{2\pi npq}} \exp\left[-\frac{(k-np)^2}{2npq}\right]. \tag{2.42}$$

The binomial pmf becomes a Gaussian density with $\mu = np$ and $\sigma^2 = npq$, the binomial mean and variance.

The derivation takes two steps. First, replace each factorial by Stirling's formula. Second, Taylor-expand the exponent about the peak. *Step one.* Insert (2.40) for $n!$, $k!$ and $(n-k)!$:

$$\begin{aligned} \ln \mathrm{P}(k) &\overset{(1)}{=} \ln n! - \ln k! - \ln(n-k)! + k \ln p + (n-k) \ln q \\ &\overset{(2)}{=} \left[n \ln n - k \ln k - (n-k) \ln(n-k) + k \ln p + (n-k) \ln q \right] - \tfrac{1}{2} \ln \frac{2\pi k(n-k)}{n} \end{aligned}$$

$$\overset{(3)}{=} -k\ln\frac{k}{np} - (n-k)\ln\frac{n-k}{nq} - \tfrac{1}{2}\ln\frac{2\pi k(n-k)}{n}. \tag{2.43}$$

(1) The logarithm of the binomial pmf.
(2) Stirling for the three factorials. The $-n+k+(n-k)$ from the three "$-m$" terms cancels. The three $\frac{1}{2}\ln(2\pi m)$ terms combine into the last logarithm.
(3) Write $n\ln n = k\ln n + (n-k)\ln n$ and collect: $k\ln n - k\ln k + k\ln p = -k\ln\frac{k}{np}$, and the same for the $n-k$ terms.

Step two. Measure k from the centre of the binomial, $k = np+\delta$, so that δ is the distance from the peak. Then $n-k = nq-\delta$, $k/(np) = 1+\delta/(np)$ and $(n-k)/(nq) = 1-\delta/(nq)$. Each logarithm in (2.43) now has the form $\ln(1+x)$ with x small, so we use $\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \dots$:

$$\begin{aligned}(np+\delta)\ln\Big(1+\frac{\delta}{np}\Big) &\overset{(1)}{=} (np+\delta)\Big(\frac{\delta}{np} - \frac{\delta^2}{2n^2p^2} + \frac{\delta^3}{3n^3p^3}\Big) + O\Big(\frac{\delta^4}{n^3}\Big)\\ &\overset{(2)}{=} \delta + \frac{\delta^2}{2np} - \frac{\delta^3}{6n^2p^2} + O\Big(\frac{\delta^4}{n^3}\Big),\\ (nq-\delta)\ln\Big(1-\frac{\delta}{nq}\Big) &\overset{(3)}{=} -\delta + \frac{\delta^2}{2nq} + \frac{\delta^3}{6n^2q^2} + O\Big(\frac{\delta^4}{n^3}\Big).\end{aligned} \tag{2.44}$$

(1) Taylor series of the logarithm to third order.
(2) Multiply out: δ^2 terms $\frac{\delta^2}{np} - \frac{\delta^2}{2np} = \frac{\delta^2}{2np}$, and δ^3 terms $-\frac{\delta^3}{2n^2p^2} + \frac{\delta^3}{3n^2p^2} = -\frac{\delta^3}{6n^2p^2}$.
(3) The same calculation with $p \to q$ and $\delta \to -\delta$.

The exponent in (2.43) is minus the sum of these two lines:

$$\begin{aligned}-k\ln\frac{k}{np} - (n-k)\ln\frac{n-k}{nq} &\overset{(1)}{=} -\frac{\delta^2}{2n}\Big(\frac{1}{p}+\frac{1}{q}\Big) + \frac{\delta^3}{6n^2}\Big(\frac{1}{p^2}-\frac{1}{q^2}\Big) + \dots\\ &\overset{(2)}{=} -\frac{\delta^2}{2npq} + \frac{\delta^3}{6n^2}\,\frac{q^2-p^2}{p^2q^2} + \dots\end{aligned} \tag{2.45}$$

(1) The $\pm\delta$ terms cancel. This cancellation is what puts the peak at np.
(2) $\frac{1}{p}+\frac{1}{q} = \frac{p+q}{pq} = \frac{1}{pq}$ because $p+q=1$.

Which terms survive? Only the quadratic one. The probability sits where δ is of order $\sqrt{npq}$, so we judge each term there. The quadratic term $\delta^2/(2npq)$ is of order 1. The cubic term is of order $n^{3/2}/n^2 = n^{-1/2}$, which goes to zero. The prefactor also simplifies: $k(n-k)/n = npq + \delta(q-p) - \delta^2/n = npq\,[1+O(n^{-1/2})]$. Dropping the terms that vanish gives (2.42).

The cubic term also says how fast the limit is reached, in two ways. For $p = q = \frac{1}{2}$ it is exactly zero, so the symmetric binomial converges faster. And the leading corrections are *odd* in δ: they skew the curve but leave its height at the peak unchanged. So the largest pointwise error is the relative correction $n^{-1/2}$ times the height of the pmf, also of order $n^{-1/2}$, which gives n^{-1}.

Watch out: The approximation is local

(2.42) was derived for $\delta = O(\sqrt{n})$. Far in the tails, say $\delta \sim n$, the neglected cubic and higher terms are *not* small and the relative error can be enormous, even though both probabilities are tiny. A "5σ" tail probability computed from the Gaussian approximation to a binomial or Poisson count can be off by a large factor. For tail probabilities use the exact distribution (`scipy.stats.binom.sf`).

Example: The continuity correction. An integer count has to be matched to a smooth curve, and the match is best when we cut half-way between integers. Let $X \sim \text{Binomial}(25, 0.6)$, so $\mu = 15$ and $\sigma = \sqrt{25 \cdot 0.6 \cdot 0.4} = 2.449$. We want $\mathrm{P}(X \le 13)$. Think of each integer k as a bar of width one centred on k, so the bar for $k = 13$ reaches up to 13.5. The plain Gaussian approximation cuts at 13 and misses half of that bar. The **continuity correction** cuts at 13.5 instead:

$$\mathrm{P}(X \le 13) \overset{(1)}{=} \mathrm{P}(X \le 13.5) \overset{(2)}{=} \Phi\left(\frac{13.5 - 15}{2.449}\right) = 0.270.$$

(1) X is an integer, so $X \le 13$ and $X \le 13.5$ are the same event. The bar picture tells us to use the second when replacing the pmf by a density.
(2) de Moivre–Laplace with $\mu = np$, $\sigma^2 = npq$, then standardise.

The exact binomial sum is 0.268 and the uncorrected approximation $\Phi\big((13 - 15)/2.449\big) = 0.207$. The half-bar is worth six percentage points here. In general, $\mathrm{P}(X \le x) \approx \Phi\big((x + \tfrac{1}{2} - \mu)/\sigma\big)$ and $\mathrm{P}(X \ge x) \approx 1 - \Phi\big((x - \tfrac{1}{2} - \mu)/\sigma\big)$.

Example: How many photo-electrons? Back to the scintillator bar of the opening: $n = 10^4$ photons, each making a photo-electron with probability $p = 0.3$, so $\mu = np = 3000$ and $\sigma = \sqrt{npq} = \sqrt{2100} = 45.83$. We wanted $\mathrm{P}(k \le 2950)$:

$$\mathrm{P}(k \le 2950) \overset{(1)}{=} \Phi\left(\frac{2950.5 - 3000}{45.83}\right) = \Phi(-1.080) \overset{(2)}{=} 0.140.$$

(1) de Moivre–Laplace with the continuity correction.
(2) Table value. The exact binomial sum (`binom.cdf(2950, 10000, 0.3)`) agrees to three decimal places.

The same numbers give the bar's energy resolution. The relative width $\sigma/\mu = \sqrt{q/(np)} = 1.5\%$ is its photo-statistics term. It falls like $1/\sqrt{\text{number of photo-electrons}}$, the famous $a/\sqrt{E}$ term of calorimetry.

Exercise: when is n large enough. CasellaBerger's rule of thumb is $\min(np, nq) \ge 5$ [CasellaBerger p. 85]. Use (2.45) to explain why the rule involves the *smaller* of np and nq. (Answer: the relative size of the cubic term at $\delta = \sigma$ is $\frac{\sigma^3}{6n^2}\frac{|q^2 - p^2|}{p^2q^2} = \frac{|q-p|}{6\sqrt{npq}}$, which is large when either np or nq is small, i.e. when the distribution is squeezed against $k = 0$ or $k = n$.)

2.9.4 Python: watching the binomial become a Gaussian

The script computes Stirling's ratio, overlays the Gaussian on three binomials, and measures the largest pointwise error as n grows from 10 to 10^4. The first key lines check Stirling's formula:

Listing 2.10: Stirling's formula against the exact $\ln n!$ (`code/ch02/20_demoivre_laplace.py`)

```
29 # ------------------------------------------------------------------ Stirling
30 n_st = np.arange(1, 31)
31 ln_fact = special.gammaln(n_st + 1)                       # ln n!, exact
32 ln_stirling = n_st * np.log(n_st) - n_st + 0.5 * np.log(2 * np.pi * n_st)
33 ratio = np.exp(ln_fact - ln_stirling)                     # n! / Stirling
```

Line 31 uses `gammaln`, the logarithm of the Gamma function, because $n!$ itself overflows a double for $n > 170$. Any likelihood code works with logarithms of factorials for the same reason. The next lines measure how far the binomial is from its Gaussian:

Listing 2.11: The largest error of the de Moivre–Laplace approximation (code/ch02/20_demoivre_laplace.py)

```
54 fig.legend(h[::-1], l[::-1], loc="upper center", bbox_to_anchor=(0.5, 0.0), ncol=2,
       frameon=False)
55 savefig(fig, "ch02", "demoivre_panels")
56
57 # ------------------------------------------------------------------ error vs n
58 n_grid = np.unique(np.round(np.logspace(1, 4, 25)).astype(int))
59 maxerr = []
60 for n in n_grid:
61     k = np.arange(0, n + 1)
62     mu, sd = n * p, np.sqrt(n * p * (1 - p))
```

For each n on a logarithmic grid, line 60 compares the exact pmf with the Gaussian density at the same integers and keeps the worst case. Line 62 fits a straight line in log–log space. The slope comes out as -1.00, the n^{-1} predicted above from the odd (skewing) corrections. At $n = 10$ the largest error is 0.0168, at $n = 10000$ it is 1.8×10^{-5} (Figure 2.25b).

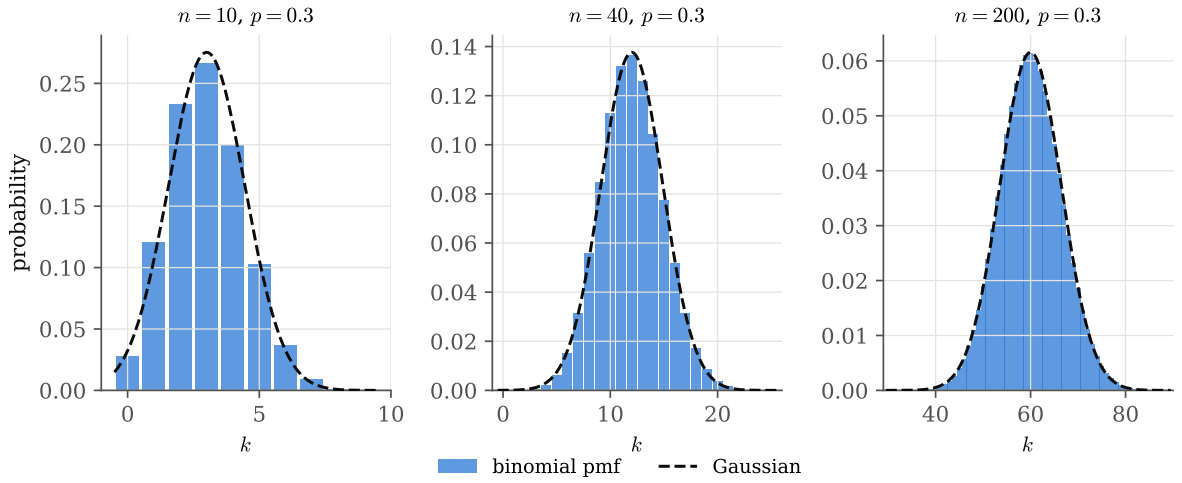


Figure 2.24: Binomial$(n, 0.3)$ probabilities (bars) with the Gaussian of the same mean and variance (dashed) for $n = 10, 40, 200$. At $n = 10$ the bars lean to the right of the curve: the distribution is squeezed against $k = 0$ and is skewed. By $n = 200$ the skew is invisible by eye.

What this bought us. The Gaussian appears whenever a distribution is sharply peaked and we look on the scale of its width: the logarithm is a parabola there. For the binomial the parabola has curvature $1/(npq)$, which fixes the Gaussian completely. The first correction is a skew of relative size $|q - p|/\sqrt{npq}$, and it is the skew, not the width, that tells us whether n is large enough.

How this is used in real research: Verifying a limit versus relying on it

Here the plot only *verifies* a known limit: the exact pmf is cheap, so nothing hangs on the approximation. Research uses the same Gaussian replacement where no exact answer is at hand. A binned χ^2 fit to a histogram treats each bin count, a binomial or Poisson number, as Gaussian with variance equal to its mean. That is fine with thousands of counts per bin and biased with a handful (the skew term above). High-energy physicists therefore switch to the exact Poisson likelihood for sparse bins (Chapters 3 and 4). CMB analyses meet the same issue at low multipoles, where each $\widehat{C}_\ell$ is a sum of only $2\ell + 1$ squared Gaussians and is visibly skewed (Chapter 7).

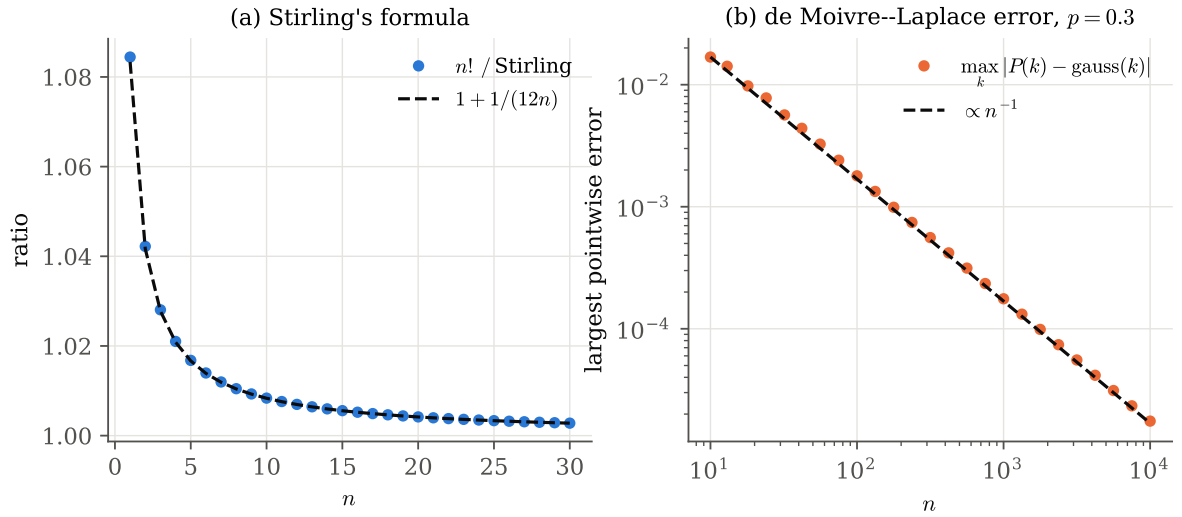


Figure 2.25: (a) The ratio of the exact $n!$ to Stirling's formula approaches one from above, following $1 + 1/(12n)$. (b) The largest pointwise difference between the binomial pmf and its Gaussian approximation falls as $1/n$ (dashed line).

Reading the numbers in the literature. AoS states the density as [Wasserman (2.3)] and works Example 2.17 (our worked example above), remarking only that "many phenomena in nature have approximately Normal distributions". Cowan's (2.22)–(2.27) is the same density with φ, Φ and the quantiles $x_\alpha = \Phi^{-1}(\alpha)$. CasellaBerger quotes the $1, 2, 3\sigma$ contents as .6826, .9544, .9974 (truncated rather than rounded), and in Example 3.3.2 the values .206, .267, .271 for the plain, exact and continuity-corrected results. Those differ slightly from ours because the book rounds z to -0.82 and -0.61 before entering the table.

2.10 The Gaussian, second way: the central limit theorem

A binomial count is the sum of n Bernoulli variables, each equal to 1 for a success and 0 for a failure. The de Moivre–Laplace result is therefore a statement about the distribution of a *sum*. The central limit theorem extends the argument beyond Bernoulli trials and explains why a Gaussian can approximate measurement errors that accumulate from many small contributions. Its assumptions tell us when that approximation is justified.

Where a timing error comes from. A drift-chamber hit time is shifted by the drift of the ionisation cluster, the rise time of the amplifier, the discriminator walk, cable delays that depend on temperature, the jitter of the clock, and a dozen smaller effects. None of these is Gaussian. A walk correction leaves a residual that is roughly uniform; a photon arrival delay is roughly exponential. Yet the histogram of timing residuals in a well-calibrated detector is a clean bell curve, until we look far out in its tails.

So the sum seems to forget the shapes of its ingredients. Three questions follow. Why does it forget? How fast does it forget, so that we know how many ingredients count as "many"? And which ingredients are never forgotten? A Breit–Wigner line shape and a Landau energy loss turn out to be two of them.

Convolution smooths, and Fourier space turns it into multiplication. Adding independent variables smooths their density, and the smoothing always ends in a Gaussian. *Why smoothing?*

The density of a sum of two independent variables is the *convolution* of their densities: every way of splitting the total between the two contributes. Convolving a box with a box gives a triangle. A triangle with a box gives three glued parabolas. After a handful of steps the shape is a smooth hump (Figure 2.26). *Why a Gaussian?* Follow the sum into Fourier space, one step at a time. There the convolution becomes a product. Taking logarithms turns the product into a sum. Each term of that sum is a power series in the frequency, and its coefficients are the *cumulants* (mean, variance, skewness, . . .). When we rescale the sum to unit width, every cumulant beyond the second is suppressed by a power of n. What survives is a quadratic in the frequency, and a function whose logarithm is quadratic is a Gaussian. This is the diffusion equation in disguise: a drop of ink of any initial shape spreads into a Gaussian cloud.

The plan.

1. What do we add? n independent copies of one random variable with finite mean μ and variance σ^2.
2. What happens to the sum without rescaling? Its mean grows like n and its width like $\sqrt{n}$, so we centre it and divide by $\sigma\sqrt{n}$.
3. Which tool makes sums easy? The characteristic function, because it turns convolutions into products.
4. What does the characteristic function of the rescaled sum tend to? Expand for small frequency, raise to the power n, take the limit.
5. How fast is the limit reached, and what breaks it? Look at the first term we threw away, and at parents whose variance is infinite.

2.10.1 Characteristic functions: the tool that turns sums into products

The density of a sum is a convolution, which is awkward to handle directly. The Fourier transform of a density, called its characteristic function, turns the convolution into a product. We need its definition and three properties.

Working definition 2.14 (characteristic function)

The **characteristic function** of a random variable X with density $p(x)$ is

$$\phi_X(k) = \mathbb{E}\big(e^{ikX}\big) = \int_{-\infty}^{\infty} e^{ikx} p(x)\,\mathrm{d}x, \tag{2.46}$$

the Fourier transform of the density [with the sign $+ikx$ and no 2π in the exponent] [Cowan (10.1)]. It exists for *every* distribution, because $|e^{ikx}| = 1$. The density is recovered by the inverse transform $p(x) = \frac{1}{2\pi}\int \phi_X(k) e^{-ikx}\mathrm{d}k$ [Cowan (10.3)], so ϕ_X and p carry the same information. (A refresher of Chapter 1: the moment generating function $\mathbb{E}\,e^{tX}$ is the same object at imaginary frequency, $t = ik$, but it can be infinite.)

Three properties do all the work. First, the characteristic function of a sum of independent variables is the product of their characteristic functions. Let $X_1, \ldots, X_n$ be independent with

densities $p_1, \dots, p_n$ and characteristic functions $\phi_1, \dots, \phi_n$, and let $S = \sum_j X_j$. Then

$$\phi_S(k) \overset{(1)}{=} \int \cdots \int e^{ik(x_1+\cdots+x_n)}\, p_1(x_1) \cdots p_n(x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n$$
$$\overset{(2)}{=} \int e^{ikx_1} p_1(x_1) \mathrm{d}x_1 \;\cdots\; \int e^{ikx_n} p_n(x_n) \mathrm{d}x_n = \phi_1(k) \cdots \phi_n(k). \tag{2.47}$$

(1) Definition (2.46) for S, with the joint density of independent variables equal to the product of the marginals.
(2) The exponential of a sum is a product of exponentials, and each factor depends on one variable only, so the n-fold integral factorises.

Second, shifting multiplies by a phase and rescaling rescales the frequency: for constants a and b, $\phi_{aX+b}(k) = \mathbb{E}\, e^{ik(aX+b)} = e^{ikb}\phi_X(ak)$. Third, derivatives at the origin give moments, $\frac{\mathrm{d}^m \phi_X}{\mathrm{d}k^m}\big|_{k=0} = i^m\, \mathbb{E}(X^m)$ [Cowan (10.4)], so near $k = 0$

$$\phi_X(k) = 1 + ik\,\mathbb{E}\,X - \frac{k^2}{2}\,\mathbb{E}\,X^2 - \frac{ik^3}{6}\,\mathbb{E}\,X^3 + \dots \tag{2.48}$$

The characteristic function of a Gaussian. For $Z \sim \mathcal{N}(0,1)$ differentiate under the integral and integrate by parts:

$$\phi_Z'(k) \overset{(1)}{=} \int iz\, e^{ikz} \varphi(z)\, \mathrm{d}z \overset{(2)}{=} -i \int e^{ikz} \varphi'(z)\, \mathrm{d}z \overset{(3)}{=} i \int \big(ik\, e^{ikz}\big) \varphi(z)\, \mathrm{d}z = -k\, \phi_Z(k). \tag{2.49}$$

(1) Differentiate e^{ikz} with respect to k.
(2) $z\varphi(z) = -\varphi'(z)$.
(3) Integrate by parts. The boundary term vanishes because $\varphi \to 0$, and $\frac{\mathrm{d}}{\mathrm{d}z} e^{ikz} = ik\, e^{ikz}$ moves to the other factor with a minus sign.

The solution of $\phi' = -k\phi$ with $\phi(0) = 1$ is $\phi_Z(k) = e^{-k^2/2}$. For $X = \mu + \sigma Z$ the scaling rule gives

$$\phi_X(k) = \exp\big(i\mu k - \tfrac{1}{2}\sigma^2 k^2\big), \tag{2.50}$$

the entry in Cowan's Table 10.1. The Fourier transform of a Gaussian is a Gaussian, with the widths inverted: a narrow density has a wide characteristic function. A sum of independent Gaussians is therefore Gaussian, with the means and the variances adding [Cowan (10.11)]. The reason is the product rule: the product of two functions of the form (2.50) is again of that form, with $\mu = \mu_1 + \mu_2$ and $\sigma^2 = \sigma_1^2 + \sigma_2^2$. This is AoS's fact (iii) [Wasserman p. 28].

2.10.2 The theorem and its proof

The theorem asks three things of the terms in the sum: they are independent, they share one distribution, and that distribution has a finite variance. It asks nothing about the shape.

Key idea: Central limit theorem

Let $X_1, X_2, \dots$ be independent and identically distributed with mean μ and *finite* variance $\sigma^2 > 0$. Let $\bar{X}_n = \frac{1}{n}\sum_{j=1}^n X_j$ and

$$Z_n = \frac{\sum_{j=1}^n X_j - n\mu}{\sigma\sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}. \tag{2.51}$$

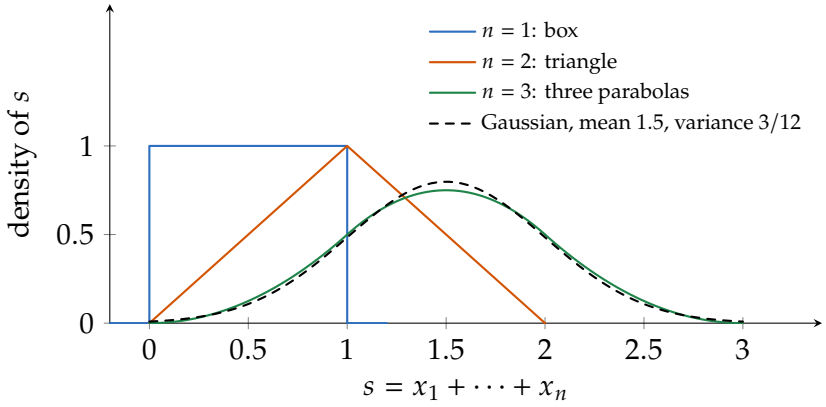


Figure 2.26: Repeated convolution of a flat density on $[0,1]$. One uniform is a box, the sum of two is a triangle, the sum of three is made of three parabolas joined smoothly. Already at $n = 3$ the green curve is hard to tell from the Gaussian with the same mean and variance (dashed). Each convolution also makes the curve one derivative smoother.

Then Z_n **converges in distribution** [the probabilities $\mathrm{P}(Z_n \le z)$ converge, for every z] to $\mathcal{N}(0,1)$:

$$\lim_{n\to\infty} \mathrm{P}(Z_n \le z) = \Phi(z) \qquad \text{for all } z. \tag{2.52}$$

Equivalently, $\bar{X}_n \approx \mathcal{N}(\mu, \sigma^2/n)$ and $\sum_j X_j \approx \mathcal{N}(n\mu, n\sigma^2)$ for large n.

The theorem approximates probabilities, not the random variable: "It's the probability statements that we are approximating, not the random variable itself" [Wasserman p. 77]. Z_n does not *become* a Gaussian variable. What happens is that the numbers $\mathrm{P}(Z_n \le z)$ approach the numbers $\Phi(z)$.

Proof. The plan is to show that the characteristic function of Z_n tends to $e^{-k^2/2}$, the one of $\mathcal{N}(0,1)$. Let $Y_j = (X_j - \mu)/\sigma$, so $\mathbb{E}\,Y_j = 0$, $\mathbb{E}\,Y_j^2 = 1$, and $Z_n = \sum_j Y_j/\sqrt{n}$. Because the second moment is finite, (2.48) holds to second order with a remainder that is small compared to k^2: $\phi_Y(k) = 1 - \frac{1}{2}k^2 + o(k^2)$. Then

$$\begin{aligned}
\phi_{Z_n}(k) &\overset{(1)}{=} \left[\phi_Y\!\left(\frac{k}{\sqrt{n}}\right)\right]^n \\
&\overset{(2)}{=} \left[1 - \frac{k^2}{2n} + o\!\left(\frac{1}{n}\right)\right]^n \\
&\overset{(3)}{=} \exp\left\{n \ln\left[1 - \frac{k^2}{2n} + o\!\left(\frac{1}{n}\right)\right]\right\} \\
&\overset{(4)}{=} \exp\left\{n\left[-\frac{k^2}{2n} + o\!\left(\frac{1}{n}\right)\right]\right\} \xrightarrow{\;n\to\infty\;} e^{-k^2/2}.
\end{aligned} \tag{2.53}$$

(1) Product rule (2.47) for n identical factors, and the scaling rule with $a = 1/\sqrt{n}$.
(2) The second-order expansion of ϕ_Y at the small frequency $k/\sqrt{n}$ (k is held fixed).
(3) $a^n = e^{n \ln a}$.
(4) $\ln(1+x) = x + O(x^2)$, and $x^2 = O(1/n^2)$ so $nx^2 \to 0$. This is the step AoS writes as $(1 + a_n/n)^n \to e^a$ when $a_n \to a$.

The limit is the characteristic function of $\mathcal{N}(0,1)$. Finally, **Lévy's continuity theorem** [if characteristic functions converge at every k to a function continuous at $k = 0$, the distributions converge] turns this into (2.52). ∎

Look at what the proof used. Independence, to multiply characteristic functions. A finite variance, to expand to second order. Identical distributions, only for convenience. It did *not* use anything about the shape of $p(x)$. That is the whole content of "regardless of the form of the individual p.d.f.s" [Cowan p. 33].

Why characteristic functions and not moment generating functions. AoS proves the CLT with the moment generating function $\psi(t) = \mathbb{E}\, e^{tX}$ and assumes it is finite near $t = 0$ [Wasserman p. 81–82]. That assumption excludes perfectly good cases. A Student t_3 variable (Section 2.14) has variance 3 but $\mathbb{E}\, e^{tX} = \infty$ for every $t \neq 0$, because its tails fall like a power. The CLT holds for it; the MGF proof cannot show it. The characteristic function always exists, which is why Cowan and all rigorous treatments use it.

Not identically distributed. Terms with different variances still give a Gaussian sum, provided no single term dominates. If the X_j are independent with means μ_j and variances σ_j^2, the sum tends to $\mathcal{N}\big(\sum \mu_j, \sum \sigma_j^2\big)$ [Cowan p. 147–148] under "somewhat more restrictive conditions". The condition (Lindeberg's) says in words: *no single term may carry a finite fraction of the total variance*. A sum of one huge error and many tiny ones is as non-Gaussian as the huge error.

Vectors. The theorem also holds for vectors. If $\boldsymbol{X}_j$ are iid vectors with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}$, then $\sqrt{n}(\bar{\boldsymbol{X}}_n - \boldsymbol{\mu})$ tends to the multivariate Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{C})$ of Section 2.12 [Wasserman p. 78–79, Thm. 5.12]. This is why a histogram with many entries, or a set of band-powers built from many modes, has a Gaussian *joint* distribution.

Unknown σ: the studentised version. Replacing the unknown σ by its estimate from the data leaves the limit unchanged. In practice σ is not known, and we use the sample standard deviation S_n, with $S_n^2 = \frac{1}{n-1}\sum_j (X_j - \bar{X}_n)^2$. Then still $\sqrt{n}(\bar{X}_n - \mu)/S_n \to \mathcal{N}(0,1)$ [Wasserman p. 79, Thm. 5.10]. *Why?* This ratio equals $Z_n \cdot (\sigma/S_n)$, and $S_n^2 = \frac{n}{n-1}\big[\frac{1}{n}\sum X_j^2 - \bar{X}_n^2\big] \to \mathbb{E}\, X^2 - \mu^2 = \sigma^2$ by the law of large numbers, so the second factor tends to the constant 1. A quantity that tends to $\mathcal{N}(0,1)$ times one that tends to 1 still tends to $\mathcal{N}(0,1)$ (Slutsky's argument). For small n and Gaussian data the exact law of this ratio is Student's t (Section 2.14).

2.10.3 How fast? Cumulants, skewness and the Berry–Esseen bound

How large n must be is decided by the k^3 term that the proof threw away. To keep track of it, we take logarithms of characteristic functions.

Working definition 2.15 (cumulants)

The **cumulants** κ_r of X are the Taylor coefficients of the logarithm of its characteristic function,

$$\ln \phi_X(k) = \sum_{r \geq 1} \kappa_r \frac{(ik)^r}{r!}. \tag{2.54}$$

κ_1 is the mean, κ_2 the variance, $\kappa_3 = \mathbb{E}(X-\mu)^3$ the third central moment [it measures asymmetry]. The **skewness** is $\gamma_1 = \kappa_3/\sigma^3$ and the **excess kurtosis** $\gamma_2 = \kappa_4/\sigma^4$ [zero for a Gaussian, positive for heavier tails]. For a Gaussian, $\ln\phi = i\mu k - \sigma^2 k^2/2$, so *all* cumulants beyond the second vanish.

Two rules fix how cumulants behave in a sum. Cumulants of independent variables add, because logarithms turn the product (2.47) into a sum. And rescaling by a multiplies κ_r by a^r,

by the scaling rule $\phi_{aX}(k) = \phi_X(ak)$. So for the standardised sum (2.51),

$$\kappa_r(Z_n) \overset{(1)}{=} n\,\kappa_r\Big(\frac{X-\mu}{\sigma\sqrt{n}}\Big) \overset{(2)}{=} n\,\frac{\kappa_r(X)}{\sigma^r n^{r/2}} = \frac{\kappa_r(X)}{\sigma^r}\,n^{1-r/2} \qquad (r \geq 2). \tag{2.55}$$

(1) Z_n is a sum of n independent copies of $(X-\mu)/(\sigma\sqrt{n})$, and cumulants add.
(2) Shifting changes only κ_1. Scaling by $1/(\sigma\sqrt{n})$ multiplies κ_r by $(\sigma\sqrt{n})^{-r}$.

This is the CLT again, now with rates. For $r = 2$ the result is 1 for every n (unit variance, by construction). For $r = 3$ it says the skewness of Z_n is $\gamma_1/\sqrt{n}$. For $r = 4$ the excess kurtosis is γ_2/n. Higher cumulants die faster still.

Example: Predicting how skewed a sum of exponentials is. For $X \sim \mathrm{Exp}(1)$ the characteristic function is $\phi(k) = (1-ik)^{-1}$ [Cowan Table 10.1], so

$$\ln\phi(k) \overset{(1)}{=} -\ln(1-ik) \overset{(2)}{=} \sum_{r\geq 1}\frac{(ik)^r}{r} \overset{(3)}{=} \sum_{r\geq 1}(r-1)!\,\frac{(ik)^r}{r!},$$

(1) Take the logarithm.
(2) $-\ln(1-u) = \sum_{r\geq 1} u^r/r$ with $u = ik$.
(3) Multiply and divide by $r!$ to read off the cumulants in the form (2.54).

so $\kappa_r = (r-1)!$: mean 1, variance 1, $\kappa_3 = 2$, skewness $\gamma_1 = 2$. By (2.55) a sum of n exponentials has skewness $2/\sqrt{n}$. The simulation (below) measures 2.01 at $n = 1$, 0.91 at $n = 5$ (predicted 0.894) and 0.366 at $n = 30$ (predicted 0.365). A sum of thirty exponentials is still visibly lopsided.

The first correction to the Gaussian (Edgeworth). The skewness term gives the leading error of the Gaussian approximation. Keep the $r = 3$ cumulant in the exponent and expand: $\phi_{Z_n}(k) \approx e^{-k^2/2}\big[1 + \frac{\gamma_1}{6\sqrt{n}}(ik)^3\big]$. Under the inverse transform each factor ik becomes $-\mathrm{d}/\mathrm{d}z$ (differentiate e^{-ikz}), so

$$\begin{aligned}\mathrm{P}(Z_n \leq z) &\overset{(1)}{=} \Phi(z) + \frac{\gamma_1}{6\sqrt{n}}\int_{-\infty}^{z}\Big(-\frac{\mathrm{d}}{\mathrm{d}z'}\Big)^3\varphi(z')\,\mathrm{d}z' + O\Big(\frac{1}{n}\Big)\\ &\overset{(2)}{=} \Phi(z) - \frac{\gamma_1}{6\sqrt{n}}\,\varphi''(z) + O\Big(\frac{1}{n}\Big) \overset{(3)}{=} \Phi(z) - \frac{\gamma_1}{6\sqrt{n}}\,(z^2-1)\varphi(z) + O\Big(\frac{1}{n}\Big).\end{aligned} \tag{2.56}$$

(1) Inverse Fourier transform of the two terms of the expanded characteristic function, then integrate the density up to z.
(2) $(-\mathrm{d}/\mathrm{d}z)^3 = -\mathrm{d}^3/\mathrm{d}z^3$, and integrating φ''' from $-\infty$ gives φ'' (it vanishes at $-\infty$).
(3) $\varphi' = -z\varphi$, so $\varphi'' = (z^2-1)\varphi$.

The correction is largest at $z = 0$, where it has size $\gamma_1\varphi(0)/(6\sqrt{n}) = 0.0665\,\gamma_1/\sqrt{n}$. We check this against the exact worst-case gap, the **Kolmogorov distance** $D_n = \sup_z |\,\mathrm{P}(Z_n \leq z) - \Phi(z)|$, which the script computes without Monte Carlo noise. For exponentials ($\gamma_1 = 2$) and $n = 32$ the prediction is 0.0235 and the exact value is 0.0235. A *symmetric* parent behaves differently. Its skewness γ_1 is zero, so the $n^{-1/2}$ term is absent, and the error starts at order $1/n$ with a coefficient set by the kurtosis. For uniforms ($\gamma_2 = -6/5$) the next Edgeworth term predicts 8.6×10^{-4} at $n = 32$, and the exact value is 8.7×10^{-4}. The fitted slopes of $\ln D_n$ against $\ln n$ are -0.50 (exponential) and -1.02 (uniform), Figure 2.28.

A guaranteed bound. The Edgeworth term is an estimate; a proof needs a bound. The Berry–Esseen inequality gives one, with no expansion [Wasserman p. 78, (5.4)]:

$$\sup_z \left|\mathrm{P}(Z_n \le z) - \Phi(z)\right| \le C\,\frac{\mathbb{E}\,|X-\mu|^3}{\sigma^3\sqrt{n}}, \tag{2.57}$$

with $C = 33/4$ in AoS. For Exp(1), $\mathbb{E}\,|X-1|^3 = \int_0^1 (1-x)^3 e^{-x}\mathrm{d}x + \int_1^\infty (x-1)^3 e^{-x}\mathrm{d}x = (6/e-2) + 6/e = 12/e - 2 = 2.415$. At $n = 30$ the AoS bound is 3.64, larger than 1 and therefore useless, while the true distance is 0.0243. The best constant known today is below 0.48, which gives 0.209: a valid guarantee, but still an order of magnitude pessimistic. The bound is for proofs, the Edgeworth term for estimates.

Example: Programs with Poisson bugs. The number of errors in a computer program is Poisson with mean 5. We have 125 programs, and $\bar{X}_{125}$ is the average number of errors per program. What is the probability that this average is below 5.5, $\mathrm{P}(\bar{X}_{125} < 5.5)$? For Poisson($\lambda$) the mean and variance are both λ, so here $\mu = 5$ and $\sigma = \sqrt{5}$, and

$$\mathrm{P}(\bar{X}_n < 5.5) \overset{(1)}{=} \mathrm{P}\Big(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} < \frac{\sqrt{125}\,(5.5-5)}{\sqrt{5}}\Big) \overset{(2)}{=} \mathrm{P}(Z_n < 2.5) \overset{(3)}{=} \Phi(2.5) = 0.9938.$$

(1) Subtract the mean, multiply by $\sqrt{n}/\sigma$ on both sides.
(2) $\sqrt{125}\times 0.5/\sqrt{5} = \sqrt{25}\times 0.5 = 2.5$.
(3) CLT (2.52).

This is [Wasserman p. 77–78, Example 5.9]. Here we can check the CLT against the exact answer. The sum $\sum X_j$ is itself Poisson(625), so the exact answer is $\mathrm{P}(\sum X_j \le 687) = 0.9932$. The CLT is good to six parts in ten thousand, because the Poisson skewness $1/\sqrt{625} = 0.04$ is small.

Example: Poisson becomes Gaussian, in Fourier space. The same limit can be watched directly in Fourier space. Let $n \sim \text{Poisson}(\nu)$, with characteristic function $\phi_n(k) = \exp[\nu(e^{ik}-1)]$ [Cowan Table 10.1]. Standardise, $x = (n-\nu)/\sqrt{\nu}$:

$$\phi_x(k) \overset{(1)}{=} e^{-ik\sqrt{\nu}}\,\phi_n\Big(\frac{k}{\sqrt{\nu}}\Big) \overset{(2)}{=} \exp\Big[\nu\Big(e^{ik/\sqrt{\nu}} - 1\Big) - ik\sqrt{\nu}\Big]$$
$$\overset{(3)}{=} \exp\Big[\nu\Big(\frac{ik}{\sqrt{\nu}} - \frac{k^2}{2\nu} - \frac{ik^3}{6\nu^{3/2}} + \dots\Big) - ik\sqrt{\nu}\Big] \overset{(4)}{=} \exp\Big[-\frac{k^2}{2} - \frac{ik^3}{6\sqrt{\nu}} + \dots\Big] \xrightarrow{\nu\to\infty} e^{-k^2/2}.$$

(1) Scaling rule with $a = 1/\sqrt{\nu}$, $b = -\sqrt{\nu}$.
(2) Insert the Poisson characteristic function.
(3) Expand $e^u = 1 + u + u^2/2 + u^3/6 + \dots$ with $u = ik/\sqrt{\nu}$.
(4) The $ik\sqrt{\nu}$ terms cancel. The cubic term is $\kappa_3(x)(ik)^3/6$ with $\kappa_3(x) = 1/\sqrt{\nu}$, the skewness of a Poisson count.

This is Cowan's (10.8)–(10.10). The Poisson mean plays the role of n: a Poisson count is a sum of ν unit-mean Poisson counts.

Example: Inverting a characteristic function by residues. Sometimes the inverse transform can be done exactly, and then we get the distribution of the sum for every n, not only its limit. Take n independent lifetimes, each Exp(ξ) with mean ξ. Each has $\phi_x(k) = (1-ik\xi)^{-1}$, so the sum $z = \sum x_j$ has $\phi_z(k) = (1-ik\xi)^{-n}$ [Cowan (10.21)–(10.22)]. The inverse transform $g(z) = \frac{1}{2\pi}\int e^{-ikz}(1-ik\xi)^{-n}\mathrm{d}k$ has a pole of order n at $k = -i/\xi$ in the lower half plane. For

$z > 0$, e^{-ikz} decays when $\operatorname{Im} k < 0$, so we close the contour below (clockwise, a factor $-2\pi i$). With $(1 - ik\xi)^{-n} = (-i\xi)^{-n}(k + i/\xi)^{-n}$,

$$g(z) \overset{(1)}{=} \frac{-2\pi i}{2\pi}\,(-i\xi)^{-n}\,\frac{1}{(n-1)!}\,\frac{\mathrm{d}^{n-1}}{\mathrm{d}k^{n-1}} e^{-ikz}\Big|_{k=-i/\xi} \overset{(2)}{=} \frac{-i\,(-iz)^{n-1}}{(-i\xi)^n (n-1)!}\,e^{-z/\xi} \overset{(3)}{=} \frac{1}{(n-1)!}\,\frac{z^{n-1}}{\xi^n} e^{-z/\xi}.$$

(1) Residue of a pole of order n: $\frac{1}{(n-1)!}$ times the $(n-1)$-th derivative of the rest of the integrand.

(2) Each derivative of e^{-ikz} brings down $-iz$, and $e^{-i(-i/\xi)z} = e^{-z/\xi}$.

(3) $-i\,(-i)^{n-1}/(-i)^n = 1$.

This is the Gamma (Erlang) density of the waiting time for the n-th event, met in the counting part of this chapter [Cowan (10.24)]. So the estimate $\hat{\xi} = z/n$ of the mean lifetime has an exactly known, skewed distribution. Its skewness is the $2/\sqrt{n}$ found above for a sum of n exponentials.

2.10.4 Python: sums of uniforms and exponentials

Listing 2.12: Standardised sums of n uniforms and n exponentials (`code/ch02/21_clt_sums.py`)

```
for row, (name, (draw, mu, sd)) in enumerate(parents.items()):
    for col, n in enumerate(ns):
        s = draw((NSAMP, n)).sum(axis=1)                    # the sum of n copies
        zn = (s - n * mu) / (sd * np.sqrt(n))               # standardised sum
        if name == "exponential":
            skew_meas[n] = stats.skew(zn)
```

Line 44 draws an array of shape $(200000, n)$ and sums along each row, so each row is one "experiment" with n ingredients. Line 45 standardises with the *exact* parent mean and standard deviation, which is (2.51). The histograms are in Figure 2.27.

Listing 2.13: Exact Kolmogorov distances, without Monte Carlo noise (`code/ch02/21_clt_sums.py`)

```
D_exp = []
for n in n_rate:
    # S_n ~ Gamma(n,1): P(Z_n <= z) = GammaCDF(n + z sqrt(n))
    D_exp.append(np.max(np.abs(stats.gamma.cdf(n + zz * np.sqrt(n), a=n) -
    stats.norm.cdf(zz))))
D_exp = np.array(D_exp)

dx = 2e-4                                   # grid step for the uniform convolutions
base = np.ones(int(round(1 / dx))) * 1.0    # density of U(0,1) sampled on the grid
dens = base.copy()
D_uni = []
for n in range(1, n_rate.max() + 1):
    if n > 1:
        dens = signal.fftconvolve(dens, base) * dx
    if n in n_rate:
        # U(0,1) is sampled at (i+1/2)dx, so the n-fold sum lives at (m + n/2)dx
        x = (np.arange(dens.size) + 0.5 * n) * dx
        cdf = np.cumsum(dens) * dx - 0.5 * dens * dx   # trapezoidal CDF at the grid points
        zn = (x - n * 0.5) / (np.sqrt(n / 12))
        D_uni.append(np.max(np.abs(cdf - stats.norm.cdf(zn))))
D_uni = np.array(D_uni)
```

A Monte Carlo histogram cannot resolve a distance of 10^{-3}, so the rate is computed exactly. For exponentials the sum is Gamma$(n, 1)$ (line 67). For uniforms, lines 74–82 build the density of the sum by repeated numerical convolution on a fine grid (FFT convolution, line 76), exactly the operation pictured in Figure 2.26, and integrate it to a CDF.

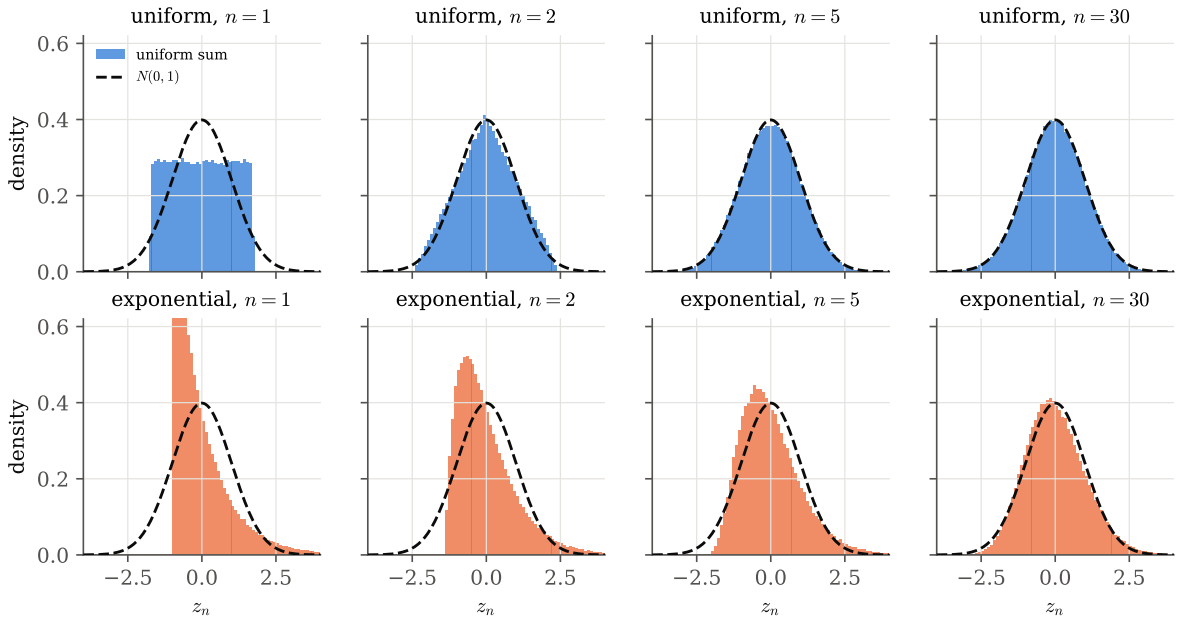


Figure 2.27: Histograms of the standardised sum z_n of n uniforms (top) and n exponentials (bottom), 200000 experiments each, with the standard normal density (dashed). The uniform sum is already bell-shaped at $n = 2$ to 5. The exponential sum keeps a visible skew, a steep left side and a long right tail, even at $n = 30$.

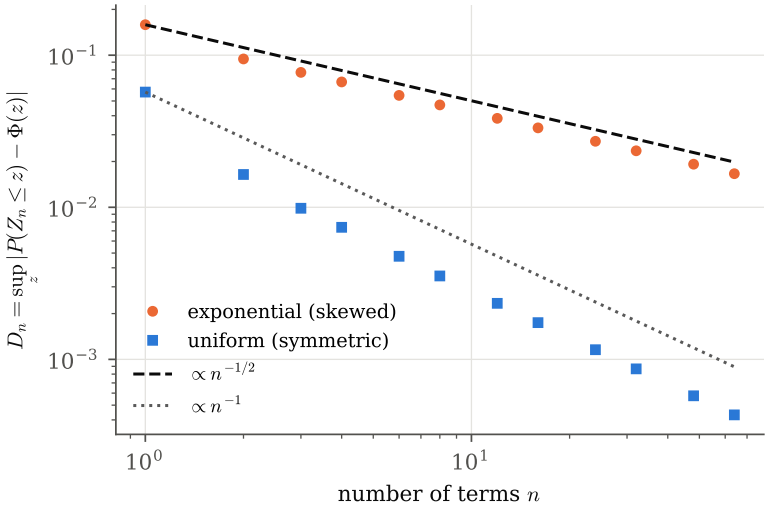


Figure 2.28: The worst-case gap between the distribution of z_n and Φ, as a function of the number of terms. For the skewed exponential parent it falls as $n^{-1/2}$ (dashed), for the symmetric uniform parent as n^{-1} (dotted). At $n = 30$ the symmetric case is more than twenty times closer to Gaussian.

What this bought us. "Large n" is not a number; it depends on the skewness of what is being added. The error of the Gaussian approximation is about $0.07\,\gamma_1/\sqrt{n}$ in cumulative probability. Symmetric ingredients converge like $1/n$, lopsided ones like $1/\sqrt{n}$.

2.10.5 Where the theorem fails: Breit–Wigner resonances

Measuring the Z mass with a running average. The invariant mass of e^+e^- pairs from Z decays follows, to a first approximation, the (non-relativistic) **Breit–Wigner** or **Cauchy** line

shape

$$p(m \mid M, \Gamma) = \frac{1}{\pi} \frac{\Gamma/2}{(m - M)^2 + \Gamma^2/4}, \tag{2.58}$$

with $M = 91.1876\,\text{GeV}$ the mass and $\Gamma = 2.4952\,\text{GeV}$ the full width at half maximum [Cowan (2.41)]. (We ignore detector resolution and radiation to keep the point clean.) A student averages the masses of the first N events and expects the average to settle down like $1/\sqrt{N}$.

The student's expectation fails, because the Breit–Wigner has no mean. It is a proper density: with $u = 2(m - M)/\Gamma$, $\int \frac{\mathrm{d}u}{\pi(1+u^2)} = \frac{1}{\pi}[\arctan u]_{-\infty}^{\infty} = 1$ [Wasserman p. 30]. But the two halves of the mean, $\int_M^{\infty} m\, p(m)\mathrm{d}m$ and $\int_{-\infty}^{M} m\, p(m)\mathrm{d}m$, diverge separately, because the integrand falls only like $1/m$. Their difference is therefore undefined [Cowan p. 36–37]; CasellaBerger's (3.3.20) states it as $\mathbb{E}\,|X| = \infty$. The variance is infinite too, so the CLT does not apply.

Instead, the mean of Cauchy variables is as wide as one of them. If $X_1, \ldots, X_n$ are iid Cauchy with centre M and half-width $\gamma = \Gamma/2$, then $\bar{X}_n$ has *exactly the same* Cauchy distribution, for every n [CasellaBerger p. 177]. Averaging does not help at all.

The proof is two lines once we have the characteristic function. For the standard Cauchy $p(x) = 1/[\pi(1 + x^2)]$ and $k > 0$, close the contour in the upper half plane, where e^{ikx} decays, around the pole at $x = i$:

$$\phi(k) \overset{(1)}{=} \frac{1}{\pi} \oint \frac{e^{ikx}}{(x - i)(x + i)} \mathrm{d}x \overset{(2)}{=} \frac{1}{\pi}\, 2\pi i\, \frac{e^{ik\cdot i}}{2i} = e^{-k}, \tag{2.59}$$

(1) Factor $1 + x^2$. The semicircle at infinity contributes nothing for $k > 0$.
(2) Residue theorem: $2\pi i$ times the residue at $x = i$, which is $e^{ik\,i}/(i + i)$.

and by symmetry $\phi(k) = e^{-|k|}$ for all k [Cowan Table 10.1]. For centre M and scale γ, $\phi(k) = e^{iMk - \gamma|k|}$. Then

$$\phi_{\bar{X}_n}(k) \overset{(1)}{=} \left[\phi\left(\frac{k}{n}\right)\right]^n \overset{(2)}{=} \left[e^{iMk/n - \gamma|k|/n}\right]^n = e^{iMk - \gamma|k|}, \tag{2.60}$$

(1) Product and scaling rules with $a = 1/n$.
(2) The Cauchy characteristic function at frequency k/n.

which is the characteristic function of a single event. *Where did the CLT proof break?* At the expansion. The kink of $e^{-|k|}$ at $k = 0$ is the Fourier-space signature of the missing variance: ϕ has no second derivative there, so the expansion (2.48) does not exist [Cowan p. 145]. The width is controlled by $|k|$ rather than k^2. So in the exponent it scales as $n \times \frac{1}{n} = 1$, instead of the $n \times \frac{1}{n^2}$ that would have made it shrink.

One event can move the average anywhere. With a Gaussian, an event 10σ away is essentially impossible. With a Breit–Wigner, the probability of landing further than t from the peak falls only like $1/t$. Among N events, the most extreme one is typically at distance $\sim N\gamma$ from the peak, so its contribution to the average, $\sim N\gamma/N = \gamma$, never becomes small. The average is always at the mercy of the latest outlier. The *median* does not care how far out an outlier is, only on which side it lies.

The simulation shows both behaviours. The script draws 10^5 Breit–Wigner masses in three independent streams. After 10^5 events one running mean sits at 94.52 GeV, 3.3 GeV from the true mass, and all three keep jumping (Figure 2.29a). The running median of the same events is 91.186 GeV. Its standard error, $\pi\gamma/(2\sqrt{N}) = 0.0062\,\text{GeV}$, does fall like $1/\sqrt{N}$: the median of any continuous density has asymptotic variance $1/[4Np(M)^2]$, and $p(M) = 1/(\pi\gamma)$ here. Panel (b) repeats the average of 1000 events twenty thousand times: its interquartile range is 2.463 GeV, equal to that of a *single* event, $2\gamma = \Gamma = 2.495\,\text{GeV}$.

Listing 2.14: Running mean and running median of Breit–Wigner masses (code/ch02/22_cauchy_landau.py)

```
for s in range(3):
    m_bw = M + gam * rng.standard_cauchy(N)                # Breit-Wigner masses
    ax.semilogx(nn, np.cumsum(m_bw) / nn, color=SERIES[s], lw=1.0,
                label="Breit--Wigner, running mean" if s == 0 else None)
m_g = M + sig_g * rng.standard_normal(N)
ax.semilogx(nn, np.cumsum(m_g) / nn, color="k", lw=1.2, label="Gaussian, running mean")
# running median of the last Breit-Wigner stream, on a sparse grid
grid = np.unique(np.round(np.logspace(0, 5, 60)).astype(int))
ax.semilogx(grid, [np.median(m_bw[:k]) for k in grid], color=SERIES[3], ls="--",
            lw=1.4, label="Breit--Wigner, running median")
```

Line 42 generates Breit–Wigner masses as $M + \gamma\, C$, where $\gamma = \Gamma/2$ is the half-width and C is a standard Cauchy variable. Numpy draws C with `rng.standard_cauchy`. (A standard Cauchy variable is the same thing as the ratio of two independent standard normals.) Line 43 computes all running means at once with a cumulative sum. Line 49 evaluates the running median on a logarithmic grid of N values.

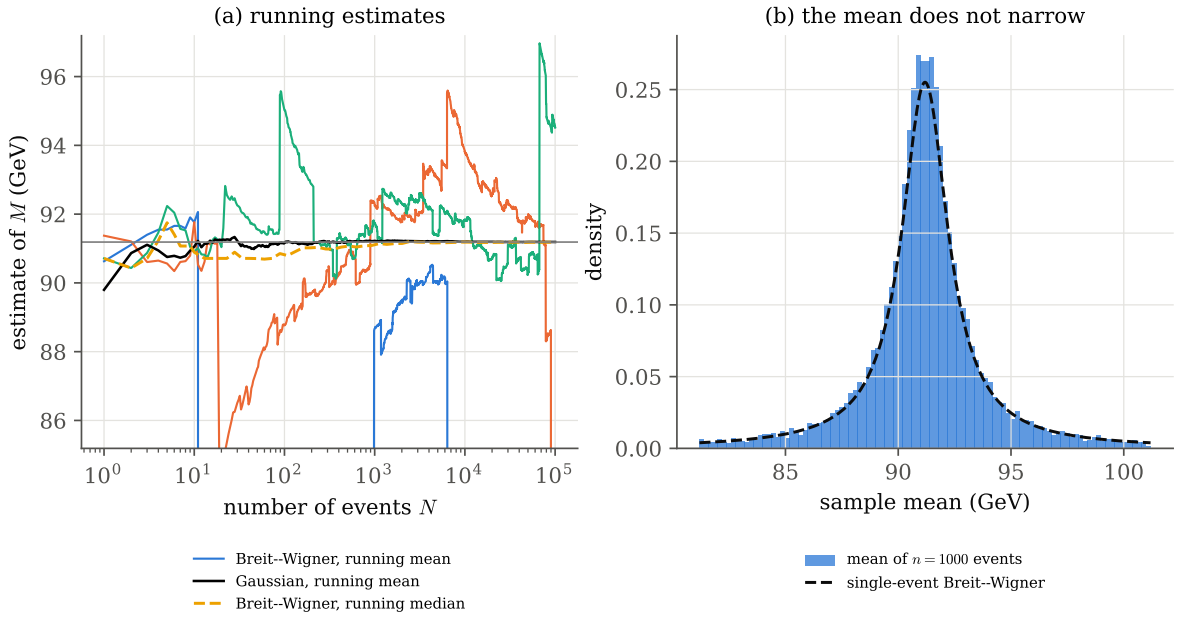


Figure 2.29: (a) Running estimates of the Z mass from Breit–Wigner events. The three coloured running means jump by GeV every time an extreme event arrives and never settle. The running mean of Gaussian events of the same half-width (black) and the running median of the Breit–Wigner events (dashed) both converge to the true mass (grey line). (b) The distribution of the average of 1000 events (histogram) is the single-event Breit–Wigner (dashed): no narrowing at all.

Watch out: The width convention

Cowan's Γ in (2.58) is the *full* width at half maximum, the convention of particle physics, where $\Gamma = \hbar/\tau$ is the decay rate. Many statistics texts write the Cauchy scale γ as the *half* width, and `scipy.stats.cauchy(loc, scale)` expects $\gamma = \Gamma/2$ [Cowan p. 37, footnote].

Strictly, real resonances do have a mean. The mass cannot be negative or exceed the collision energy, and the relativistic line shape with phase space and parton luminosities cuts the tails [Cowan p. 38–39]. But the cut sits so far out that the practical lesson stands: sample means of resonance masses converge painfully slowly and are dominated by a few events. Fit the line shape, or use the median or a truncated mean.

2.10.6 Where the theorem is slow: Landau energy loss

A charged particle crossing a thin layer loses energy Δ in many soft collisions and, occasionally, one hard collision that knocks out a fast δ-electron. The distribution of Δ, derived by Landau, is [Cowan (2.42)–(2.46)]

$$f(\Delta) = \frac{1}{\xi}\,\phi_L(\lambda), \qquad \lambda = \frac{1}{\xi}\left[\Delta - \xi\left(\ln\frac{\xi}{\epsilon'} + 1 - \gamma_E\right)\right], \qquad \phi_L(\lambda) = \frac{1}{\pi}\int_0^\infty e^{-u\ln u - \lambda u}\,\sin \pi u\,\mathrm{d}u, \tag{2.61}$$

where ξ grows with the layer thickness and falls like $1/\beta^2$ (β is the particle's speed in units of c), ϵ' depends on the ionisation energy of the material and on $\beta\gamma$, and $\gamma_E = 0.5772$ is Euler's constant. The high tail falls like $1/\lambda^2$, following the Rutherford cross-section for hard knock-ons. So, as for the Breit–Wigner, the mean and variance do not exist. The useful number is the most probable loss, $\Delta_{\mathrm{mp}} = \xi[\ln(\xi/\epsilon') + 0.198]$ [Cowan (2.48)]. It depends on the velocity, and it is the basis of particle identification by $\mathrm{d}E/\mathrm{d}x$.

Adding layers does not make the Landau tail Gaussian. Cowan's example [Cowan p. 148]: the ionisation in 1 m of argon is the sum over 25 layers of 4 cm, and one might hope the sum is Gaussian. The script adds 25 standard Landau variables and compares the result with a Gaussian of the same median and interquartile range (Figure 2.30). The core matches. The tail does not: a fraction 6.1×10^{-2} of the sums lie more than five matched standard deviations above the median, against 2.9×10^{-7} for a Gaussian. Rare hard collisions dominate the sum however many layers we add. (Mathematically the Landau distribution is *stable*: like the Cauchy, the suitably shifted and rescaled sum of Landau variables is again Landau.) Practical $\mathrm{d}E/\mathrm{d}x$ estimators therefore use a truncated mean, discarding the largest 30–40% of the samples.

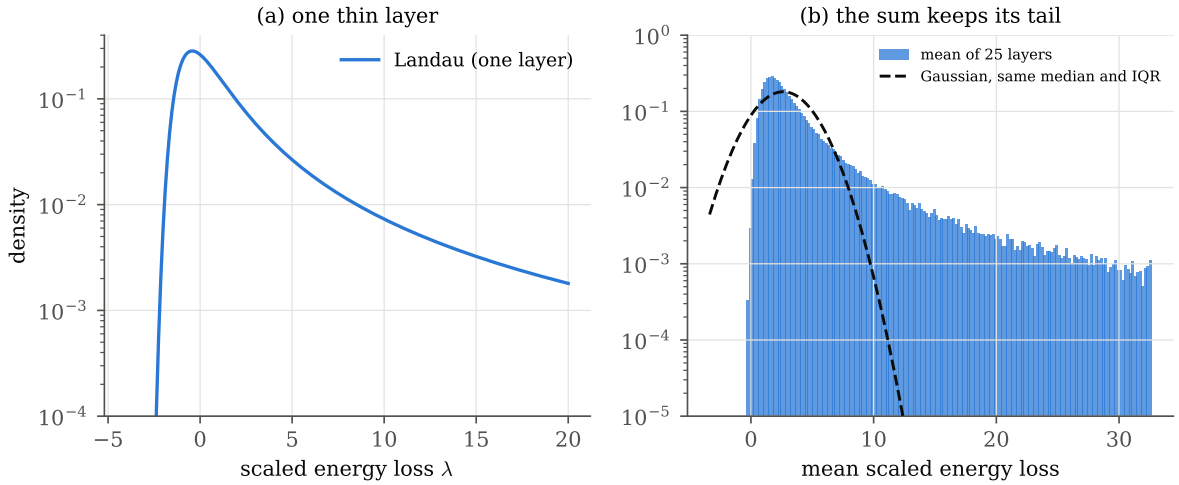


Figure 2.30: (a) The Landau density for one thin layer, on a logarithmic scale: a sharp peak and a long power-law tail to high energy loss. (b) The mean over 25 layers (histogram) against a Gaussian with the same median and interquartile range (dashed). The Gaussian describes the core and misses the tail by many orders of magnitude.

How research uses this. The sums of uniforms and exponentials *verify* a theorem whose answer we know. The same machinery becomes the research instrument when the ingredients are the actual detector effects. Cowan's warning [Cowan p. 148] is that a quoted "1σ" interval covers 68.3% only if the errors are Gaussian in their tails. The only way to learn the tails of a real resolution function (multiple Coulomb scattering has a Gaussian core and a single-scattering Rutherford tail) is a detailed Monte Carlo of the individual contributions. In cosmology the band-power $\widehat{C}_\ell$ is a sum of $2\ell+1$ squared Gaussian modes. At high ℓ the CLT makes it Gaussian;

Taxonomy: Does the CLT apply to my sum?

situation	what the sum does
independent, identical, finite variance	Gaussian; error $\approx 0.07\gamma_1/\sqrt{n}$ in probability
independent, different variances, none dominant	Gaussian (Lindeberg), variance $\sum \sigma_j^2$
one term dominates the variance	looks like that term
infinite variance (Breit–Wigner, Landau, power-law tails)	no Gaussian limit; a *stable* law with the same heavy tail
correlated terms	Gaussian only if correlations die off, with an *effective* n smaller than n
a *product* of positive factors	log-normal: apply the CLT to the logarithm (Section 2.15)
the *maximum*, not the sum	not Gaussian at all: extreme-value laws (Gumbel, Fréchet, Weibull)

at low ℓ it is visibly skewed, and the likelihood of the CMB power spectrum has to be built with that skewness in it [16] (Chapters 7 and 8). Gravitational-wave detector noise is Gaussian in bulk, but "glitches" play the role of the Landau tail and must be vetoed or modelled.

2.11 The Gaussian, third way: the least committal assignment

A calibration note says: "the timing resolution of the chamber is 2 ns rms." Nothing more. No histogram, no list of contributions, no statement about tails. Yet the next step of every analysis is to write down a probability density for the timing residual, because a likelihood needs one. Which density should we write? The first two derivations of the Gaussian do not help here. De Moivre–Laplace needs a binomial, and the central limit theorem needs a sum of many independent pieces. Here we know neither. We know one number.

The third reason the Gaussian is everywhere is of a different kind. It is not about how the noise was produced. It is about what we *know*. The principle of maximum entropy picks out the one distribution that agrees with the facts we have and assumes nothing else. With it we can write a density when all we have is a mean, or a variance, or a covariance matrix. It shows why least squares, which is a Gaussian likelihood, is the natural fit when all we know is the size of the error bars. And it is the same variational principle that produces the Boltzmann distribution in statistical mechanics, so it rests on physics we already trust.

Among all distributions that satisfy a set of known averages, which one should we use, and what does it look like when the known average is a variance?

2.11.1 A die with a loaded average

Start with Sivia's example, because it is discrete and we can see every number [Sivia p. 113–114]. A six-sided die has been rolled very many times. We are told only that the average number of dots was 4.5. What probabilities $p_1, \dots, p_6$ should we assign to the six faces?

The information is a single linear constraint,

$$\sum_{i=1}^{6} i\, p_i = 4.5, \qquad \sum_{i=1}^{6} p_i = 1. \tag{2.62}$$

The fair die, $p_i = 1/6$, has average 3.5, so it is ruled out. But infinitely many assignments remain. Here are three of them:

- $p = (0, 0, 0, \frac{1}{2}, \frac{1}{2}, 0)$: only faces 4 and 5 ever appear. The average is 4.5.
- $p = (0, 0, \frac{1}{2}, 0, 0, \frac{1}{2})$: only faces 3 and 6 appear. The average is $\frac{3}{2} + 3 = 4.5$.
- Something smooth that rises steadily from face 1 to face 6.

The first two are absurd. They claim that faces 1, 2, 3 and 6 (or 1, 2, 4 and 5) *never* appear, and nothing we were told supports that. They contain information we do not have. The third kind feels right. It leans towards the high faces just enough to move the average from 3.5 to 4.5, and otherwise treats the faces evenhandedly. To turn "feels right" into a rule, we need a number that measures how much an assignment commits to beyond the constraints. Then we pick the assignment for which that number is smallest.

2.11.2 Why the measure is the entropy: counting the ways

Sivia's "monkey argument" [Sivia p. 116–118] derives that number by counting, exactly as Boltzmann counted microstates. Imagine M boxes, one per possible outcome, and a team of monkeys throwing N pennies into them completely at random, each penny into each box with probability $1/M$. After the throw, box i holds n_i pennies, with $\sum_i n_i = N$, and we read off the candidate assignment

$$p_i = \frac{n_i}{N}. \tag{2.63}$$

If the candidate violates the constraint we throw it away. If it satisfies it we keep it. Repeat many times. The assignment the monkeys produce most often is the one we should use. *Why?* The monkeys have no preferences. So whatever structure the favoured assignment has beyond the constraint, it has only because overwhelmingly many throws produce it.

How often do the monkeys produce a given set of occupation numbers $\{n_i\}$? There are M^N equally likely sequences of throws (each of N pennies picks one of M boxes). The number of sequences that put n_1 pennies in box 1, n_2 in box 2, and so on, is the multinomial coefficient of Section 2.3. So the frequency is

$$F(\{p_i\}) = \frac{1}{M^N} \, \frac{N!}{n_1!\, n_2! \cdots n_M!}. \tag{2.64}$$

Now let N be very large, so that every n_i is large too, and use Stirling's formula in its crudest form $\ln n! \approx n \ln n - n$ (the leading terms of (2.40)):

$$\begin{aligned}
\ln F &\overset{(1)}{=} -N \ln M + \ln N! - \sum_{i=1}^{M} \ln n_i! \\
&\overset{(2)}{=} -N \ln M + \big(N \ln N - N\big) - \sum_{i=1}^{M} \big(n_i \ln n_i - n_i\big) \\
&\overset{(3)}{=} -N \ln M + N \ln N - \sum_{i=1}^{M} n_i \ln n_i \\
&\overset{(4)}{=} -N \ln M - \sum_{i=1}^{M} n_i \ln \frac{n_i}{N} \\
&\overset{(5)}{=} -N \ln M - N \sum_{i=1}^{M} p_i \ln p_i .
\end{aligned} \tag{2.65}$$

(1) Logarithm of (2.64): a quotient becomes a difference and a product becomes a sum.
(2) Stirling, $\ln n! \approx n \ln n - n$, for $N!$ and for each $n_i!$. The neglected terms grow only like $\ln n$, which is negligible next to n.
(3) The $-N$ and the $+\sum_i n_i = +N$ cancel.
(4) Write $N \ln N = \sum_i n_i \ln N$ (again because $\sum_i n_i = N$) and combine the two logarithms.
(5) Substitute $n_i = N p_i$ from (2.63).

The first term, $-N \ln M$, is the same for every candidate. So the monkeys produce the assignment $\{p_i\}$ with a frequency proportional to e^{NS}, where

$$S = -\sum_{i=1}^{M} p_i \ln p_i. \tag{2.66}$$

Because N is huge, e^{NS} is extremely sharply peaked. Two candidates whose values of S differ by only 0.01 are produced with frequencies in the ratio $e^{0.01N}$, which for $N = 10^4$ pennies is e^{100}. Among all candidates that pass the constraint, essentially only the one with the largest S ever comes out. This is the same mathematics that makes the equilibrium macrostate of a gas overwhelmingly the most probable: Boltzmann's $S = k_B \ln W$ with W the number of microstates is (2.65) with $k_B = 1$.

Unequal boxes. When the possible outcomes are not equally likely to begin with, the entropy acquires a reference distribution. So far all boxes were the same size. Suppose instead that the proposition "the face shows 3, 4, 5 or 6" is one box and "the face shows 1" and "shows 2" are two more. Then a fair monkey should hit the big box four times as often [Sivia p. 118–119]. If box i is hit with probability m_i (with $\sum_i m_i = 1$), the frequency (2.64) becomes the full multinomial probability $\frac{N!}{\prod_i n_i!} \prod_i m_i^{n_i}$, and the same Stirling steps give

$$\frac{1}{N} \ln F = -\sum_{i=1}^{M} p_i \ln \frac{p_i}{m_i}. \tag{2.67}$$

The only new term is $+\sum_i n_i \ln m_i = N \sum_i p_i \ln m_i$, and it joins the logarithm. When all $m_i = 1/M$ we recover (2.66) up to the constant $\ln M$.

Working definition 2.16 (entropy of a distribution, and the MaxEnt principle)

The **entropy** of a discrete assignment $\{p_i\}$ relative to a **measure** $\{m_i\}$ [the probabilities we would assign if we knew nothing at all, often equal] is

$$S = -\sum_i p_i \ln \frac{p_i}{m_i}.$$

For a continuous variable with density $p(x)$ and measure $m(x)$ it is [Sivia (5.29)]

$$S[p] = -\int p(x) \ln \frac{p(x)}{m(x)} \, \mathrm{d}x. \tag{2.68}$$

The **principle of maximum entropy** (MaxEnt): given **testable information** [constraints of the form $\mathbb{E}\, f_k(X) = F_k$, which any proposed distribution either obeys or violates], assign the distribution that maximises S subject to those constraints and to normalisation. *Operational recipe:* write $p \propto m\, e^{-\sum_k \lambda_k f_k}$ and solve the constraint equations for the multipliers λ_k.

The continuous formula needs the measure $m(x)$ because a density has units. If x is a time in ns and we switch to seconds, $p(x)$ is multiplied by 10^9, and $-\int p \ln p \, \mathrm{d}x$ shifts by $-\ln 10^9$.

The ratio p/m has no units, since p and m transform the same way under a change of variable. So (2.68) is the same number in any coordinates [Sivia p. 119]. What is m? Maximising S with only the normalisation constraint gives $p = m$ (Sivia's (5.30), and the case $K = 0$ of the recipe below). So m is the distribution of complete ignorance. For a location variable on the real line it is uniform, and then we may drop it. From now on we do so unless we say otherwise.

2.11.3 The recipe: Lagrange multipliers give exponentials

Maximising the entropy under known averages always gives an exponential. To see why, maximise S with K testable constraints $\sum_i p_i f_k(x_i) = F_k$, $k = 1, \dots, K$. We handle the constraints with Lagrange multipliers [one extra unknown per constraint, chosen at the end so the constraint holds]: find the stationary point of

$$Q = -\sum_i p_i \ln \frac{p_i}{m_i} + \lambda_0\Big(1 - \sum_i p_i\Big) + \sum_{k=1}^{K} \lambda_k \Big(F_k - \sum_i p_i f_k(x_i)\Big).$$

Differentiate with respect to one particular p_j. Only the terms containing p_j contribute:

$$\begin{aligned} \frac{\partial Q}{\partial p_j} &\overset{(1)}{=} -\ln \frac{p_j}{m_j} - 1 - \lambda_0 - \sum_{k=1}^{K} \lambda_k f_k(x_j) \overset{(2)}{=} 0 \\ \Longrightarrow \quad p_j &\overset{(3)}{=} m_j\, e^{-(1+\lambda_0)} \exp\Big[-\sum_{k=1}^{K} \lambda_k f_k(x_j)\Big] \overset{(4)}{=} \frac{m_j}{Z} \exp\Big[-\sum_{k=1}^{K} \lambda_k f_k(x_j)\Big]. \end{aligned} \tag{2.69}$$

(1) $\frac{\mathrm{d}}{\mathrm{d}p}(-p\ln(p/m)) = -\ln(p/m) - 1$, and each constraint is linear in p_j with coefficient -1 or $-f_k(x_j)$.
(2) A maximum of Q is a stationary point.
(3) Exponentiate and multiply by m_j.
(4) Normalisation fixes $e^{1+\lambda_0} = Z \equiv \sum_j m_j \exp[-\sum_k \lambda_k f_k(x_j)]$, the **partition function**.

Is it a maximum and not a minimum? Yes, and the only one. The second derivatives are $\partial^2 Q/\partial p_j \partial p_l = -\delta_{jl}/p_j < 0$, so Q is strictly concave, and a concave function on the convex set of allowed $\{p_i\}$ has exactly one stationary point, its global maximum. There is no risk of picking a saddle.

Every physicist has seen (2.69) before. With one constraint, the mean energy $\sum_i p_i E_i = U$, it reads $p_j = e^{-\beta E_j}/Z$, the Boltzmann distribution, with the Lagrange multiplier β playing the role of $1/k_B T$. Temperature *is* the Lagrange multiplier that enforces a given mean energy. Jaynes's reading of statistical mechanics, which Sivia summarises [Sivia p. 118], is exactly this: the canonical ensemble is the least committal assignment compatible with a known average energy.

Example: the die. For the die, $m_i = 1/6$ and there is one constraint, $f(i) = i$, $F = 4.5$. Writing $\lambda \equiv -\lambda_1$ for convenience, (2.69) says

$$p_i = \frac{e^{\lambda i}}{Z(\lambda)}, \qquad Z(\lambda) = \sum_{j=1}^{6} e^{\lambda j}, \qquad \sum_{i=1}^{6} i\, p_i = \frac{\mathrm{d}\ln Z}{\mathrm{d}\lambda} = 4.5. \tag{2.70}$$

This equation for λ has exactly one solution. The mean $\mathrm{d}\ln Z/\mathrm{d}\lambda$ increases with λ, because its derivative $\mathrm{d}^2 \ln Z/\mathrm{d}\lambda^2 = \sum i^2 p_i - (\sum i p_i)^2$ is a variance and so positive. At $\lambda = 0$ the mean is

3.5, and as $\lambda \to \infty$ it tends to 6. So it passes 4.5 exactly once, and a root finder gives $\lambda = 0.3710$. The probabilities are

$$(p_1, \dots, p_6) = (0.0544,\ 0.0788,\ 0.1142,\ 0.1654,\ 0.2398,\ 0.3475),$$

a geometric progression with ratio $e^{\lambda} \approx 1.45$: every step up in dots multiplies the probability by the same factor (Figure 2.31a, which reproduces Sivia's Fig. 5.2). The entropy is $S = 1.6136$, below the fair-die value $\ln 6 = 1.7918$: the constraint costs entropy, as it must.

Checking by brute force. No random die with the same average beats the MaxEnt die. The script draws 4×10^6 random assignments uniformly over all possible dice (a Dirichlet$(1, \dots, 1)$ distribution, which we meet in Section 2.15). It keeps the 7888 whose mean lies within 0.005 of 4.5, and Figure 2.31b histograms their entropies. The largest entropy among them is 1.6139 (for a die of mean 4.4966), against 1.6136 for the MaxEnt die. The two are very close, and a random die can even come out a hair ahead. That is not a defeat: the window lets in means slightly below 4.5, closer to the fair 3.5, and a smaller tilt costs less entropy. Comparing like with like, each candidate against the MaxEnt die of *its own* mean, every one falls short, by at least 9.6×10^{-4}.

Listing 2.15: The MaxEnt die, and random dice with the same mean (`code/ch02/23_maxent.py`)

```
38 def die_mean(lam):
39     w = np.exp(lam * faces)
40     return np.sum(faces * w) / np.sum(w)
41
42
43 lam = optimize.brentq(lambda l: die_mean(l) - target, -5, 5)
44 p_me = np.exp(lam * faces)
45 p_me /= p_me.sum()
46 S_me = -np.sum(p_me * np.log(p_me))
47
48 # random distributions on the 6-simplex with mean 4.5 (accept within 0.005)
49 cand = rng.dirichlet(np.ones(6), size=4_000_000)
50 keep = cand[np.abs(cand @ faces - target) < 0.005]
51 S_rand = -np.sum(np.where(keep > 0, keep * np.log(keep), 0.0), axis=1)
```

Lines 38–40 compute the mean $\sum_i i\, p_i$ of the exponential family (2.70) for a given λ. Line 43 solves "mean = 4.5" for λ with Brent's bracketing root finder; the monotonicity argument above is what guarantees that the bracket $[-5, 5]$ contains exactly one root. Lines 49–51 draw random dice, keep those with the right mean, and compute their entropies, using the convention $0 \ln 0 = 0$.

Example: the kangaroo problem. Is entropy really special, or would any function that rewards spreading out do? Sivia's kangaroo problem answers this [Sivia p. 114–116]. We are told that one third of all kangaroos have blue eyes and one third are left-handed. What fraction are both? Call the four probabilities p_1 (blue, left), p_2 (blue, right), p_3 (not blue, left), p_4 (not blue, right). The two facts and normalisation leave one free number, $x = p_1$ (Figure 2.32a):

$$p_1 = x, \quad p_2 = \tfrac{1}{3} - x, \quad p_3 = \tfrac{1}{3} - x, \quad p_4 = \tfrac{1}{3} + x, \qquad 0 \le x \le \tfrac{1}{3}.$$

Common sense says $x = 1/9$: with no information about a link between eye colour and handedness, we should not invent one, and independence gives $\frac{1}{3} \times \frac{1}{3}$. Maximise the entropy:

$$\frac{\mathrm{d}S}{\mathrm{d}x} \overset{(1)}{=} \frac{\mathrm{d}}{\mathrm{d}x}\Big[-x \ln x - 2\big(\tfrac{1}{3} - x\big) \ln\big(\tfrac{1}{3} - x\big) - \big(\tfrac{1}{3} + x\big) \ln\big(\tfrac{1}{3} + x\big)\Big]$$

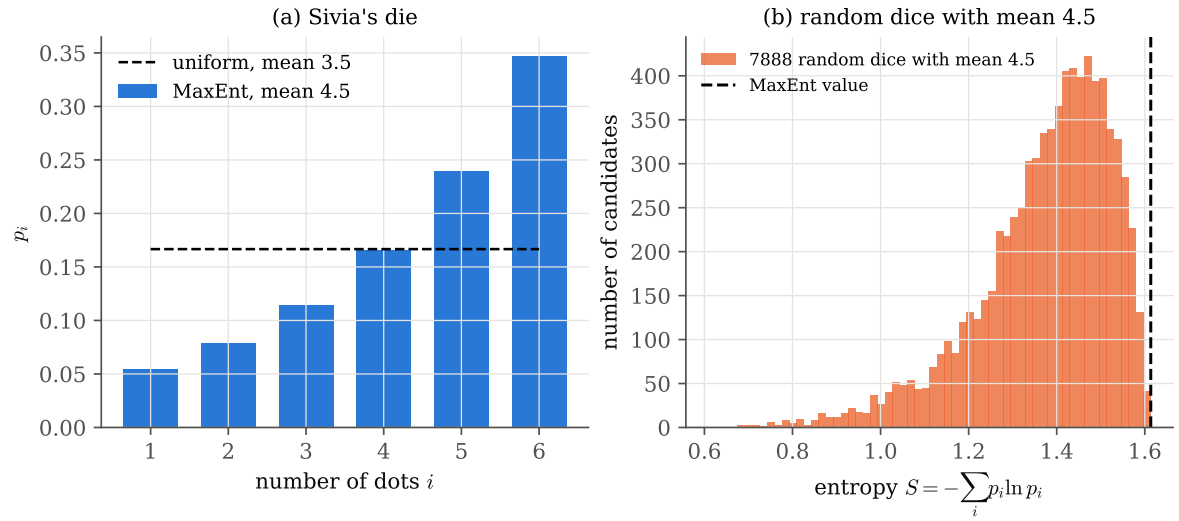


Figure 2.31: (a) The maximum-entropy die for a known average of 4.5 (bars) against the fair die (dashed). The probabilities grow by a constant factor from face to face. (b) Entropies of several thousand randomly drawn dice whose average is also 4.5. They pile up to the left of the MaxEnt value (dashed line), and compared with the MaxEnt die of its own average every one falls short: every other assignment assumes something extra.

$$\begin{aligned} &\overset{(2)}{=} -\ln x - 1 + 2\ln\left(\tfrac{1}{3} - x\right) + 2 - \ln\left(\tfrac{1}{3} + x\right) - 1 \\ &\overset{(3)}{=} \ln \frac{\left(\frac{1}{3} - x\right)^2}{x\left(\frac{1}{3} + x\right)}. \end{aligned} \tag{2.71}$$

(1) Entropy (2.66) of the four-cell table.
(2) $\frac{\mathrm{d}}{\mathrm{d}x}(-u \ln u) = -(\ln u + 1)\frac{\mathrm{d}u}{\mathrm{d}x}$, with $\frac{\mathrm{d}u}{\mathrm{d}x} = +1, -1, -1, +1$ for the four cells.
(3) The constants cancel ($-1 + 2 - 1 = 0$) and the logarithms combine.

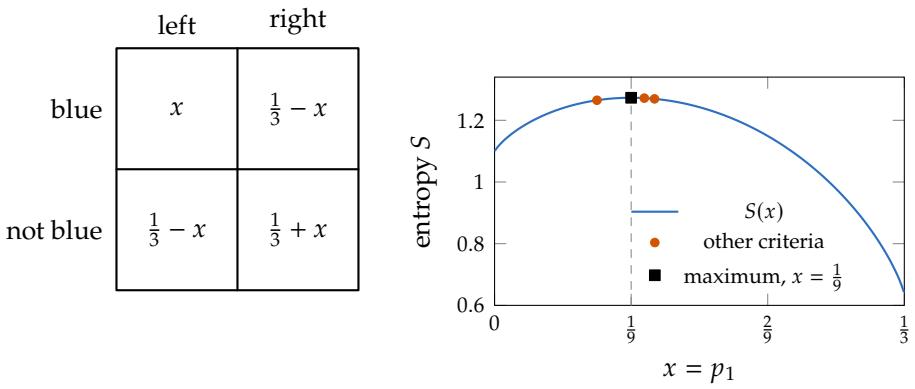
Setting this to zero gives $(\frac{1}{3} - x)^2 = x(\frac{1}{3} + x)$, that is $\frac{1}{9} - \frac{2}{3}x + x^2 = \frac{1}{3}x + x^2$, so $x = \frac{1}{9}$. Entropy reproduces independence exactly.

Other plausible-looking criteria do not. Maximising $-\sum p_i^2$ gives $\frac{\mathrm{d}}{\mathrm{d}x}\left[x^2 + 2(\frac{1}{3} - x)^2 + (\frac{1}{3} + x)^2\right] = 8x - \frac{2}{3} = 0$, so $x = \frac{1}{12}$: a negative correlation nobody asked for. The script maximises all four of Sivia's candidates numerically and finds $x = 0.1111$ (entropy), 0.0833 ($-\sum p_i^2$), 0.1301 ($\sum \ln p_i$) and 0.1218 ($\sum \sqrt{p_i}$), reproducing his Table 5.1. Only functions monotonically related to S give 1/9 (Sivia quotes Skilling for the proof), and that is the strongest argument that entropy is the right measure of "assuming nothing extra".

2.11.4 Continuous MaxEnt: a variance gives a Gaussian

Back to the timing residual of the calibration note, where all we were given was an rms of 2 ns. The testable information is that the residual x has mean μ and variance σ^2, and x can be any real number. The question: which density has the largest entropy under these two constraints? The continuous version of the recipe replaces sums by integrals and the partial derivative by a functional derivative [the change of Q when p is changed by a small bump δp at one point x, divided by the area of the bump]. We need

$$Q[p] = -\int p \ln p \,\mathrm{d}x + \lambda_0\left(1 - \int p \,\mathrm{d}x\right) + \lambda_1\left(\sigma^2 - \int (x - \mu)^2 p \,\mathrm{d}x\right)$$



(a) **(b)**

Figure 2.32: The kangaroo problem. (a) The two known marginals, one third blue-eyed and one third left-handed, leave a single free number x in the 2×2 table. (b) The entropy of the table as a function of x. It peaks at $x = 1/9$, the independent answer. The orange dots mark the maximisers of Sivia's three alternative criteria, which all build in a correlation that the information does not contain.

to be stationary [Sivia p. 121–122]. Then

$$\frac{\delta Q}{\delta p(x)} \overset{(1)}{=} -\ln p(x) - 1 - \lambda_0 - \lambda_1 (x-\mu)^2 \overset{(2)}{=} 0 \quad \Longrightarrow \quad p(x) = e^{-1-\lambda_0}\, e^{-\lambda_1 (x-\mu)^2},$$

$$1 \overset{(3)}{=} e^{-1-\lambda_0} \int_{-\infty}^{\infty} e^{-\lambda_1 (x-\mu)^2} \mathrm{d}x \overset{(4)}{=} e^{-1-\lambda_0} \sqrt{\frac{\pi}{\lambda_1}},$$

$$\sigma^2 \overset{(5)}{=} \sqrt{\frac{\lambda_1}{\pi}} \int_{-\infty}^{\infty} (x-\mu)^2 e^{-\lambda_1 (x-\mu)^2} \mathrm{d}x \overset{(6)}{=} \frac{1}{2\lambda_1} \quad \Longrightarrow \quad p(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(x-\mu)^2/2\sigma^2}. \tag{2.72}$$

(1) The same differentiation as in (2.69), applied pointwise: changing p near x changes each integral by its integrand at x times the size of the change.
(2) Stationarity; then solve for p. For the integral to exist we need $\lambda_1 > 0$.
(3) The normalisation constraint.
(4) The Gaussian integral (2.36), rescaled: substitute $z = \sqrt{2\lambda_1}(x - \mu)$.
(5) The variance constraint, with the constant from line (4), $e^{-1-\lambda_0} = \sqrt{\lambda_1/\pi}$.
(6) The normalised density $\sqrt{\lambda_1/\pi}\, e^{-\lambda_1 (x-\mu)^2}$ is a Gaussian of variance $1/(2\lambda_1)$, by (2.38). Solving $\sigma^2 = 1/(2\lambda_1)$ gives $\lambda_1 = 1/(2\sigma^2)$.

So the maximum-entropy density with a given mean and variance is the Gaussian $\mathcal{N}(\mu, \sigma^2)$ of (2.35). Nothing was assumed about how the residual was produced.

A second proof, and a proof that it is the maximum. The Lagrange argument finds a stationary point. A short inequality shows directly that no other density with the same variance does better. We use one elementary fact, $\ln t \le t - 1$ for $t > 0$ (the tangent line at $t = 1$ lies above the concave logarithm). Let p be any density with mean μ and variance σ^2, and let g be the Gaussian with the same two numbers. Then

$$\begin{aligned} S[p] - S[g] &\overset{(1)}{=} -\int p \ln p \,\mathrm{d}x + \int g \ln g \,\mathrm{d}x \\ &\overset{(2)}{=} -\int p \ln p \,\mathrm{d}x + \int p \ln g \,\mathrm{d}x \\ &\overset{(3)}{=} \int p \ln \frac{g}{p} \,\mathrm{d}x \end{aligned}$$

$$\overset{(4)}{\leq} \int p\Big(\frac{g}{p} - 1\Big)\mathrm{d}x \overset{(5)}{=} \int g\,\mathrm{d}x - \int p\,\mathrm{d}x = 0. \tag{2.73}$$

(1) Definition (2.68) with uniform measure.
(2) $\ln g(x) = -\frac{1}{2}\ln(2\pi\sigma^2) - (x-\mu)^2/2\sigma^2$ is a constant plus a multiple of $(x-\mu)^2$. Its average is therefore fixed by the normalisation and the variance alone, and p and g share both.
(3) Combine the two integrals.
(4) $\ln t \leq t - 1$ with $t = g/p$, at every x where $p > 0$.
(5) Both densities integrate to one.

Equality in step (4) needs $g/p = 1$ everywhere, so the Gaussian is the *unique* maximiser. The quantity $\int p \ln(p/g)\,\mathrm{d}x \geq 0$ we just met is the **Kullback–Leibler divergence** of p from g. It will return in Chapter 3 (maximum likelihood minimises it) and in Chapter 5. The value itself is

$$S[g] = -\int g \ln g\,\mathrm{d}x \overset{(1)}{=} \tfrac{1}{2}\ln(2\pi\sigma^2) + \frac{\mathbb{E}(x-\mu)^2}{2\sigma^2} = \tfrac{1}{2}\ln\big(2\pi e\sigma^2\big),$$

where (1) inserts $\ln g$ from step (2) above.

Key idea: The third origin of the Gaussian

Of all densities on the real line with a given mean μ and variance σ^2, the Gaussian $\mathcal{N}(\mu, \sigma^2)$ has the largest entropy, $\frac{1}{2}\ln(2\pi e\sigma^2)$. Assigning it claims nothing about the data except the mean and the variance. Assigning anything else claims more.

Example: five densities with unit variance. The script evaluates the entropy of five familiar densities, each scaled to variance one: uniform 1.2425, Laplace 1.3466, triangular 1.3959, logistic 1.4046, Gaussian $1.4189 = \frac{1}{2}\ln(2\pi e)$. The Gaussian wins, but the others are close. The logistic is only 0.014 behind: two densities with nearly the same core can have nearly the same entropy. Figure 2.33 shows three of the five, and why they lose. The uniform "wastes" its variance by having no tails at all. The Laplace concentrates probability in a cusp and pays for its variance with long exponential tails. The Gaussian spreads its probability in the most even way the variance allows.

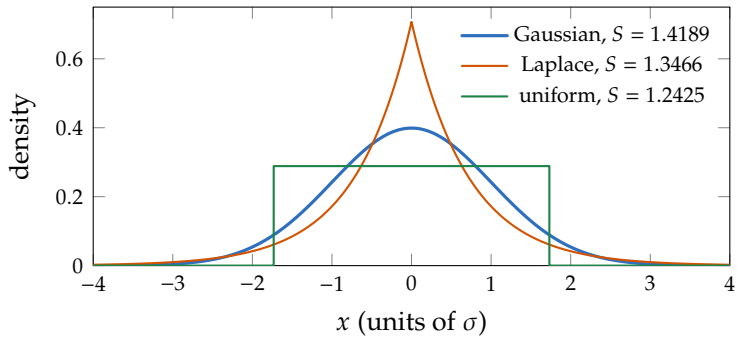


Figure 2.33: Three densities with the same mean (0) and variance (1). The uniform box stops dead at $\pm\sqrt{3}$; the Laplace density has a sharp peak and fat exponential tails. Among all densities with this variance the Gaussian has the largest entropy, and the numbers in the legend show by how much it beats these two.

2.11.5 The family tree: other constraints, other distributions

Each common distribution is the MaxEnt answer to some piece of partial knowledge. The recipe (2.69) works like a machine: feed in the constraint functions f_k and the range of x, and

out comes $p \propto m\, e^{-\sum \lambda_k f_k}$. Sivia runs it for the common cases [Sivia p. 120–125]:

- *Nothing but the range* $[a, b]$. Then $p = m$, the uniform density $1/(b-a)$ (Bernoulli's principle of insufficient reason).
- *A mean* μ*, with* $x \geq 0$. One constraint $f = x$, so $p \propto e^{-\lambda_1 x}$. Normalisation on $[0, \infty)$ gives $\lambda_1 e^{-\lambda_1 x}$ and the mean gives $\lambda_1 = 1/\mu$: the exponential $p(x) = \mu^{-1} e^{-x/\mu}$ [Sivia (5.33)], the waiting-time law of (2.18). The same computation on the line with a mean alone fails: $e^{-\lambda_1 x}$ cannot be normalised on $(-\infty, \infty)$, which is the recipe's way of saying that a mean alone does not specify a spread.
- *A variance on the whole line.* The Gaussian, (2.72).
- *A mean absolute deviation* $\mathbb{E}\,|x-\mu| = \epsilon$. $f = |x-\mu|$ gives the Laplace (double-exponential) density $\frac{1}{2\epsilon} e^{-|x-\mu|/\epsilon}$ [Sivia (5.39)–(5.40)]. Its log-likelihood is $-\sum |x_k - \mu_k|/\epsilon_k$: the ℓ_1 "least absolute deviations" fit, which is more robust to outliers than least squares.
- *Many variables with known individual variances.* With N data x_k, predictions μ_k and known $\mathbb{E}(x_k - \mu_k)^2 = \sigma_k^2$, the multidimensional version gives the product of Gaussians [Sivia (5.36)–(5.38)]

$$p(\{x_k\}) = \prod_{k=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_k} \exp\Big[-\frac{(x_k-\mu_k)^2}{2\sigma_k^2}\Big], \tag{2.74}$$

  which is exactly the likelihood behind least squares, $-2\ln p = \chi^2 + \text{const}$. From the MaxEnt point of view least squares does not assume that the noise *is* Gaussian. It is the honest likelihood when all we know is the size of each error bar. If we also know the covariances $\mathbb{E}(x_j - \mu_j)(x_k - \mu_k)$, the same recipe returns the correlated multivariate Gaussian of the next section.
- *A mean count* $\mathbb{E}\,N = \mu$ *of events in an interval.* Here the measure matters. Cut the interval into M tiny cells, each of which can hold an event or not. The number of cell patterns giving N events among M indistinguishable cells is $M^N/N!$ for $M \gg N$, so $m(N) \propto M^N/N!$ [Sivia (5.43)]. Then

$$p(N) \propto \frac{M^N}{N!} e^{-\lambda N} = \frac{(Me^{-\lambda})^N}{N!} \quad \Longrightarrow \quad p(N) = e^{-\mu}\frac{\mu^N}{N!},$$

  where normalisation, $\sum_N (Me^{-\lambda})^N/N! = e^{Me^{-\lambda}}$, and the mean condition identify $Me^{-\lambda} = \mu$. The unknown M drops out, and we recover the Poisson distribution of Section 2.4 from knowledge of the mean alone.
- *The mean kinetic energy of a gas molecule.* The constraint $\mathbb{E}\,\frac{1}{2}m|v|^2 = \frac{3}{2}k_B T$ with uniform measure on velocity space gives a Gaussian for each velocity component with variance $k_B T/m$. The speed distribution that follows is the Maxwell–Boltzmann law of Section 2.15.

Watch out: MaxEnt is a statement about knowledge, not about the noise

The MaxEnt Gaussian is the best assignment *given only* a variance. If the real residuals have a Landau tail (Section 2.10.6) or a Breit–Wigner shape, the Gaussian will misjudge the tails badly, and the cure is not to argue about entropy but to put the extra knowledge in as a constraint (a known tail fraction, a known fourth moment) or to model the mechanism. And when the variance is infinite, as for a Breit–Wigner, there is no variance constraint to impose in the first place.

How the three origins fit together. The three reasons for the Gaussian agree, and that is not a coincidence. De Moivre–Laplace and the central limit theorem are statements about *mechanisms*: many small independent contributions added together. MaxEnt is a statement about *information*: we know a variance and nothing else. The link is that adding independent pieces and rescaling to fixed variance can only increase the entropy (convolution smooths). So the central limit theorem is the dynamics that drives a sum towards the maximum-entropy state, like a gas relaxing to equilibrium. When a physicist writes "Gaussian errors" she is usually relying on all of them at once: the CLT for why the errors should be close to Gaussian, and MaxEnt for why, knowing only the error bar, nothing else would be more honest.

The same principle supplies the forward model $P(\boldsymbol{d} \mid \boldsymbol{\theta})$, the probability of the data $\boldsymbol{d}$ given the parameters $\boldsymbol{\theta}$, whenever only the covariance of the noise is known. The prior assignments of Chapter 5 use it too. And the MaxEnt die's exponential form is the first example of the exponential families that make maximum likelihood and conjugate priors tractable (Chapters 3 and 5).

2.12 Many correlated Gaussians: the multivariate normal

Most measurements come as several numbers that move together, and one object describes them all: the multivariate Gaussian. A straight-line track fit returns two numbers, an intercept and a slope, and they are never independent: tilt the line a little and the intercept has to move to keep the line going through the hits. A fit of a CMB power spectrum returns $\Omega_m h^2$ and n_s, and they, too, move together. A CMB temperature map is a list of 10^6 pixel values, each Gaussian, and neighbouring pixels are strongly correlated. A stretch of gravitational-wave strain is a list of 10^5 samples whose correlations are set by the detector's noise spectrum. The Fisher matrix of Chapter 10 approximates every posterior by a multivariate Gaussian, and a Gaussian random field (Chapter 7) is a multivariate Gaussian with very many components.

Four things about it are used constantly. How to *sample* it, which follows from building it out of independent standard normals. The shape of its contours, the ellipses of every parameter-estimation plot. What fraction of the probability each ellipse contains. And two operations: *marginalising* (averaging out a nuisance variable) and *conditioning* (fixing one variable at a known value).

What is the joint density of several correlated Gaussian quantities, how do we draw from it, and what are the distributions of a subset of them, either ignoring the rest or with the rest held fixed?

2.12.1 Building it: a linear map of independent standard normals

Start from what we know. Let $Z_1, \dots, Z_k$ be independent $\mathcal{N}(0,1)$ variables and collect them into a column vector $\boldsymbol{Z}$. Independence means the joint density is the product of the one-dimensional densities [Wasserman p. 40]:

$$p_{\boldsymbol{Z}}(\boldsymbol{z}) = \prod_{j=1}^{k} \frac{e^{-z_j^2/2}}{\sqrt{2\pi}} = \frac{1}{(2\pi)^{k/2}} \exp\!\left(-\tfrac{1}{2}\boldsymbol{z}^{\mathsf{T}}\boldsymbol{z}\right). \tag{2.75}$$

Its contours, $\boldsymbol{z}^{\mathsf{T}}\boldsymbol{z} = \text{const}$, are spheres, and its components are uncorrelated. Correlations appear when we mix the components linearly. Take a fixed invertible $k \times k$ matrix $\mathbf{A}$ and a fixed vector $\boldsymbol{\mu}$ and define

$$\boldsymbol{X} = \boldsymbol{\mu} + \mathbf{A}\boldsymbol{Z}. \tag{2.76}$$

Its mean is $\mathbb{E}\,\boldsymbol{X} = \boldsymbol{\mu} + \mathbf{A}\,\mathbb{E}\,\boldsymbol{Z} = \boldsymbol{\mu}$. Its covariance matrix [the $k \times k$ table of all $\mathrm{Cov}(X_i, X_j)$, with the variances on the diagonal] is

$$\mathbf{C} \equiv \mathbb{E}\big[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^{\mathsf{T}}\big] \overset{(1)}{=} \mathbb{E}\big[\mathbf{A}\boldsymbol{Z}\boldsymbol{Z}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}}\big] \overset{(2)}{=} \mathbf{A}\,\mathbb{E}\big[\boldsymbol{Z}\boldsymbol{Z}^{\mathsf{T}}\big]\mathbf{A}^{\mathsf{T}} \overset{(3)}{=} \mathbf{A}\mathbf{A}^{\mathsf{T}}. \tag{2.77}$$

(1) $\boldsymbol{X} - \boldsymbol{\mu} = \mathbf{A}\boldsymbol{Z}$, and $(\mathbf{A}\boldsymbol{Z})^{\mathsf{T}} = \boldsymbol{Z}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}}$.
(2) $\mathbf{A}$ is not random, and expectation is linear, so it can be taken outside on both sides.
(3) $\mathbb{E}(Z_i Z_j) = \delta_{ij}$: unit variances and, by independence, zero covariances. So $\mathbb{E}\,\boldsymbol{Z}\boldsymbol{Z}^{\mathsf{T}} = \mathbb{1}$.

We know the mean and covariance of $\boldsymbol{X}$; now we want its density. Change variables. The map $\boldsymbol{z} \mapsto \boldsymbol{x} = \boldsymbol{\mu} + \mathbf{A}\boldsymbol{z}$ is linear with inverse $\boldsymbol{z} = \mathbf{A}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$. A small box of volume $\mathrm{d}^k z$ is mapped to a parallelepiped of volume $|\det \mathbf{A}|\,\mathrm{d}^k z$, so probability conservation, $p_{\boldsymbol{X}}\,\mathrm{d}^k x = p_{\boldsymbol{Z}}\,\mathrm{d}^k z$, gives

$$\begin{aligned} p_{\boldsymbol{X}}(\boldsymbol{x}) &\overset{(1)}{=} p_{\boldsymbol{Z}}\big(\mathbf{A}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\big)\,\frac{1}{|\det \mathbf{A}|} \\ &\overset{(2)}{=} \frac{1}{(2\pi)^{k/2}|\det \mathbf{A}|}\exp\Big[-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathsf{T}}(\mathbf{A}^{-1})^{\mathsf{T}}\mathbf{A}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\Big] \\ &\overset{(3)}{=} \frac{1}{(2\pi)^{k/2}(\det \mathbf{C})^{1/2}}\exp\Big[-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathsf{T}}\mathbf{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\Big]. \end{aligned} \tag{2.78}$$

(1) The multidimensional change-of-variables rule: the Jacobian of the inverse map is $\det \mathbf{A}^{-1} = 1/\det \mathbf{A}$.
(2) Insert (2.75) with $\boldsymbol{z} = \mathbf{A}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$, so $\boldsymbol{z}^{\mathsf{T}}\boldsymbol{z} = (\boldsymbol{x}-\boldsymbol{\mu})^{\mathsf{T}}(\mathbf{A}^{-1})^{\mathsf{T}}\mathbf{A}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$.
(3) $(\mathbf{A}^{-1})^{\mathsf{T}}\mathbf{A}^{-1} = (\mathbf{A}\mathbf{A}^{\mathsf{T}})^{-1} = \mathbf{C}^{-1}$, and $\det \mathbf{C} = \det \mathbf{A}\,\det \mathbf{A}^{\mathsf{T}} = (\det \mathbf{A})^2$.

The matrix $\mathbf{A}$ has disappeared: only $\mathbf{C} = \mathbf{A}\mathbf{A}^{\mathsf{T}}$ is left. So the distribution is fixed by the mean and the covariance alone, and any two matrices with the same $\mathbf{A}\mathbf{A}^{\mathsf{T}}$ give the same distribution. That is the density quoted by all our sources [Wasserman (2.10)], [Cowan (2.28)].

Working definition 2.17 (multivariate normal distribution)

A random vector $\boldsymbol{X} \in \mathbb{R}^k$ is **multivariate normal**, $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$, if its density is (2.78), where $\boldsymbol{\mu}$ is its mean vector and $\mathbf{C}$ its covariance matrix, which must be symmetric and **positive definite** [$\boldsymbol{a}^{\mathsf{T}}\mathbf{C}\boldsymbol{a} > 0$ for every nonzero vector $\boldsymbol{a}$. This says that every linear combination $\boldsymbol{a}^{\mathsf{T}}\boldsymbol{X}$ has a nonzero variance]. It has k means plus $k(k+1)/2$ independent entries of $\mathbf{C}$ as parameters [Cowan p. 33]. *Operational recipe (sampling):* factor $\mathbf{C} = \mathbf{L}\mathbf{L}^{\mathsf{T}}$ (Cholesky, $\mathbf{L}$ lower triangular, or the symmetric square root $\mathbf{C}^{1/2}$ of [Wasserman Thm 2.43]), draw k independent standard normals $\boldsymbol{z}$, and return $\boldsymbol{\mu} + \mathbf{L}\boldsymbol{z}$. *Inverse (whitening):* $\mathbf{L}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \sim \mathcal{N}(0, \mathbb{1})$.

Why must $\mathbf{C}$ be positive definite? Because $\boldsymbol{a}^{\mathsf{T}}\mathbf{C}\boldsymbol{a} = \mathrm{Var}(\boldsymbol{a}^{\mathsf{T}}\boldsymbol{X})$, and a variance cannot be negative. If it is zero for some $\boldsymbol{a} \neq 0$, then $\boldsymbol{a}^{\mathsf{T}}\boldsymbol{X}$ is a constant: one component is an exact linear function of the others, the distribution lives on a lower-dimensional plane, $\det \mathbf{C} = 0$, and (2.78) does not exist. (A fitted $\mathbf{C}$ that is singular or nearly singular almost always means a parameter that the data cannot distinguish from a combination of the others.)

Example: two dimensions, written out. For $k = 2$ write $\mathbf{C} = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$, where $\rho = \mathrm{Cov}(x,y)/(\sigma_x\sigma_y)$ is the **correlation coefficient**, $-1 < \rho < 1$. The inverse of a 2×2 matrix is its adjugate over its determinant, so with $u = (x-\mu_x)/\sigma_x$ and $v = (y-\mu_y)/\sigma_y$,

$$(\boldsymbol{x}-\boldsymbol{\mu})^{\mathsf{T}}\mathbf{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \overset{(1)}{=} \frac{1}{\sigma_x^2\sigma_y^2(1-\rho^2)}\begin{pmatrix} x-\mu_x & y-\mu_y \end{pmatrix}\begin{pmatrix} \sigma_y^2 & -\rho\sigma_x\sigma_y \\ -\rho\sigma_x\sigma_y & \sigma_x^2 \end{pmatrix}\begin{pmatrix} x-\mu_x \\ y-\mu_y \end{pmatrix}$$

$$\overset{(2)}{=} \frac{u^2 - 2\rho uv + v^2}{1-\rho^2}, \tag{2.79}$$

(1) $\det \mathbf{C} = \sigma_x^2\sigma_y^2 - \rho^2\sigma_x^2\sigma_y^2 = \sigma_x^2\sigma_y^2(1-\rho^2)$, and the adjugate swaps the diagonal entries and flips the sign of the off-diagonal ones.
(2) Multiply out: the three terms are $\sigma_y^2(x-\mu_x)^2$, $-2\rho\sigma_x\sigma_y(x-\mu_x)(y-\mu_y)$ and $\sigma_x^2(y-\mu_y)^2$; dividing each by $\sigma_x^2\sigma_y^2$ turns them into u^2, $-2\rho uv$, v^2.

and $(\det \mathbf{C})^{1/2} = \sigma_x\sigma_y\sqrt{1-\rho^2}$. This gives the bivariate normal density of [Cowan (2.30)] and [CasellaBerger Def. 4.5.10],

$$p(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\Big[-\frac{u^2 - 2\rho uv + v^2}{2(1-\rho^2)}\Big]. \tag{2.80}$$

Our running example, used by the script and in the rest of this section, has $\sigma_x = 1$, $\sigma_y = 2$, $\rho = 0.8$ (and means $\mu_x = 1$, $\mu_y = 2$), so $\mathbf{C} = \begin{pmatrix} 1 & 1.6 \\ 1.6 & 4 \end{pmatrix}$ and $\det \mathbf{C} = 4 - 2.56 = 1.44$. Its Cholesky factor is $\mathbf{L} = \begin{pmatrix} 1 & 0 \\ 1.6 & 1.2 \end{pmatrix}$; check: $\mathbf{L}\mathbf{L}^\mathsf{T} = \begin{pmatrix} 1 & 1.6 \\ 1.6 & 1.6^2 + 1.2^2 \end{pmatrix} = \mathbf{C}$. So a sample is $x = \mu_x + z_1$, $y = \mu_y + 1.6z_1 + 1.2z_2$: the y component copies 80% of the x fluctuation, in units of its own width, and adds fresh noise of its own. That sentence is what a correlation of 0.8 *means* mechanically.

2.12.2 The shape: ellipses and how much they contain

The contours of the multivariate Gaussian are ellipsoids, and rotating to their axes removes the correlations. To see this, note that the density (2.78) depends on $\boldsymbol{x}$ only through the quadratic form

$$Q(\boldsymbol{x}) = (\boldsymbol{x}-\boldsymbol{\mu})^\mathsf{T}\mathbf{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}),$$

so its contours are the surfaces $Q = c$. To find their shape, diagonalise $\mathbf{C}$. A real symmetric matrix has an orthonormal basis of eigenvectors: $\mathbf{C} = \mathbf{R}\boldsymbol{\Lambda}\mathbf{R}^\mathsf{T}$ with $\mathbf{R}$ a rotation whose columns are the eigenvectors and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$, all $\lambda_i > 0$ by positive definiteness. In rotated coordinates $\boldsymbol{y} = \mathbf{R}^\mathsf{T}(\boldsymbol{x}-\boldsymbol{\mu})$,

$$Q \overset{(1)}{=} \boldsymbol{y}^\mathsf{T}\boldsymbol{\Lambda}^{-1}\boldsymbol{y} \overset{(2)}{=} \sum_{i=1}^{k} \frac{y_i^2}{\lambda_i} = c, \tag{2.81}$$

(1) $\mathbf{C}^{-1} = \mathbf{R}\boldsymbol{\Lambda}^{-1}\mathbf{R}^\mathsf{T}$ because $\mathbf{R}^{-1} = \mathbf{R}^\mathsf{T}$.
(2) $\boldsymbol{\Lambda}^{-1}$ is diagonal.

an ellipsoid with principal axes along the eigenvectors and semi-axes $\sqrt{c\lambda_i}$. In the y coordinates the components are independent, $y_i \sim \mathcal{N}(0, \lambda_i)$: rotating to the eigenbasis *removes* the correlations. This is the same step as finding the normal modes of coupled oscillators.

For the running example ($\sigma_x = 1$, $\sigma_y = 2$, $\rho = 0.8$), the eigenvalues solve $\lambda^2 - (\mathrm{tr}\,\mathbf{C})\lambda + \det\mathbf{C} = \lambda^2 - 5\lambda + 1.44 = 0$, so $\lambda = \frac{1}{2}\big(5 \pm \sqrt{25 - 5.76}\big)$, that is $\lambda_1 = 4.693$ and $\lambda_2 = 0.307$. The major axis, from $(1-\lambda_1)v_1 + 1.6v_2 = 0$, has slope $v_2/v_1 = (\lambda_1 - 1)/1.6 = 2.31$, an angle of 66.6° from the x axis.

How much probability is inside an ellipse? The question reduces to one about the length of a standard normal vector. Whiten: $\boldsymbol{Z} = \mathbf{L}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$, with $\mathbf{L}$ the Cholesky factor of $\mathbf{C}$, is a vector of independent standard normals, and $Q = \boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$ (the calculation in step (3) of (2.78), read backwards). So the probability inside $Q \le c$ is the probability that a k-dimensional standard

normal vector has squared length at most c. In Section 2.13 we will show that this squared length has the χ^2_k distribution [Wasserman Thm 2.44(4)]. For $k = 2$ we can do it at once in polar coordinates, exactly as in (2.36):

$$\mathrm{P}(Q \le c) \overset{(1)}{=} \int_0^{2\pi} \int_0^{\sqrt{c}} \frac{e^{-r^2/2}}{2\pi}\, r\, \mathrm{d}r\, \mathrm{d}\phi \overset{(2)}{=} \left[-e^{-r^2/2}\right]_0^{\sqrt{c}} = 1 - e^{-c/2}. \tag{2.82}$$

(1) $Q = z_1^2 + z_2^2 = r^2$, and the density (2.75) with $k = 2$ is $e^{-r^2/2}/2\pi$.

(2) The angular integral gives 2π, and $r\, e^{-r^2/2}$ has antiderivative $-e^{-r^2/2}$.

Setting $1 - e^{-c/2}$ equal to the one-dimensional 1σ and 2σ probabilities, 0.6827 and 0.9545, gives $c = -2\ln(1-0.6827) = 2.296$ and $c = 6.180$. These are the familiar $\Delta\chi^2 = 2.30$ and 6.18 contours of two-parameter plots. So the ellipse $Q = 1$, which is often drawn as the "1σ ellipse", is too small: it contains only $1 - e^{-1/2} = 39.3\%$. In k dimensions the required c grows further. A random vector in many dimensions is almost never near its centre, because the χ^2_k distribution of its squared length has mean k.

Reading the ellipse: marginal box, conditional chord. One ellipse carries two different error bars for y, and they must not be confused. Figure 2.34 is drawn with the true numbers of the running example. *The box.* The ellipse $Q = 1$ is tangent to the box $|x - \mu_x| \le \sigma_x$, $|y - \mu_y| \le \sigma_y$. (Check the top point: with $x - \mu_x = z_1$, $y - \mu_y = 1.6z_1 + 1.2z_2$ and $z_1 = \cos t$, $z_2 = \sin t$, y is largest when $\tan t = 1.2/1.6$, giving $y - \mu_y = 1.6 \times 0.8 + 1.2 \times 0.6 = 2 = \sigma_y$.) So the projection of the $Q = 1$ ellipse on each axis is the marginal $\pm1\sigma$ range. *The chord.* The vertical chord through the centre is shorter: at $x = \mu_x$ the ellipse has $y - \mu_y = \pm 1.2 = \pm\sigma_y\sqrt{1-\rho^2}$, which, as Section 2.12.4 shows, is the 1σ range of y when x is *known*. *The tilt.* The points where the ellipse has vertical tangents, $(\mu_x \pm 1, \mu_y \pm 1.6)$, lie on the line $y - \mu_y = \rho(\sigma_y/\sigma_x)(x - \mu_x)$, the conditional mean. This line is *not* the major axis: its angle is $\arctan 1.6 = 58°$, not 66.6°.

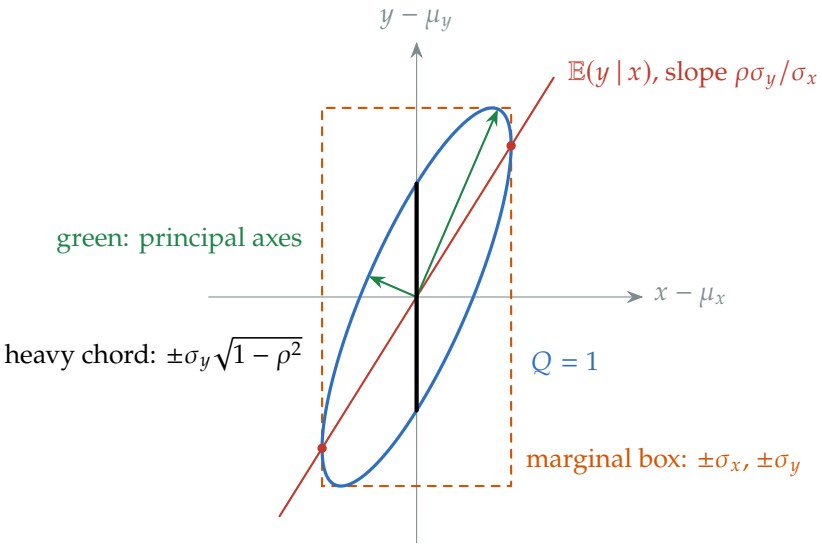


Figure 2.34: The $Q = 1$ contour of the example ($\sigma_x = 1$, $\sigma_y = 2$, $\rho = 0.8$), drawn exactly. It just touches the dashed box of the marginal $\pm1\sigma$ ranges. The heavy vertical chord at $x = \mu_x$ is the much shorter $\pm1\sigma$ range of y once x is fixed. The red line of conditional means passes through the two points of vertical tangency (dots), and it is tilted less steeply than the long principal axis (green).

2.12.3 Linear combinations and marginals

Every linear function of a Gaussian vector is again Gaussian, and marginals are a special case. The quickest proof uses the characteristic function, the k-dimensional version of (2.46), $\phi_X(\boldsymbol{t}) = \mathbb{E}\, e^{i\boldsymbol{t}^\top \boldsymbol{X}}$. With $\boldsymbol{X} = \boldsymbol{\mu} + \mathbf{A}\boldsymbol{Z}$,

$$\phi_X(\boldsymbol{t}) \overset{(1)}{=} e^{i\boldsymbol{t}^\top\boldsymbol{\mu}}\, \mathbb{E}\exp\!\big(i(\mathbf{A}^\top\boldsymbol{t})^\top \boldsymbol{Z}\big) \overset{(2)}{=} e^{i\boldsymbol{t}^\top\boldsymbol{\mu}} \prod_{j=1}^{k} e^{-(\mathbf{A}^\top\boldsymbol{t})_j^2/2} \overset{(3)}{=} \exp\!\Big(i\boldsymbol{t}^\top\boldsymbol{\mu} - \tfrac{1}{2}\boldsymbol{t}^\top\mathbf{C}\,\boldsymbol{t}\Big). \tag{2.83}$$

(1) $\boldsymbol{t}^\top\mathbf{A}\boldsymbol{Z} = (\mathbf{A}^\top\boldsymbol{t})^\top\boldsymbol{Z}$, and the constant phase comes out of the expectation.
(2) The Z_j are independent, so the expectation factorises, and each factor is the one-dimensional Gaussian characteristic function $e^{-s^2/2}$ (2.50) at $s = (\mathbf{A}^\top\boldsymbol{t})_j$.
(3) $\sum_j (\mathbf{A}^\top\boldsymbol{t})_j^2 = \boldsymbol{t}^\top\mathbf{A}\mathbf{A}^\top\boldsymbol{t} = \boldsymbol{t}^\top\mathbf{C}\boldsymbol{t}$.

Now let $\mathbf{B}$ be any $m \times k$ matrix and $\boldsymbol{Y} = \mathbf{B}\boldsymbol{X}$. Its characteristic function is $\mathbb{E}\, e^{i\boldsymbol{s}^\top\mathbf{B}\boldsymbol{X}} = \phi_X(\mathbf{B}^\top\boldsymbol{s}) = \exp(i\boldsymbol{s}^\top\mathbf{B}\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{s}^\top\mathbf{B}\mathbf{C}\mathbf{B}^\top\boldsymbol{s})$, which is again of the form (2.83). Since the characteristic function determines the distribution,

$$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C}) \quad\Longrightarrow\quad \mathbf{B}\boldsymbol{X} \sim \mathcal{N}\big(\mathbf{B}\boldsymbol{\mu},\ \mathbf{B}\mathbf{C}\mathbf{B}^\top\big). \tag{2.84}$$

Two special cases are the ones everybody uses [Wasserman Thm 2.44], [CasellaBerger p. 144].

- A single linear combination, $\mathbf{B} = \boldsymbol{a}^\top$: $\boldsymbol{a}^\top\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{a}^\top\boldsymbol{\mu}, \boldsymbol{a}^\top\mathbf{C}\boldsymbol{a})$. For two variables, $ax + by \sim \mathcal{N}(a\mu_x + b\mu_y,\ a^2\sigma_x^2 + b^2\sigma_y^2 + 2ab\rho\sigma_x\sigma_y)$. This is the exact form of the error-propagation formula for linear functions.
- The marginal of a subset. Split $\boldsymbol{X} = (\boldsymbol{X}_a, \boldsymbol{X}_b)$ and $\boldsymbol{\mu}$, $\mathbf{C}$ into the matching blocks $\boldsymbol{\mu}_a$, $\mathbf{C}_{aa}$, $\mathbf{C}_{ab}$, and so on. Choosing $\mathbf{B} = (\mathbb{1}\ \ 0)$ picks out $\boldsymbol{X}_a$, and $\mathbf{B}\mathbf{C}\mathbf{B}^\top = \mathbf{C}_{aa}$. So $\boldsymbol{X}_a \sim \mathcal{N}(\boldsymbol{\mu}_a, \mathbf{C}_{aa})$.

The second statement is marginalising in exactly the sense of Chapter 1. To average out the nuisance variables $\boldsymbol{X}_b$ of a Gaussian, *delete their rows and columns* from $\boldsymbol{\mu}$ and $\mathbf{C}$. No integral needs to be done. In the running example the marginal of y is $\mathcal{N}(\mu_y, 4)$ whatever the value of ρ; the script finds a sample standard deviation of 1.997 for y.

2.12.4 Conditionals: fixing one variable

Suppose we learn the value of $\boldsymbol{X}_a$ exactly. What is the distribution of $\boldsymbol{X}_b$ now? Split $\boldsymbol{X}_b$ into the part that $\boldsymbol{X}_a$ can predict and a residual that it cannot. Define the residual

$$\boldsymbol{W} = \boldsymbol{X}_b - \mathbf{C}_{ba}\mathbf{C}_{aa}^{-1}\boldsymbol{X}_a.$$

$\boldsymbol{W}$ and $\boldsymbol{X}_a$ are uncorrelated:

$$\operatorname{Cov}(\boldsymbol{W}, \boldsymbol{X}_a) \overset{(1)}{=} \operatorname{Cov}(\boldsymbol{X}_b, \boldsymbol{X}_a) - \mathbf{C}_{ba}\mathbf{C}_{aa}^{-1}\operatorname{Cov}(\boldsymbol{X}_a, \boldsymbol{X}_a) \overset{(2)}{=} \mathbf{C}_{ba} - \mathbf{C}_{ba}\mathbf{C}_{aa}^{-1}\mathbf{C}_{aa} = 0. \tag{2.85}$$

(1) Covariance is linear in each argument, and $\mathbf{C}_{ba}\mathbf{C}_{aa}^{-1}$ is a constant matrix.
(2) The blocks of $\mathbf{C}$ are exactly these covariances.

For Gaussians, uncorrelated means independent. The vector $(\boldsymbol{W}, \boldsymbol{X}_a)$ is a linear function of $\boldsymbol{X}$, so by (2.84) it is jointly Gaussian, and its covariance matrix is block diagonal by (2.85). A block-diagonal $\mathbf{C}$ has a block-diagonal inverse and a determinant equal to the product of the blocks' determinants, so the density (2.78) factorises into a function of $\boldsymbol{w}$ times a function of $\boldsymbol{x}_a$. That is independence.

So knowing $\boldsymbol{X}_a = \boldsymbol{x}_a$ tells us nothing about $\boldsymbol{W}$, and $\boldsymbol{X}_b = \boldsymbol{W} + \mathbf{C}_{ba}\mathbf{C}_{aa}^{-1}\boldsymbol{x}_a$ is a Gaussian vector shifted by a constant. Its mean and covariance are

$$\mathbb{E}(\boldsymbol{X}_b \mid \boldsymbol{x}_a) \overset{(1)}{=} \mathbb{E}\,\boldsymbol{W} + \mathbf{C}_{ba}\mathbf{C}_{aa}^{-1}\boldsymbol{x}_a \overset{(2)}{=} \boldsymbol{\mu}_b + \mathbf{C}_{ba}\mathbf{C}_{aa}^{-1}(\boldsymbol{x}_a - \boldsymbol{\mu}_a),$$
$$\mathrm{Cov}(\boldsymbol{X}_b \mid \boldsymbol{x}_a) \overset{(3)}{=} \mathrm{Cov}(\boldsymbol{W}) \overset{(4)}{=} \mathbf{C}_{bb} - 2\,\mathbf{C}_{ba}\mathbf{C}_{aa}^{-1}\mathbf{C}_{ab} + \mathbf{C}_{ba}\mathbf{C}_{aa}^{-1}\mathbf{C}_{aa}\mathbf{C}_{aa}^{-1}\mathbf{C}_{ab} = \mathbf{C}_{bb} - \mathbf{C}_{ba}\mathbf{C}_{aa}^{-1}\mathbf{C}_{ab}. \quad (2.86)$$

(1) $\boldsymbol{W}$ is independent of $\boldsymbol{X}_a$, so conditioning does not change its mean.
(2) $\mathbb{E}\,\boldsymbol{W} = \boldsymbol{\mu}_b - \mathbf{C}_{ba}\mathbf{C}_{aa}^{-1}\boldsymbol{\mu}_a$.
(3) A constant shift does not change a covariance, and $\boldsymbol{W}$ is unaffected by the conditioning.
(4) Expand $\mathrm{Cov}(\boldsymbol{X}_b - \mathbf{M}\boldsymbol{X}_a) = \mathbf{C}_{bb} - \mathbf{M}\mathbf{C}_{ab} - \mathbf{C}_{ba}\mathbf{M}^{\mathsf{T}} + \mathbf{M}\mathbf{C}_{aa}\mathbf{M}^{\mathsf{T}}$ with $\mathbf{M} = \mathbf{C}_{ba}\mathbf{C}_{aa}^{-1}$, using $\mathbf{C}_{ab}^{\mathsf{T}} = \mathbf{C}_{ba}$ and the symmetry of $\mathbf{C}_{aa}^{-1}$. The two middle terms are equal.

Key idea: Marginal and conditional of a Gaussian

For $\boldsymbol{X} = (\boldsymbol{X}_a, \boldsymbol{X}_b) \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$ [Wasserman Thm 2.44]:

- *marginal*: $\boldsymbol{X}_a \sim \mathcal{N}(\boldsymbol{\mu}_a, \mathbf{C}_{aa})$ (delete the other rows and columns);
- *conditional*: $\boldsymbol{X}_b \mid \boldsymbol{X}_a = \boldsymbol{x}_a \;\sim\; \mathcal{N}\big(\boldsymbol{\mu}_b + \mathbf{C}_{ba}\mathbf{C}_{aa}^{-1}(\boldsymbol{x}_a - \boldsymbol{\mu}_a),\; \mathbf{C}_{bb} - \mathbf{C}_{ba}\mathbf{C}_{aa}^{-1}\mathbf{C}_{ab}\big)$.

The conditional mean moves linearly with $\boldsymbol{x}_a$; the conditional covariance does not depend on $\boldsymbol{x}_a$ at all, and is never larger than the marginal one.

For two variables the conditional is [CasellaBerger p. 145]

$$y \mid x \;\sim\; \mathcal{N}\Big(\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x),\; \sigma_y^2(1 - \rho^2)\Big), \quad (2.87)$$

since here $\mathbf{C}_{ba}\mathbf{C}_{aa}^{-1} = \rho\sigma_x\sigma_y/\sigma_x^2$ and $\mathbf{C}_{bb} - \mathbf{C}_{ba}\mathbf{C}_{aa}^{-1}\mathbf{C}_{ab} = \sigma_y^2 - \rho^2\sigma_y^2$. As $|\rho| \to 1$ the conditional width goes to zero: y becomes a deterministic linear function of x.

Example: the slab. In the running example ($\mu_x = 1$, $\mu_y = 2$, $\sigma_x = 1$, $\sigma_y = 2$, $\rho = 0.8$), fix $x = \mu_x + 1 = 2$. Then (2.87) predicts $y \mid x \sim \mathcal{N}(2 + 0.8 \times 2 \times 1,\; 4 \times 0.36)$, a mean of 3.600 and a standard deviation of 1.200, compared with the marginal standard deviation of 2. The script cannot condition on an exact value, so it keeps the 9588 samples (out of 4×10^5) in the thin slab $|x - 2| < 0.05$ and finds mean 3.614 and standard deviation 1.187. Panel (b) of Figure 2.35 shows both histograms against the predicted Gaussians.

Example: conditional versus marginal errors. The conditional variance of one variable can be read straight off the inverse covariance: it is $1/(\mathbf{C}^{-1})_{bb}$, while the marginal variance is $\mathbf{C}_{bb}$. The reason is the block-inverse (Schur complement) formula, which says that the bb block of $\mathbf{C}^{-1}$ is $(\mathbf{C}_{bb} - \mathbf{C}_{ba}\mathbf{C}_{aa}^{-1}\mathbf{C}_{ab})^{-1}$, the inverse of the conditional covariance. In the running example $(\mathbf{C}^{-1})_{yy} = 1/1.44$, so the conditional variance is 1.44 and the conditional standard deviation 1.2, as above. The same distinction matters for Fisher forecasts (Chapter 10), where the Fisher matrix $\mathbf{F}$ plays the role of $\mathbf{C}^{-1}$. There $1/\sqrt{F_{ii}}$ is the error on parameter i with all the others *fixed* (the chord), and $\sqrt{(\mathbf{F}^{-1})_{ii}}$ is the error with the others *marginalised* (the box). Quoting the first when you mean the second is a classic way to overstate a forecast.

Watch out: Gaussian marginals do not make a Gaussian vector

Every property above starts from *joint* normality, and the derivation does not run backwards [CasellaBerger p. 145]. Let $X \sim \mathcal{N}(0, 1)$ and let $S = \pm 1$ be an independent fair sign. Then

$Y = SX$ is also $\mathcal{N}(0,1)$, and $\mathrm{Cov}(X,Y) = \mathbb{E}(S)\,\mathbb{E}(X^2) = 0$. But $X+Y$ equals 0 with probability $\frac{1}{2}$, which no Gaussian does, and $|Y| = |X|$ exactly, so the two are uncorrelated but completely dependent. "Uncorrelated implies independent" holds only for jointly Gaussian variables.

Python: sampling and checking. The script draws 20000 points by the Cholesky recipe, computes Q for each, and checks the ellipse contents and the conditional.

Listing 2.16: Drawing a bivariate Gaussian and measuring the content of its ellipses (`code/ch02/24_mvn.py`)

```
30 mu = np.array([1.0, 2.0])
31 sx, sy, rho = 1.0, 2.0, 0.8
32 C = np.array([[sx ** 2, rho * sx * sy], [rho * sx * sy, sy ** 2]])
33 L = np.linalg.cholesky(C)                    # C = L L^T
34 N = 20_000
35 x = mu + rng.standard_normal((N, 2)) @ L.T  # each row: mu + L z
36
37 Cinv = np.linalg.inv(C)
38 r = x - mu
39 chi2 = np.einsum("ni,ij,nj->n", r, Cinv, r)
40 lev1, lev2 = stats.chi2.ppf([stats.chi2.cdf(1, 1), stats.chi2.cdf(4, 1)], df=2)
41 f1, f2 = np.mean(chi2 < lev1), np.mean(chi2 < lev2)
```

Line 33 computes the Cholesky factor $\mathbf{L}$, with $\mathbf{L}\mathbf{L}^{\mathsf{T}} = \mathbf{C}$. Line 35 is the whole sampler: each row of the standard-normal array is a $\boldsymbol{z}$, and multiplying by $\mathbf{L}^{\mathsf{T}}$ from the right is the row-vector version of $\mathbf{L}\boldsymbol{z}$. Line 39 evaluates $Q = \boldsymbol{r}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{r}$ for every sample at once with `np.einsum`. Line 40 converts the one-dimensional 1σ and 2σ probabilities into χ^2_2 quantiles, the $\Delta\chi^2$ levels, and line 41 counts the samples inside: 0.6873 and 0.9559, against 0.6827 and 0.9545. Is that agreement? With 20000 samples the binomial standard error of a fraction near 0.68 is $\sqrt{0.68 \times 0.32/20000} \approx 0.003$, so both agree within about 1.5 standard errors.

The same object in three sciences. The multivariate Gaussian is the common core of the three kinds of data this book keeps returning to. *Colliders.* For Poisson counts in many bins it is the large-count limit of the product of Poissons. Its $\mathbf{C}$ is diagonal, with variances equal to the means, until systematic uncertainties shared between bins add correlations. *The CMB.* For a CMB map (Chapter 7) the vector is the list of pixel temperatures or of the harmonic coefficients $a_{\ell m}$, and $\mathbf{C}$ is built from the power spectrum C_ℓ. For an isotropic Gaussian field the $a_{\ell m}$ are independent with variance C_ℓ, so in harmonic space $\mathbf{C}$ is diagonal. *Gravitational waves.* For detector noise the vector is the time series and $\mathbf{C}$ is set by the power spectral density. Stationarity makes it a Toeplitz matrix (constant along each diagonal), and the Fourier transform diagonalises it, just as the eigenbasis did in (2.81). In each case the step that makes the likelihood computable is the same: whiten, and then everything is a sum of independent squares. That sum of squares is the χ^2 variable of Section 2.13. When the tails of real data are heavier than Gaussian, the multivariate Student-t, a Gaussian with a randomly rescaled covariance (Section 2.14), is the standard robust replacement [Lambert §8.4.11].

2.13 Sums of squared Gaussians: the χ^2 distribution

We fit a straight line to ten points with error bars and get $\chi^2 = \sum_i (x_i - \mu_i)^2/\sigma_i^2 = 14.2$, where x_i is a measured value, μ_i the line's prediction and σ_i the error bar. Is the fit good? The number means nothing until we know how χ^2 is distributed when the model is right. Sums of squared Gaussians turn up throughout physics. A CMB band power $\widehat{C}_\ell = \frac{1}{2\ell+1}\sum_m |a_{\ell m}|^2$ is a sum of

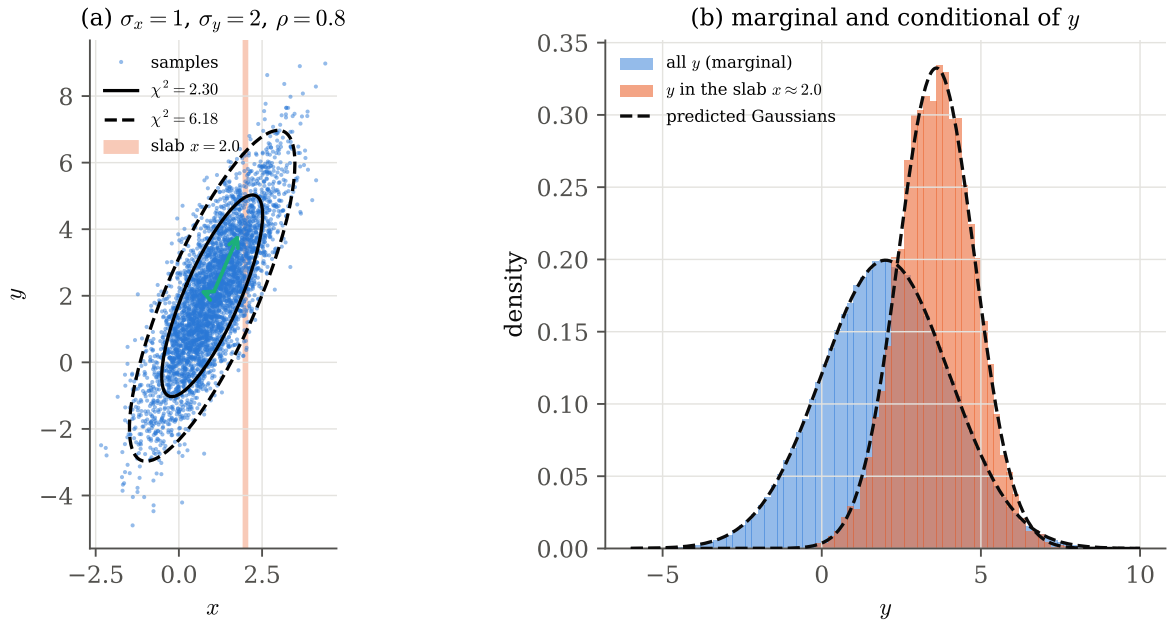


Figure 2.35: (a) Samples of the bivariate Gaussian with $\sigma_x = 1$, $\sigma_y = 2$, $\rho = 0.8$, with the $\Delta\chi^2 = 2.30$ and 6.18 ellipses and the principal axes (arrows, of length one standard deviation along each axis). The shaded strip is the slab used for conditioning. (b) All y values (the marginal) and the y values inside the slab (the conditional), each with its predicted Gaussian. Fixing x shifts the centre and shrinks the width.

squares of Gaussian coefficients, and its scatter is cosmic variance. The power of Gaussian detector noise in one frequency bin of a gravitational-wave stream is the sum of two squares, the real and imaginary parts. The quadratic form in the exponent of the multivariate Gaussian, whose contours we just drew, is a sum of squares after whitening. And the sample variance of n repeated measurements is a sum of squared deviations.

Two results answer these questions. A sum of ν squared standard normals has a definite distribution, χ^2_ν, with mean ν and variance 2ν. And the sample variance divides by $n-1$, not n, because estimating the mean uses up exactly one "degree of freedom" in a sense the geometry makes precise.

If ν independent standard normal numbers are squared and added, what is the density of the total? And what changes when the numbers are deviations from their own average rather than from the true mean?

2.13.1 One square

Start with a single square, $Q = Y^2$ with $Y \sim \mathcal{N}(0,1)$. *When is Q small?* For $q > 0$, $Q \le q$ happens exactly when $-\sqrt{q} \le Y \le \sqrt{q}$. So we can get the density of Q by differentiating the probability of that interval:

$$f_1(q) = \frac{\mathrm{d}}{\mathrm{d}q}\,\mathrm{P}(Q \le q) \overset{(1)}{=} \frac{\mathrm{d}}{\mathrm{d}q}\left[\Phi(\sqrt{q}) - \Phi(-\sqrt{q})\right] \overset{(2)}{=} 2\varphi(\sqrt{q})\,\frac{1}{2\sqrt{q}} \overset{(3)}{=} \frac{1}{\sqrt{2\pi q}}\,e^{-q/2}. \tag{2.88}$$

(1) The probability of an interval is the difference of the cumulative distribution at its ends.
(2) Chain rule, $\frac{\mathrm{d}}{\mathrm{d}q}\sqrt{q} = \frac{1}{2\sqrt{q}}$, and the symmetry $\varphi(-y) = \varphi(y)$. The two ends contribute equally: this is the factor two of [Cowan (10.14)], which counts the positive and the negative root.

(3) $\varphi(\sqrt{q}) = e^{-q/2}/\sqrt{2\pi}$.

The density diverges like $q^{-1/2}$ at the origin, because Y is most likely near zero and squaring crowds those values towards $q = 0$. One square is a gamma variable, Gamma$(\frac{1}{2}, 2)$ [CasellaBerger Lemma 5.3.2]. To check, recall the gamma density (2.20): with shape $\alpha = \frac{1}{2}$ and scale $\beta = 2$ it is $q^{-1/2}e^{-q/2}/(\Gamma(\frac{1}{2})\sqrt{2})$, and $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ (2.37) makes this identical to (2.88).

2.13.2 Many squares: two derivations

The geometric route: integrate the joint density over a ball. *The question in words:* what is the probability that $Q_\nu \le q$, whatever the individual values of $Y_1, \dots, Y_\nu$ happen to be? We know the joint density of all the Y_j, and we do not care about them one by one, only about the one combination Q_ν. So we do what we always do with variables we do not care about (Section 1.10.2): integrate the joint density over every set of values $y_1, \dots, y_\nu$ that gives $Q_\nu \le q$, and then change variables to Q. It goes in four steps.

Step 1: the joint density. The Y_j are independent, so their joint density is the product of the ν one-dimensional densities. Writing $\boldsymbol{y} = (y_1, \dots, y_\nu)$ and $|\boldsymbol{y}|^2 = y_1^2 + \cdots + y_\nu^2$,

$$f(\boldsymbol{y}) = \prod_{j=1}^{\nu} \frac{e^{-y_j^2/2}}{\sqrt{2\pi}} = \frac{e^{-(y_1^2+\cdots+y_\nu^2)/2}}{(2\pi)^{\nu/2}} = \frac{e^{-|\boldsymbol{y}|^2/2}}{(2\pi)^{\nu/2}}, \tag{2.89}$$

since a product of exponentials is the exponential of the sum. The density depends on $\boldsymbol{y}$ only through its length $r = |\boldsymbol{y}|$: it is the same at every point of a sphere centred on the origin.

Step 2: the event $Q_\nu \le q$ is a ball. $Q_\nu = |\boldsymbol{Y}|^2$ is the squared distance of the random point $\boldsymbol{Y}$ from the origin. So $Q_\nu \le q$ happens exactly when $\boldsymbol{Y}$ lands inside the ball of radius $\sqrt{q}$, and its probability is the joint density integrated over all $\boldsymbol{y}$ in that ball:

$$\mathrm{P}(Q_\nu \le q) = \int_{|\boldsymbol{y}|^2 \le q} \frac{e^{-|\boldsymbol{y}|^2/2}}{(2\pi)^{\nu/2}} \, \mathrm{d}^\nu y.$$

This is the one-square argument of Section 2.13.1 in ν dimensions: there the ball was the interval $[-\sqrt{q}, \sqrt{q}]$.

Step 3: spherical coordinates. Because the integrand depends only on r, we slice the ball into thin spherical shells. A shell of radius r and thickness $\mathrm{d}r$ has volume $S_{\nu-1} r^{\nu-1} \mathrm{d}r$, where $S_{\nu-1}$ is the area of the unit sphere in ν dimensions ($S_1 = 2\pi$, the circumference of the unit circle; $S_2 = 4\pi$). The power $r^{\nu-1}$ is there because a sphere of radius r is the unit sphere stretched by r in each of its $\nu - 1$ directions. Integrating over the angles of each shell leaves a one-dimensional integral,

$$\mathrm{P}(Q_\nu \le q) = \frac{S_{\nu-1}}{(2\pi)^{\nu/2}} \int_0^{\sqrt{q}} r^{\nu-1} e^{-r^2/2} \, \mathrm{d}r. \tag{2.90}$$

We still need $S_{\nu-1}$. The Gaussian integral gives it, by computing one integral in two ways. The integral of $e^{-|\boldsymbol{y}|^2/2}$ over all of $\mathbb{R}^\nu$ factorises into ν one-dimensional Gaussian integrals, each equal to $\sqrt{2\pi}$; in shells it is $S_{\nu-1} \int_0^\infty r^{\nu-1} e^{-r^2/2} \mathrm{d}r$. Setting the two equal,

$$(2\pi)^{\nu/2} \overset{(1)}{=} \int_{\mathbb{R}^\nu} e^{-|\boldsymbol{y}|^2/2} \, \mathrm{d}^\nu y \overset{(2)}{=} S_{\nu-1} \int_0^\infty r^{\nu-1} e^{-r^2/2} \, \mathrm{d}r \overset{(3)}{=} S_{\nu-1} \, 2^{\nu/2-1} \, \Gamma\left(\tfrac{\nu}{2}\right) \quad \Longrightarrow \quad S_{\nu-1} = \frac{2\pi^{\nu/2}}{\Gamma(\nu/2)}. \tag{2.91}$$

(1) Cartesian coordinates: $e^{-|\boldsymbol{y}|^2/2} = \prod_j e^{-y_j^2/2}$, and each $\int_{-\infty}^{\infty} e^{-y_j^2/2} \mathrm{d}y_j = \sqrt{2\pi}$.

(2) Spherical coordinates, as in step 3, with the ball extended to all of space.
(3) Substitute $t = r^2/2$, so $r = \sqrt{2t}$, $\mathrm{d}r = \mathrm{d}t/\sqrt{2t}$ and $r^{\nu-1}\mathrm{d}r = (2t)^{\nu/2-1}\mathrm{d}t$; the t integral is $\int_0^\infty t^{\nu/2-1}e^{-t}\mathrm{d}t = \Gamma(\nu/2)$.

Divide by $2^{\nu/2-1}\Gamma(\nu/2)$ and use $(2\pi)^{\nu/2}/2^{\nu/2-1} = 2\pi^{\nu/2}$. (Check: $\nu = 2$ gives $2\pi/\Gamma(1) = 2\pi$, and $\nu = 3$ gives $2\pi^{3/2}/(\sqrt{\pi}/2) = 4\pi$.)

Step 4: from the radius to Q, and differentiate. In (2.90) change the integration variable from r to $s = r^2$, the value of Q on the shell, and then differentiate with respect to the upper limit:

$$\mathrm{P}(Q_\nu \le q) \overset{(1)}{=} \frac{S_{\nu-1}}{(2\pi)^{\nu/2}} \int_0^q s^{(\nu-1)/2}\, e^{-s/2}\, \frac{\mathrm{d}s}{2\sqrt{s}} \overset{(2)}{=} \frac{1}{2^{\nu/2}\,\Gamma(\nu/2)} \int_0^q s^{\nu/2-1}\, e^{-s/2}\, \mathrm{d}s,$$

$$f_\nu(q) = \frac{\mathrm{d}}{\mathrm{d}q}\,\mathrm{P}(Q_\nu \le q) \overset{(3)}{=} \frac{q^{\nu/2-1}e^{-q/2}}{2^{\nu/2}\,\Gamma(\nu/2)}, \qquad q > 0, \tag{2.92}$$

(1) $r = \sqrt{s}$, so $r^{\nu-1} = s^{(\nu-1)/2}$ and $\mathrm{d}r = \mathrm{d}s/(2\sqrt{s})$; the limits $r = 0$ and $r = \sqrt{q}$ become $s = 0$ and $s = q$.
(2) Insert $S_{\nu-1}$ from (2.91): $\frac{2\pi^{\nu/2}}{\Gamma(\nu/2)} \cdot \frac{1}{(2\pi)^{\nu/2}} \cdot \frac{1}{2} = \frac{1}{2^{\nu/2}\Gamma(\nu/2)}$, and combine the powers, $s^{(\nu-1)/2}s^{-1/2} = s^{\nu/2-1}$.
(3) The derivative of an integral with respect to its upper limit is the integrand at that limit.

the χ^2 density with ν **degrees of freedom** [Cowan (2.34)], [Wasserman p. 30]. The power $q^{\nu/2-1}$ is pure geometry: the area of a sphere of radius $\sqrt{q}$ in ν dimensions. In high dimensions this factor pushes the probability away from the origin, which is why χ^2_ν peaks near $\nu - 2$ rather than at zero. We use the same shell argument for the Maxwell–Boltzmann speed distribution in Section 2.15.

The characteristic function route. The second route draws no sphere. Characteristic functions turn the sum of squares into a product [Cowan (10.12)–(10.15)]. First transform one square. With $a = 1 - 2ik$,

$$\phi_1(k) = \mathbb{E}\, e^{ikY^2} \overset{(1)}{=} \int_{-\infty}^{\infty} \frac{e^{-y^2/2}}{\sqrt{2\pi}}\, e^{iky^2}\, \mathrm{d}y \overset{(2)}{=} \int_{-\infty}^{\infty} \frac{e^{-ay^2/2}}{\sqrt{2\pi}}\, \mathrm{d}y \overset{(3)}{=} (1-2ik)^{-1/2}. \tag{2.93}$$

(1) The expectation of a function of Y is its integral against the density of Y.
(2) Combine the exponents: $-\frac{y^2}{2} + iky^2 = -\frac{(1-2ik)y^2}{2}$.
(3) The Gaussian integral $\int e^{-ay^2/2}\mathrm{d}y = \sqrt{2\pi/a}$, which holds for complex a with positive real part (here $\mathrm{Re}\, a = 1$) with the principal square root, by analytic continuation from real a.

For $Q_\nu = Y_1^2 + \cdots + Y_\nu^2$ with independent Y_j, the product rule (2.47) gives $\phi_\nu(k) = (1-2ik)^{-\nu/2}$ [Cowan (10.15)]. Which density has this transform? One square was a gamma variable, so try the gamma density with shape α and scale β:

$$\int_0^\infty e^{ikx}\, \frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^\alpha}\, \mathrm{d}x \overset{(1)}{=} \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty x^{\alpha-1}e^{-x(1-i\beta k)/\beta}\, \mathrm{d}x \overset{(2)}{=} \frac{1}{\Gamma(\alpha)\beta^\alpha}\,\Gamma(\alpha)\Big(\frac{\beta}{1-i\beta k}\Big)^\alpha = (1-i\beta k)^{-\alpha}, \tag{2.94}$$

(1) Combine the exponentials.
(2) Substitute $t = x(1-i\beta k)/\beta$; then $\int_0^\infty x^{\alpha-1}e^{-cx}\mathrm{d}x = \Gamma(\alpha)/c^\alpha$, again valid for complex c with positive real part.

so $(1-2ik)^{-\nu/2}$ is the characteristic function of Gamma$(\nu/2, 2)$. Since a characteristic function determines its density, $f_\nu(q) = q^{\nu/2-1}e^{-q/2}/(2^{\nu/2}\Gamma(\nu/2))$, which is (2.92) again: the product of ν transforms and the integral over a ball give the same density.

Working definition 2.18 (χ^2 distribution)

$Q \sim \chi^2_\nu$ if $Q = \sum_{j=1}^{\nu} Y_j^2$ with $Y_j \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Equivalently, Q has density (2.92), the Gamma$(\nu/2, 2)$ density, and characteristic function $(1-2ik)^{-\nu/2}$. *Summary numbers:* $\mathbb{E}\,Q = \nu$, $\operatorname{Var} Q = 2\nu$, mode $\max(\nu-2, 0)$. *Additivity:* independent $\chi^2_{\nu_1}$ and $\chi^2_{\nu_2}$ add to $\chi^2_{\nu_1+\nu_2}$ (multiply the characteristic functions). *Recognise it by:* a sum of squares of independent, zero-mean, unit-variance Gaussian quantities, such as residuals divided by their true errors.

Mean and variance. The mean of χ^2_ν is ν and its variance is 2ν. For one square, $\mathbb{E}\,Y^2 = 1$ and $\mathbb{E}\,Y^4 = 3$ (differentiate the Gaussian characteristic function four times, or integrate by parts twice as in (2.38)), so $\operatorname{Var} Y^2 = 3 - 1 = 2$. Means and variances of independent terms add, so $\mathbb{E}\,Q_\nu = \nu$ and $\operatorname{Var} Q_\nu = \sum_{i=1}^{\nu} \operatorname{Var} Y_i^2 = \sum_{i=1}^{\nu} 2 = 2\nu$ [Cowan (2.36)–(2.37)]. This explains why the variance of χ^2 increases with degrees of freedom as each random variable comes with its own fluctuation and they all add up. The script sums ν squared normals 2×10^5 times and finds mean 5.010 and variance 10.076 for $\nu = 5$, and 9.993 and 19.897 for $\nu = 10$ (Figure 2.36a). For large ν the CLT makes $\chi^2_\nu \approx \mathcal{N}(\nu, 2\nu)$, but slowly: its skewness is $\sqrt{8/\nu}$, still 0.28 at $\nu = 100$ (the skewness of one square, $\sqrt{8}$, shrinks like $1/\sqrt{\nu}$, as in (2.55)).

Example: is $\chi^2 = 14.2$ for ten degrees of freedom a good fit? Yes, it is acceptable. *The question in words:* if the straight line is the right model, how often would a fit come out this bad or worse? With $\nu = 10$, $\mathbb{E}\,\chi^2 = 10$ and the standard deviation is $\sqrt{20} = 4.5$. So 14.2 is less than one standard deviation high. Exactly, $\mathrm{P}(\chi^2_{10} > 14.2) = 0.16$: one fit in six would do this badly even with the right model. The central 90% range of χ^2_{10} is 3.9 to 18.3. A fit with $\chi^2 = 24$ would have tail probability 0.008 and would be suspicious.

The same reduced χ^2/ν can mean different things for different ν. Its standard deviation is $\sqrt{2/\nu}$, which is 0.45 for ten points. So "$\chi^2/\nu = 1.4$" is unremarkable for $\nu = 10$, while for $\nu = 100$ ($\chi^2 = 140$) it has tail probability 0.005. Hypothesis tests built on this are the subject of Chapter 4.

Example: cosmic variance. A CMB band power is a χ^2 variable, and that fixes its scatter. Recall that $\widehat{C}_\ell = \frac{1}{2\ell+1}\sum_m |a_{\ell m}|^2$ is the measured power at multipole ℓ, built from the $2\ell+1$ harmonic coefficients $a_{\ell m}$, and C_ℓ is its true value. *How many independent Gaussians go into it?* Exactly $2\ell+1$ real numbers. The $m = 0$ coefficient is real. For $m > 0$ the real and imaginary parts count separately. And $a_{\ell,-m}$ adds nothing new, being fixed by $a_{\ell,-m} = (-1)^m a^*_{\ell m}$. For an isotropic Gaussian sky these $2\ell+1$ numbers are independent Gaussians whose squares add up to $\sum_m |a_{\ell m}|^2$, each with the same share of the variance. So $(2\ell+1)\widehat{C}_\ell/C_\ell \sim \chi^2_{2\ell+1}$, and

$$\operatorname{Var}\widehat{C}_\ell \overset{(1)}{=} \left(\frac{C_\ell}{2\ell+1}\right)^2 \operatorname{Var}\chi^2_{2\ell+1} \overset{(2)}{=} \frac{C_\ell^2}{(2\ell+1)^2}\,2(2\ell+1) = \frac{2C_\ell^2}{2\ell+1},$$

where (1) scales a random variable by a constant and (2) uses $\operatorname{Var}\chi^2_\nu = 2\nu$. This is the cosmic variance that Chapter 7 verifies by simulation. At $\ell = 2$, $\widehat{C}_\ell$ is a χ^2_5 variable divided by five: skewed, with a 1σ scatter of 63%.

2.13.3 Correlated Gaussians: the quadratic form

Correlations do not change the answer, as long as the sum of squares uses the full inverse covariance. If $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$ in k dimensions, then

$$Q = (\boldsymbol{X} - \boldsymbol{\mu})^{\mathsf{T}} \mathbf{C}^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) \;\sim\; \chi^2_k \tag{2.95}$$

[Cowan (2.39)], [Wasserman Thm 2.44(4)]. The proof is one line with the whitening of Section 2.12: with $\mathbf{L}$ the Cholesky factor, $\mathbf{L}\mathbf{L}^{\mathsf{T}} = \mathbf{C}$, the vector $\boldsymbol{Z} = \mathbf{L}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$ has independent standard normal components and $\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z} = (\boldsymbol{X} - \boldsymbol{\mu})^{\mathsf{T}}(\mathbf{L}\mathbf{L}^{\mathsf{T}})^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) = Q$. This is what justified the ellipse contents in (2.82).

Example: forgetting the correlations. Dropping the off-diagonal elements keeps the right mean but gives the wrong spread. The script takes a correlated three-dimensional Gaussian with covariance

$$\mathbf{C} = \begin{pmatrix} 1 & 0.6 & 0.3 \\ 0.6 & 2 & -0.5 \\ 0.3 & -0.5 & 1.5 \end{pmatrix}$$

and computes Q for 2×10^5 draws: mean 2.999 and variance 5.989, as χ^2_3 requires. It also computes the "naive" sum $\sum_i (x_i - \mu_i)^2 / C_{ii}$, which ignores the off-diagonal elements. Its mean is still 2.997, because each term has expectation 1 whatever the correlations. But its variance is 7.284 instead of 6 (Figure 2.36b), because the terms are correlated and so do not add like independent squares. A goodness-of-fit test that uses only the diagonal errors on correlated data (neighbouring bins of a smoothed spectrum, overlapping band powers) will reject correct models too often, even though its average looks fine.

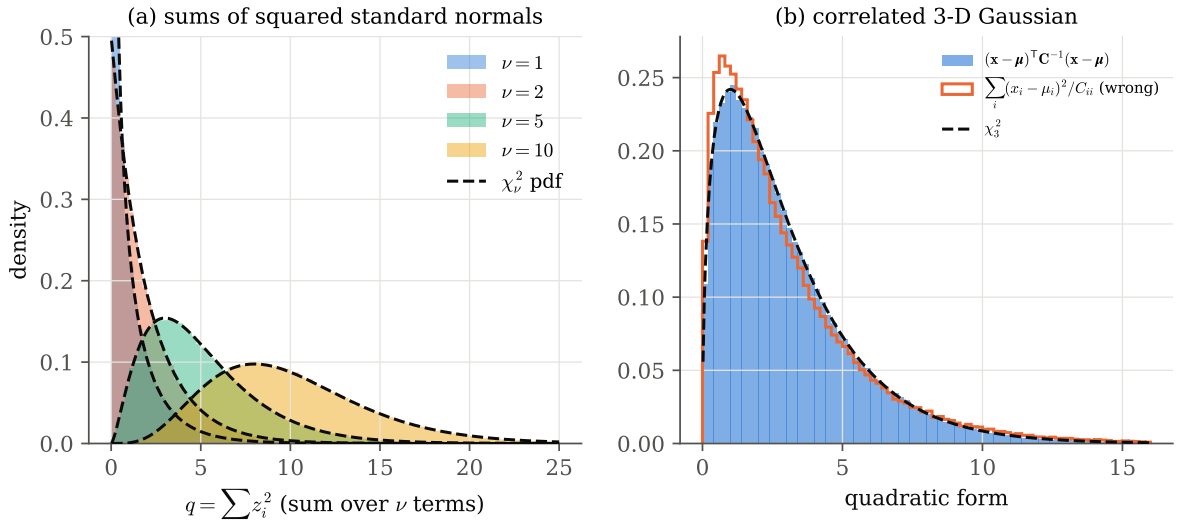


Figure 2.36: (a) Histograms of sums of ν squared standard normals for $\nu = 1, 2, 5, 10$ against the χ^2_ν densities (dashed). For $\nu = 1$ and 2 the density is largest at zero; for larger ν the peak moves out to $\nu - 2$. (b) For a correlated three-dimensional Gaussian, the full quadratic form follows χ^2_3, while the sum that uses only the diagonal variances (orange) has the right mean but the wrong shape: it piles up near zero and has a longer tail, so its variance is larger.

2.13.4 The sample variance and the lost degree of freedom

Five readings of a pendulum period, $x_1, \dots, x_5$, scatter about the unknown true value μ with unknown spread σ. We estimate μ by the sample mean $\bar{x} = \frac{1}{n}\sum x_i$, and σ^2 from the deviations

$x_i - \bar{x}$. Why divide the sum of squared deviations by $n - 1$?

The bias. Deviations from $\bar{x}$ are systematically smaller than deviations from the true μ. The reason: the sample mean is the value of a that minimises $\sum_i (x_i - a)^2$ (set the derivative $-2\sum_i (x_i - a)$ to zero), so no other centre, μ included, gives a smaller sum. By how much smaller? Write $x_i - \mu = (x_i - \bar{x}) + (\bar{x} - \mu)$ and expand:

$$\sum_{i=1}^{n} (x_i - \mu)^2 \overset{(1)}{=} \sum_{i=1}^{n} (x_i - \bar{x})^2 + 2(\bar{x} - \mu) \sum_{i=1}^{n} (x_i - \bar{x}) + n(\bar{x} - \mu)^2$$
$$\overset{(2)}{=} \sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2. \tag{2.96}$$

(1) Square the sum of two terms and sum over i; the last term does not depend on i.
(2) $\sum_i (x_i - \bar{x}) = n\bar{x} - n\bar{x} = 0$.

Take expectations for independent x_i with variance σ^2, any distribution. The left side has expectation $n\sigma^2$ and $\mathbb{E}(\bar{x} - \mu)^2 = \operatorname{Var} \bar{x} = \sigma^2/n$, so

$$\mathbb{E} \sum_{i=1}^{n} (x_i - \bar{x})^2 = n\sigma^2 - n \cdot \frac{\sigma^2}{n} = (n-1)\sigma^2. \tag{2.97}$$

Hence $S^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$ has $\mathbb{E}\, S^2 = \sigma^2$, while the "natural" $\frac{1}{n} \sum (x_i - \bar{x})^2$ is low on average by the factor $(n-1)/n$. The script, with $n = 5$ and $\sigma^2 = 4.0$, finds $\mathbb{E}\, S^2 = 4.0024$ and $\mathbb{E}\left[\frac{1}{n} \sum (x_i - \bar{x})^2\right] = 3.2019$, against the predicted 3.2000. The missing σ^2/n is exactly the variance of $\bar{x}$ around μ: the part of the scatter that the sample mean "absorbs" by following the data.

The distribution: rotate the data. For Gaussian data we can say much more: $(n-1)S^2/\sigma^2$ is exactly χ^2_{n-1}, and it is independent of $\bar{x}$. A rotation of the data shows both. Picture the n readings as one point $\boldsymbol{x} \in \mathbb{R}^n$. The sample mean is about the component of $\boldsymbol{x}$ along the diagonal direction $\boldsymbol{e} = (1, \ldots, 1)/\sqrt{n}$, since $\boldsymbol{e}^{\mathsf{T}} \boldsymbol{x} = \sqrt{n}\, \bar{x}$. The deviations $x_i - \bar{x}$ are the components of what is left over, $\boldsymbol{x} - \sqrt{n}\, \bar{x}\, \boldsymbol{e}$, which is perpendicular to $\boldsymbol{e}$ and so lives in an $(n-1)$-dimensional subspace (Figure 2.37). Choose an orthonormal basis whose first vector is $\boldsymbol{e}$, collect the basis vectors as the rows of an orthogonal matrix $\mathbf{H}$ (for $n = 3$, one choice, the **Helmert** matrix, is shown below), and rotate: $\boldsymbol{y} = \mathbf{H}(\boldsymbol{x} - \mu \mathbf{1})$, where $\mathbf{1} = (1, \ldots, 1)$. Then:

- $\boldsymbol{x} - \mu\mathbf{1} \sim \mathcal{N}(0, \sigma^2 \mathbb{1})$, so by (2.84) $\boldsymbol{y} \sim \mathcal{N}(0, \sigma^2 \mathbf{H}\mathbf{H}^{\mathsf{T}}) = \mathcal{N}(0, \sigma^2 \mathbb{1})$: the y_j are again independent with variance σ^2. An isotropic Gaussian looks the same in every rotated frame.
- $y_1 = \boldsymbol{e}^{\mathsf{T}}(\boldsymbol{x} - \mu\mathbf{1}) = \sqrt{n}(\bar{x} - \mu)$.
- A rotation preserves length, so $\sum_{j=1}^{n} y_j^2 = \sum_i (x_i - \mu)^2$, and by (2.96), $\sum_{j=2}^{n} y_j^2 = \sum_i (x_i - \mu)^2 - n(\bar{x} - \mu)^2 = \sum_i (x_i - \bar{x})^2$.

The last line says that $(n-1)S^2/\sigma^2 = \sum_{j=2}^{n} (y_j/\sigma)^2$ is a sum of $n-1$ independent squared standard normals, and it involves only $y_2, \ldots, y_n$, which are independent of y_1 and hence of $\bar{x}$.

Example: $n = 3$, by hand. The rows of

$$\mathbf{H} = \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} \end{pmatrix}$$

have unit length and are mutually orthogonal (for example, row 2 · row 3 $= \frac{1}{\sqrt{12}} - \frac{1}{\sqrt{12}} + 0 = 0$). Take the readings $\boldsymbol{x} = (1, 2, 6)$, so $\bar{x} = 3$ and $\sum(x_i - \bar{x})^2 = 4 + 1 + 9 = 14$. The second and third rotated components (the shift by $\mu\mathbf{1}$ drops out of them, because their rows sum to zero) are $y_2 = (1-2)/\sqrt{2}$ and $y_3 = (1+2-12)/\sqrt{6}$, and $y_2^2 + y_3^2 = \frac{1}{2} + \frac{81}{6} = 14$. The first component carries the mean, the other two carry the scatter.

Key idea: Sample mean and sample variance of Gaussian data

If $x_1, \ldots, x_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ [CasellaBerger Thm 5.3.1]: (a) $\bar{x}$ and S^2 are independent; (b) $\bar{x} \sim \mathcal{N}(\mu, \sigma^2/n)$; (c) $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$. One degree of freedom is spent on the direction along which $\bar{x}$ is measured; the scatter lives in the remaining $n - 1$.

Casella and Berger prove (a) by factorising the joint density after the change of variables $(\bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x})$, and (c) by induction, adding one data point at a time with the update $(n-1)S_n^2 = (n-2)S_{n-1}^2 + \frac{n-1}{n}(x_n - \bar{x}_{n-1})^2$ [CasellaBerger (5.3.1)–(5.3.2)]. They also give the fastest check of (a): $\mathrm{Cov}(\bar{x}, x_j - \bar{x}) = \sum_i \frac{1}{n}(\delta_{ij} - \frac{1}{n})\sigma^2 = 0$, and uncorrelated linear functions of independent Gaussians are independent [CasellaBerger Lemma 5.3.3]. The rotation is the same argument seen geometrically.

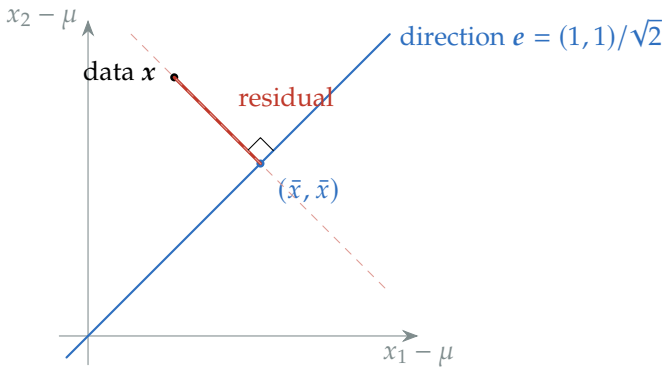


Figure 2.37: Two readings seen as one point $\boldsymbol{x}$ in the plane (origin at the true value). Projecting onto the diagonal gives the sample mean $(\bar{x}, \bar{x})$; the residual, whose length squared is $\sum(x_i - \bar{x})^2$, is perpendicular to the diagonal. All data sets with the same mean lie on the dashed line, so the residual has only $n - 1 = 1$ direction to move in. For an isotropic Gaussian the component along the diagonal and the component across it are independent.

Python: checking the theorem, and seeing it fail. The script draws 200000 samples of $n = 5$ Gaussian readings.

Listing 2.17: The sample variance: $n - 1$ degrees of freedom, and independence from the mean (`code/ch02/25_chi2_sampvar.py`)

```
n, sigma, mu = 5, 2.0, 10.0
xs = rng.normal(mu, sigma, (NREP, n))
xbar = xs.mean(axis=1)
S2 = xs.var(axis=1, ddof=1)             # 1/(n-1) estimator
S2n = xs.var(axis=1, ddof=0)            # 1/n estimator
w = (n - 1) * S2 / sigma ** 2

# known-mean version: sum (x_i - mu)^2 / sigma^2 has n dof
w_known = ((xs - mu) ** 2).sum(axis=1) / sigma ** 2

# independence of xbar and S^2: Gaussian vs exponential parent
xe = rng.exponential(1.0, (NREP, n))
```

```
79 corr_gauss = np.corrcoef(xbar, S2)[0, 1]
80 corr_exp = np.corrcoef(xe.mean(axis=1), xe.var(axis=1, ddof=1))[0, 1]
```

Line 70 computes S^2 with `ddof=1`, numpy's name for dividing by $n-1$ ("delta degrees of freedom"). Line 71 divides by n. Line 72 forms $(n-1)S^2/\sigma^2$, whose simulated mean and variance are 4.002 and 8.022, as χ^2_4 requires (4 and 8). Line 75 forms the same sum with the true mean instead of $\bar{x}$, which follows χ^2_5 instead (Figure 2.38a). Lines 78–80 test independence, and show where it fails. The correlation between $\bar{x}$ and S^2 is +0.0016 for Gaussian readings, but +0.685 for exponential readings. *Why?* For a skewed parent, a sample that happens to contain a large value has both a large mean and a large variance. The independence of $\bar{x}$ and S^2 is special to the Gaussian (it actually characterises it), and so is everything built on it in the next section.

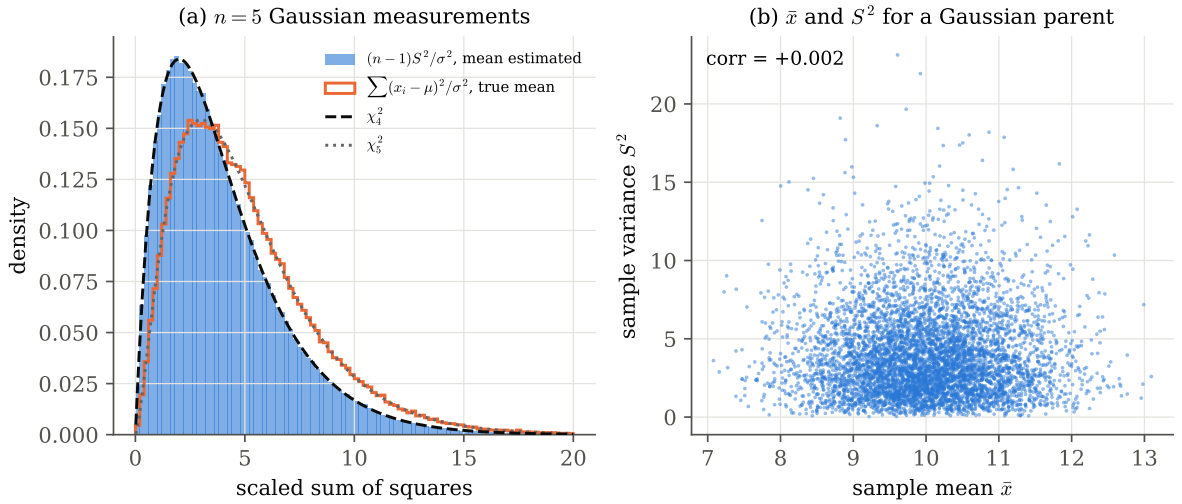


Figure 2.38: (a) For five Gaussian readings, the scaled sum of squared deviations from the sample mean (histogram) follows χ^2_4 (dashed), while deviations from the true mean (outline) follow χ^2_5 (dotted). (b) Sample mean against sample variance for Gaussian readings: no trend at all, as independence requires.

Watch out: Degrees of freedom are dimensions, not data points

The rule "ν = number of data minus number of fitted parameters" is the rotation argument again: each parameter that enters the model *linearly* and is fitted by least squares removes one direction from the residual vector, exactly as $\bar{x}$ did. The rule is exact for linear models with Gaussian errors, approximate for nonlinear ones, and wrong when the data are correlated but treated as independent, or when parameters sit at a physical boundary.

2.14 When the error bar is itself estimated: Student's t and the F ratio

We time five swings of a pendulum, compute g from each, and quote the result as $\bar{x} \pm S/\sqrt{5}$, with S the sample standard deviation. Then, to claim a "two-sigma" discrepancy with the textbook value, we check whether $|\bar{x} - g_0|$ exceeds $1.96\,S/\sqrt{5}$. This is wrong, and not by a little. The rule 1.96 comes from the Gaussian, and it assumes that we know the true σ. We do not. Our error bar is itself a random number, computed from the same five readings, and sometimes it comes out much too small. When it does, the ratio $(\bar{x}-\mu)/(S/\sqrt{n})$ is large, with μ the true value. So the ratio has fatter tails than a Gaussian. W. S. Gosset, who faced exactly this with small samples at the Guinness brewery and published as "Student", derived its distribution

[CasellaBerger p. 181].

Gosset's distribution does three jobs. It is the exact distribution of the *studentised* mean for Gaussian data, so small-sample intervals and tests built on it are honest. It is also a Gaussian whose variance is itself random, which makes it the standard robust replacement for a Gaussian likelihood. And the same construction applied to two variances gives the F distribution, the tool for asking whether two detectors, two runs or two models have the same scatter.

If a Gaussian quantity is divided by an independent, estimated measure of its own spread, how is the ratio distributed, and how different is it from a Gaussian?

2.14.1 The studentised mean

Why this ratio appears. Let $x_1, \dots, x_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. If we knew σ, the standardised mean

$$U = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

would be all we need. We do not know σ, so we replace it by the sample standard deviation S, with $S^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$. Write $p = n-1$. The quantity we actually compute is the **studentised mean**, and dividing its top and bottom by $\sigma/\sqrt{n}$ shows what it is made of [CasellaBerger (5.3.5)]:

$$T = \frac{\bar{x} - \mu}{S/\sqrt{n}} \overset{(1)}{=} \frac{(\bar{x} - \mu)/(\sigma/\sqrt{n})}{\sqrt{S^2/\sigma^2}} \overset{(2)}{=} \frac{U}{\sqrt{V/p}}, \qquad V = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_p. \tag{2.98}$$

(1) Divide numerator and denominator by $\sigma/\sqrt{n}$. The $\sqrt{n}$ cancels in the denominator.
(2) $S^2/\sigma^2 = V/p$ by the definition of V. That V is χ^2_p, and that U and V are independent, are the results on the sample mean and variance of Section 2.13.4.

The unknown σ has cancelled. So the whole problem is: given two independent pieces, $U \sim \mathcal{N}(0,1)$ and $V \sim \chi^2_p$, what is the distribution of $T = U/\sqrt{V/p}$?

The two ingredients and their joint density. The density of U is the standard Gaussian, and the density of V is the χ^2_p density (2.92):

$$f_U(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}, \qquad f_V(v) = \frac{1}{2^{p/2}\Gamma(p/2)} v^{p/2-1} e^{-v/2} \quad (v > 0). \tag{2.99}$$

Because U and V are independent, their joint density is the product, $f_{U,V}(u,v) = f_U(u)\, f_V(v)$.

Change variables from (U,V) to (T,V). We keep V and trade U for T. Solving (2.98) for u gives the map and its Jacobian:

$$u = t\sqrt{\frac{v}{p}}, \quad v = v, \qquad \left|\frac{\partial(u,v)}{\partial(t,v)}\right| \overset{(1)}{=} \begin{vmatrix} \sqrt{v/p} & \dfrac{t}{2\sqrt{pv}} \\ 0 & 1 \end{vmatrix} \overset{(2)}{=} \sqrt{\frac{v}{p}}. \tag{2.100}$$

(1) The four partial derivatives: $\partial u/\partial t = \sqrt{v/p}$, $\partial u/\partial v = t/(2\sqrt{pv})$, $\partial v/\partial t = 0$, $\partial v/\partial v = 1$.
(2) The matrix is triangular, so its determinant is the product of the diagonal entries. The off-diagonal entry never matters.

The joint density of (T,V) is the old joint density, evaluated at the old variables, times this

Jacobian (the rule of Eq. (1.38)):

$$\begin{aligned} f_{T,V}(t,v) &\overset{(1)}{=} f_U\Big(t\sqrt{\tfrac{v}{p}}\Big)\, f_V(v)\,\sqrt{\frac{v}{p}} \\ &\overset{(2)}{=} \frac{1}{\sqrt{2\pi}} e^{-t^2 v/(2p)}\, \frac{v^{p/2-1}e^{-v/2}}{2^{p/2}\Gamma(p/2)}\,\sqrt{\frac{v}{p}} \\ &\overset{(3)}{=} \frac{1}{\sqrt{2\pi p}\;2^{p/2}\Gamma(p/2)}\, v^{(p+1)/2-1}\,\exp\Big[-\frac{v}{2}\Big(1+\frac{t^2}{p}\Big)\Big]. \end{aligned} \tag{2.101}$$

(1) Change of variables, with the joint density of (U,V) factorised by independence.
(2) Substitute the two densities (2.99), with $u^2 = t^2 v/p$.
(3) Collect the powers of v: $v^{p/2-1}\cdot v^{1/2} = v^{(p+1)/2-1}$. Collect the exponentials: $e^{-t^2v/(2p)}\,e^{-v/2} = e^{-\frac{v}{2}(1+t^2/p)}$. The constants are $1/\sqrt{2\pi}$, $1/\sqrt{p}$ and $1/(2^{p/2}\Gamma(p/2))$.

The question that finishes the job. What is the probability density for $T = t$, regardless of what V happens to be? That is a marginal density: we have the joint density of (T,V), and V is a nuisance we do not care about, so we integrate it out (Section 1.10.2). Write $A = 1 + t^2/p$ to keep the formula short. Then

$$\begin{aligned} f_T(t) &\overset{(1)}{=} \int_0^\infty f_{T,V}(t,v)\,\mathrm{d}v \\ &\overset{(2)}{=} \frac{1}{\sqrt{2\pi p}\;2^{p/2}\Gamma(p/2)} \int_0^\infty v^{(p+1)/2-1}\, e^{-Av/2}\,\mathrm{d}v \\ &\overset{(3)}{=} \frac{1}{\sqrt{2\pi p}\;2^{p/2}\Gamma(p/2)}\,\Gamma\Big(\frac{p+1}{2}\Big)\Big(\frac{2}{A}\Big)^{(p+1)/2} \\ &\overset{(4)}{=} \frac{\Gamma\big(\frac{p+1}{2}\big)}{\Gamma\big(\frac{p}{2}\big)\sqrt{p\pi}}\,\Big(1+\frac{t^2}{p}\Big)^{-(p+1)/2}. \end{aligned} \tag{2.102}$$

(1) Marginalisation: the density of T alone is the joint density summed (integrated) over every value V can take.
(2) Insert (2.101). Everything that does not depend on v comes outside the integral.
(3) The gamma integral $\int_0^\infty x^{a-1}e^{-bx}\,\mathrm{d}x = \Gamma(a)/b^a$, with $a = \frac{p+1}{2}$ and $b = \frac{A}{2}$, so $1/b^a = (2/A)^{(p+1)/2}$.
(4) The powers of two cancel, $2^{(p+1)/2}/(2^{p/2}\sqrt{2}) = 1$, leaving $1/\sqrt{\pi p}$; then put back $A = 1+t^2/p$.

This is Student's t density with p degrees of freedom [CasellaBerger (5.3.6)], [Wasserman p. 30]. The whole construction fits in one line: a standard Gaussian U, divided by the square root of an independent χ^2_p/p, is t_p. The t distribution exists because the fixed σ of a Gaussian standardisation has been replaced by a random estimate, whose square is a scaled χ^2, and that extra randomness fattens the tails.

Working definition 2.19 (Student's t distribution)

$T \sim t_p$ if $T = U/\sqrt{V/p}$ with $U \sim \mathcal{N}(0,1)$ and $V \sim \chi^2_p$ independent; its density is (2.102). For Gaussian data, $(\bar{x}-\mu)/(S/\sqrt{n}) \sim t_{n-1}$ exactly, for every $n \geq 2$. *Equivalent picture:* a Gaussian $\mathcal{N}(0,\sigma^2)$ whose variance is random, $\sigma^2 \sim$ Inv-Gamma$(p/2, p/2)$ [so the precision $1/\sigma^2$ is a gamma variable with mean one] [Lambert §8.4.7]. *Summary numbers:* $\mathbb{E}\,T = 0$ for $p > 1$, $\mathrm{Var}\,T = p/(p-2)$ for $p > 2$; only moments of order less than p exist.

The tails, and the moments that do not exist. For large $|t|$, (2.102) falls like $|t|^{-(p+1)}$, a power law. So $\mathbb{E}\,|T|^r$ is finite only for $r < p$ [CasellaBerger (5.3.7)]. With $p = 1$ the density is $\Gamma(1)/(\Gamma(\tfrac{1}{2})\sqrt{\pi})\,(1+t^2)^{-1} = \frac{1}{\pi(1+t^2)}$, the Cauchy distribution of Section 2.10.5. So the studentised mean of *two* Gaussian measurements has no mean at all. The variance, for $p > 2$, follows from independence:

$$\operatorname{Var} T = \mathbb{E}\,T^2 \overset{(1)}{=} p\;\mathbb{E}(U^2)\;\mathbb{E}\Big(\frac{1}{V}\Big) \overset{(2)}{=} p\int_0^\infty \frac{1}{v}\,\frac{v^{p/2-1}e^{-v/2}}{2^{p/2}\Gamma(p/2)}\,\mathrm{d}v \overset{(3)}{=} p\,\frac{2^{p/2-1}\Gamma(p/2-1)}{2^{p/2}\Gamma(p/2)} \overset{(4)}{=} \frac{p}{p-2}. \tag{2.103}$$

(1) $T^2 = pU^2/V$ with U and V independent, so the expectation factorises; $\mathbb{E}\,T = 0$ by symmetry.
(2) $\mathbb{E}\,U^2 = 1$, and the expectation of $1/V$ is its integral against the χ^2_p density.
(3) The integrand is $v^{(p/2-1)-1}e^{-v/2}$, which integrates to $\Gamma(p/2-1)\,2^{p/2-1}$ provided $p > 2$ (otherwise it diverges at $v = 0$).
(4) $\Gamma(p/2) = (p/2-1)\Gamma(p/2-1)$, so the ratio is $\frac{1}{2(p/2-1)} = \frac{1}{p-2}$.

The variance of T is dominated by the rare samples whose estimated error bar S happens to be tiny. That is what the divergence of step (3) at small v means. As $p \to \infty$, $V/p \to 1$ by the law of large numbers and $(1+t^2/p)^{-(p+1)/2} \to e^{-t^2/2}$, so $t_\infty = \mathcal{N}(0,1)$ [Wasserman p. 30]. For non-Gaussian data and large n the studentised mean still tends to $\mathcal{N}(0,1)$ (the studentised CLT, [Wasserman Thm 5.10]). What the t distribution adds is the *exact* small-n answer for Gaussian data.

Example: five pendulum readings. Back to the five pendulum values of g from the opening: how often does the Gaussian rule cry "two sigma" when nothing is wrong? The script simulates 400000 experiments of $n = 5$ readings with $\sigma = 0.05\,\mathrm{m\,s^{-2}}$ and computes T for each.

Listing 2.18: The studentised mean of five Gaussian readings (`code/ch02/26_student_t_F.py`)

```
35 n, mu, sigma = 5, 9.81, 0.05                  # five pendulum readings of g, noise 0.05
36 x = rng.normal(mu, sigma, (NREP, n))
37 xbar = x.mean(axis=1)
38 S = x.std(axis=1, ddof=1)
39 T = (xbar - mu) / (S / np.sqrt(n))             # sigma replaced by its estimate
40 Z = (xbar - mu) / (sigma / np.sqrt(n))         # sigma known: exactly N(0,1)
41
42 frac_T_196 = np.mean(np.abs(T) > 1.96)         # "2-sigma" by the Gaussian table
43 frac_Z_196 = np.mean(np.abs(Z) > 1.96)
44 t975 = stats.t.ppf(0.975, n - 1)              # the correct 95% two-sided point
45 frac_T_t975 = np.mean(np.abs(T) > t975)
46 frac_T_3 = np.mean(np.abs(T) > 3.0)
```

Line 38 estimates the error bar from the same five numbers (`ddof=1` is the $n-1$ divisor). Line 39 forms T with the estimated error bar and line 40 forms the same ratio with the true σ, which is exactly $\mathcal{N}(0,1)$. Lines 42–46 count how often each exceeds various thresholds. *With the true* σ, $|Z| > 1.96$ happens in a fraction 0.0497 of experiments, the promised 5%. *With the estimated error bar*, $|T| > 1.96$ happens in 0.1209, against the t_4 prediction 0.1216. So the nominal "2σ discrepancy" occurs more than twice as often as claimed. The correct 95% point is $t_{4,0.975} = 2.776$, and $|T|$ exceeds it in a fraction 0.0493. Further out the difference explodes: $|T| > 3$ happens in 0.0396 of experiments (t_4: 0.0399), fifteen times the Gaussian 0.0027 (Figure 2.39a).

Watch out: Do not use Gaussian thresholds with an estimated error bar

With n readings and σ estimated from them, the two-sided 95% multiplier is $t_{n-1,0.975}$: 12.7 for $n = 2$, 4.30 for $n = 3$, 2.78 for $n = 5$, 2.26 for $n = 10$, 2.09 for $n = 20$, and it reaches 1.96 only as $n \to \infty$. For three-sigma or five-sigma claims the Gaussian rule is off by orders of magnitude unless n is large. And the exact t needs Gaussian readings: with a few outliers even t is too optimistic.

Example: a Gaussian with a random width. Integrating V out of (2.101) has a second reading. For a fixed value $V = v$, $T = U\sqrt{p/v}$ is just a Gaussian with variance p/v; integrating over v averages these Gaussians of different widths. So the t distribution is a Gaussian whose width is itself random, and that is also a recipe for modelling data. Lambert's schools example [Lambert §8.4.7] merges the test scores of many schools that share one mean but have different spreads. Each school on its own is Gaussian; the merged population is not. If the school variances are distributed as $\sigma^2 \sim \text{Inv-Gamma}(\nu/2, \nu/2)$, equivalently the precisions $1/\sigma^2$ as $\text{Gamma}(\nu/2, \text{scale } 2/\nu)$, which is χ^2_ν/ν, the merged distribution is exactly t_ν: "its head rises, its shoulders fall, and its tails fatten." The script draws 400000 values both ways for $\nu = 3$: as a Gaussian with a random precision (lines 51–52), and as $U/\sqrt{V/\nu}$ (line 54). Both histograms follow t_3, with Kolmogorov–Smirnov distances [the largest gap between the empirical and the exact cumulative distribution] of 0.0010 and 0.0013, as small as sampling noise allows at this sample size (Figure 2.39b).

So the t distribution is the workhorse *robust likelihood*. When a few measurements have underestimated errors, when a detector occasionally misbehaves, when the population is heterogeneous, a t_ν likelihood with ν of a few lets outliers exist without dragging the fit. The degrees of freedom need not be an integer: ν is a free, real-valued shape parameter that can itself be fitted [Lambert §8.4.7]. Its multivariate version, a multivariate Gaussian with one random scale factor for the whole covariance, gives joint tails heavy enough to see Lambert's hedge-fund "bankruptcy days" that a multivariate normal fit misses [Lambert §8.4.11].

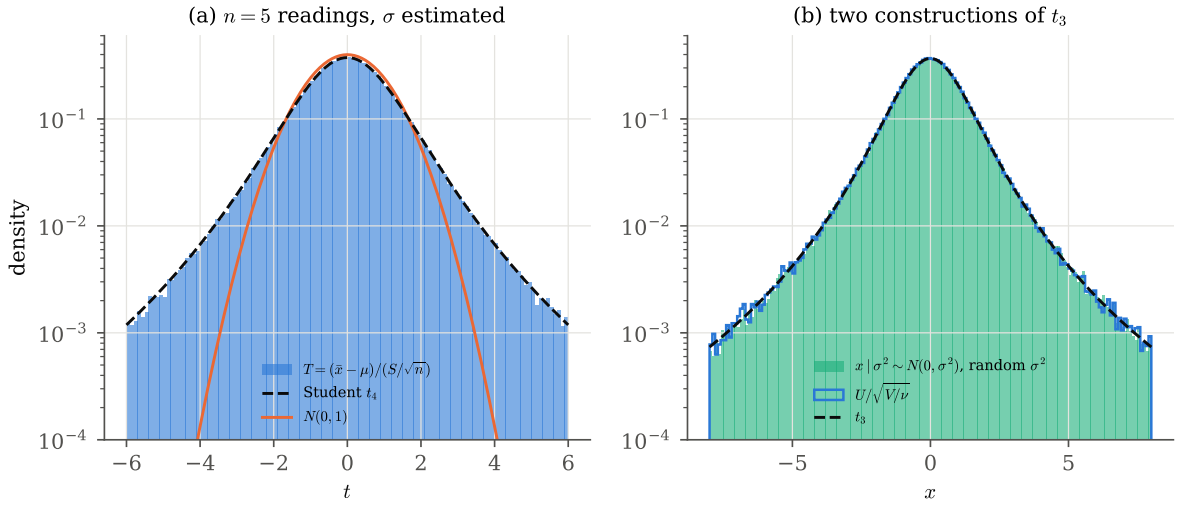


Figure 2.39: (a) The studentised mean of five Gaussian readings (histogram, logarithmic scale) follows Student's t_4 (dashed). The standard normal (solid) is far too thin in the tails. (b) Two constructions of t_3: a Gaussian whose variance is drawn afresh from an inverse-gamma distribution for every value, and the ratio $U/\sqrt{V/\nu}$. They are the same distribution.

2.14.2 Ratios of variances: Snedecor's F

Two calorimeters measure the same electron beam. Calorimeter X gives n_X readings with sample variance S_X^2, calorimeter Y gives n_Y readings with S_Y^2. Is X's resolution really worse, or did the numbers fluctuate? For Gaussian readings each $(n-1)S^2/\sigma^2$ is χ^2_{n-1}, and the two are independent, so [CasellaBerger (5.3.8)]

$$F = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} = \frac{V_1/p}{V_2/q}, \qquad V_1 \sim \chi^2_p, \quad V_2 \sim \chi^2_q, \qquad p = n_X - 1, \quad q = n_Y - 1. \tag{2.104}$$

Under the hypothesis $\sigma_X = \sigma_Y$ the unknown widths cancel and $F = S_X^2/S_Y^2$ is computable from the data. Its density comes from the same trick as for t: hold the denominator fixed, then average over it. Fix $V_2 = w$. Then $F = \frac{q}{pw} V_1$, so $V_1 = \frac{pw}{q} x$ and $f_{F\,|\,w}(x) = \frac{pw}{q} f_p\big(\frac{pwx}{q}\big)$. Averaging over w:

$$\begin{aligned} f_F(x) &\overset{(1)}{=} \int_0^\infty f_q(w)\, \frac{pw}{q}\, f_p\Big(\frac{pwx}{q}\Big)\mathrm{d}w \\ &\overset{(2)}{=} \frac{(p/q)^{p/2}\, x^{p/2-1}}{2^{(p+q)/2}\Gamma(\frac{p}{2})\Gamma(\frac{q}{2})} \int_0^\infty w^{(p+q)/2-1}\, e^{-\frac{w}{2}\left(1+\frac{px}{q}\right)}\mathrm{d}w \\ &\overset{(3)}{=} \frac{\Gamma\big(\frac{p+q}{2}\big)}{\Gamma\big(\frac{p}{2}\big)\Gamma\big(\frac{q}{2}\big)} \Big(\frac{p}{q}\Big)^{p/2} \frac{x^{p/2-1}}{\big(1+\frac{p}{q}x\big)^{(p+q)/2}}, \qquad x > 0. \end{aligned} \tag{2.105}$$

(1) Law of total probability, with the change-of-variable factor $\frac{\mathrm{d}V_1}{\mathrm{d}x} = \frac{pw}{q}$.

(2) Insert both χ^2 densities (2.92). The powers of w are $1 + (\frac{p}{2} - 1) + (\frac{q}{2} - 1) = \frac{p+q}{2} - 1$; the x-dependence outside the integral is $x^{p/2-1}$ with the constant $(p/q)^{1+p/2-1}$.

(3) The gamma integral with $a = \frac{p+q}{2}$ and $c = \frac{1}{2}(1 + px/q)$; the factor $2^{(p+q)/2}$ cancels.

This is Snedecor's $F_{p,q}$ density [CasellaBerger (5.3.9)]. Its mean is slightly above one. By the same factorisation as (2.103), $\mathbb{E}\,F = \frac{q}{p}\,\mathbb{E}\,V_1\,\mathbb{E}\,\frac{1}{V_2} = \frac{q}{q-2}$ for $q > 2$ [CasellaBerger Ex. 5.3.7]. It exceeds one because $1/V_2$ is convex, so its average exceeds one over the average of V_2.

Three facts connect F to its neighbours [CasellaBerger Thm 5.3.8], and each is visible from the definition (2.104). First, $1/F \sim F_{q,p}$: swap numerator and denominator. Second, if $T \sim t_q$ then $T^2 = U^2/(V/q) = (\chi^2_1/1)/(\chi^2_q/q) \sim F_{1,q}$. Third, $\frac{(p/q)F}{1+(p/q)F} = \frac{V_1}{V_1+V_2} \sim \mathrm{Beta}(p/2, q/2)$, the fraction of a total carried by one gamma variable, which we derive in Section 2.15. Casella and Berger also note Kelker's result: the ratio has an F distribution for any spherically symmetric parent, not only the Gaussian. So the F test is somewhat robust.

Example: two calorimeters. Small samples determine variances poorly. With six and eleven readings, one calorimeter's variance must come out more than three times the other's before we can claim, at the 5% level, that its resolution is worse. The script shows this. It takes $n_X = 6$ readings with $\sigma_X = 1.5$ and $n_Y = 11$ readings with $\sigma_Y = 0.4$ (the true widths differ, but they cancel in (2.104)), so $p = 5$ and $q = 10$. Over 400000 repetitions the mean of F is 1.252, against $q/(q-2) = 1.250$. The 95th percentile of $F_{5,10}$ is 3.326, and it is exceeded in a fraction 0.0502 of repetitions (Figure 2.40a). The script also squares the t_4 values of the pendulum experiment and compares them with $F_{1,4}$: the Kolmogorov–Smirnov distance on 10^5 values is 0.0035 (Figure 2.40b).

Where the ratios come back. Every exact small-sample test in classical statistics is built on one pattern: a Gaussian or a χ^2, divided by an independent χ^2 that estimates the unknown scale.

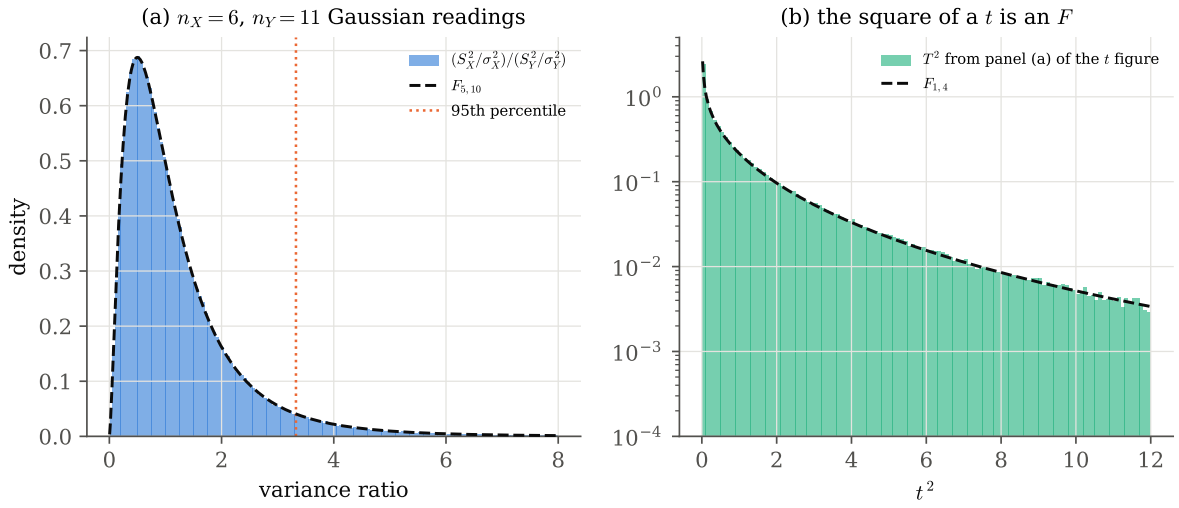


Figure 2.40: (a) The ratio of two scaled sample variances, from six and eleven Gaussian readings, follows $F_{5,10}$ (dashed); the dotted line marks its 95th percentile. (b) The squares of the studentised means of the pendulum experiment follow $F_{1,4}$, on a logarithmic scale that shows the power-law tail.

Examples are the t test for a mean, the two-sample t test, the F test for the equality of variances, and the F test that compares a model with a nested model having extra parameters by the ratio of χ^2 improvements. These are the subject of Chapter 4. Bayesian inference (Chapter 5) arrives at the same t_{n-1}. Integrating an unknown Gaussian σ out of the posterior for μ, with the prior $p(\sigma) \propto 1/\sigma$, produces it [Sivia p. 56–57, (3.43)–(3.44)]. The fattening of the tails is the price of not knowing the noise level, whichever school of inference pays it.

2.15 The rest of the zoo, each with the machine that makes it

Almost every distribution a physicist meets is produced by a small machine built from the ones we already have: a uniform phase, a sum, a product, a length, a ratio, a sorting. So a table of distributions is not a list of unrelated formulas. Recognising the machine lets us choose the right distribution for a new problem, and it tells us when a distribution will fail. Each remaining common continuous distribution is treated the same way: the machine, the density derived from it, the summary numbers, and a check by running the machine in Python.

The question for every entry, in words: *which physical mechanism produces this distribution, and what does the mechanism predict for its shape, its mean and its spread?*

2.15.1 Uniform: complete ignorance within a range

A charged particle crosses a silicon strip detector whose strips have pitch $w = 50\,\mu\text{m}$, and only the strip that fired is recorded. Where in the strip did the particle pass? With no further information every position is equally likely. The same holds for a phase that has been scrambled, the rounding error of a digitiser, or the position of an event within a time bin. All are **uniform**: $X \sim \mathrm{U}(a, b)$ has density $1/(b-a)$ on $[a, b]$ [CasellaBerger p. 81]. It is the maximum-entropy assignment when only the range is known (Section 2.11.5). Its moments are

$$\mathbb{E}\,X \overset{(1)}{=} \int_a^b \frac{x}{b-a}\,\mathrm{d}x = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2},$$

$$\operatorname{Var} X \overset{(2)}{=} \int_a^b \frac{(x-\frac{a+b}{2})^2}{b-a}\,\mathrm{d}x \overset{(3)}{=} \frac{1}{b-a}\int_{-(b-a)/2}^{(b-a)/2} u^2\,\mathrm{d}u = \frac{(b-a)^2}{12}. \tag{2.106}$$

(1) Definition of the mean, and $b^2-a^2=(b-a)(b+a)$.
(2) Definition of the variance about the mean just found.
(3) Substitute $u=x-\frac{a+b}{2}$; then $\int_{-h}^{h} u^2\mathrm{d}u = \frac{2h^3}{3}$ with $h=\frac{b-a}{2}$, which gives $\frac{(b-a)^3}{12}$.

So the strip detector has a position resolution of $w/\sqrt{12} = 14.4\,\mu\mathrm{m}$, the famous "pitch over root twelve", and a digitiser with step Δ adds quantisation noise of variance $\Delta^2/12$.

The uniform distribution is also the parent of all others: any continuous variable can be made from a uniform one, and turned back into one. If X is continuous with cumulative distribution function F, then $U=F(X)$ is uniform on $[0,1]$, because

$$\mathrm{P}(U\le u) \overset{(1)}{=} \mathrm{P}\big(X\le F^{-1}(u)\big) \overset{(2)}{=} F\big(F^{-1}(u)\big) = u,$$

where (1) uses that F is increasing and (2) is the definition of F. Read backwards, $X=F^{-1}(U)$ turns uniform random numbers into draws from any distribution: this is inverse-transform sampling, the first tool of Chapter 6. Read forwards, it says that a p-value, which is $1-F$ of the observed statistic under H_0, is uniformly distributed when H_0 is true (Chapter 4).

2.15.2 Exponential, gamma and Weibull: waiting and adding

The exponential and gamma distributions were derived in Section 2.6 from the Poisson process with rate λ: the wait for the first event is $\mathrm{Exp}(\lambda)$, and the wait for the n-th event is $\mathrm{Gamma}(n,1/\lambda)$, the Erlang density (2.19). The characteristic function (2.94) shows the whole family at once: gamma variables with a common scale add. $\mathrm{Gamma}(\alpha,\beta)$, with shape α and scale β, has $\phi(k)=(1-i\beta k)^{-\alpha}$, so for independent gamma variables with a common scale

$$\mathrm{Gamma}(\alpha_1,\beta)+\mathrm{Gamma}(\alpha_2,\beta)=\mathrm{Gamma}(\alpha_1+\alpha_2,\beta), \tag{2.107}$$

since $(1-i\beta k)^{-\alpha_1}(1-i\beta k)^{-\alpha_2}=(1-i\beta k)^{-(\alpha_1+\alpha_2)}$ [CasellaBerger Ex. 4.6.8]. Two special lines of the family we have met: $\mathrm{Exp}(\lambda)=\mathrm{Gamma}(1,1/\lambda)$, and $\chi^2_\nu=\mathrm{Gamma}(\nu/2,2)$ [CasellaBerger (3.3.10)–(3.3.11)]. So adding exponential waits and adding squared Gaussians are the same operation seen through different parameters.

The moments follow from a trick called "recognise the kernel" [CasellaBerger (3.3.7)–(3.3.8)]: multiplying by x turns one gamma density into another, whose integral we already know. Since the gamma density integrates to one, $\int_0^\infty x^{\alpha-1}e^{-x/\beta}\mathrm{d}x=\Gamma(\alpha)\beta^\alpha$ for every $\alpha>0$. Then

$$\mathbb{E}\,X \overset{(1)}{=} \frac{1}{\Gamma(\alpha)\beta^\alpha}\int_0^\infty x^{(\alpha+1)-1}e^{-x/\beta}\,\mathrm{d}x \overset{(2)}{=} \frac{\Gamma(\alpha+1)\beta^{\alpha+1}}{\Gamma(\alpha)\beta^\alpha} \overset{(3)}{=} \alpha\beta, \tag{2.108}$$

(1) Multiply the density by x: the integrand becomes the kernel of a $\mathrm{Gamma}(\alpha+1,\beta)$ density.
(2) Its integral is the normalisation of that density.
(3) $\Gamma(\alpha+1)=\alpha\Gamma(\alpha)$.

and in the same way $\mathbb{E}\,X^2=\alpha(\alpha+1)\beta^2$, so $\operatorname{Var}X=\alpha\beta^2$. For χ^2_ν this gives back ν and 2ν.

A third member of the waiting-time family is the **Weibull** distribution, $Y=X^{1/\gamma}$ with X exponential [CasellaBerger (3.3.12)]. It is the lifetime of a component whose failure rate $h(t)$ is not constant but grows (or falls) like a power of its age, $h(t)\propto t^{\gamma-1}$. So $\gamma>1$ describes wear-out, $\gamma<1$ infant mortality, and $\gamma=1$ the memoryless exponential. The Weibull is also one of the three limiting laws for the *maximum* of many variables, listed in the taxonomy of Section 2.10.6.

As a prior, the gamma distribution is the natural choice for a Poisson rate $\lambda>0$ [Lambert §8.5.3]. The reason is conjugacy, which Machine 3 of the beta distribution below shows at work.

2.15.3 Log-normal: products of many factors

A photo-electron entering a photomultiplier hits the first dynode and knocks out g_1 secondary electrons. Each of these knocks g_2 out of the next dynode, and so on through ten stages. The gain is a *product*, $G = g_1 g_2 \cdots g_{10}$. A product becomes a sum when we take logarithms: $\ln G = \sum_i \ln g_i$ is a sum of independent terms, so the central limit theorem makes $\ln G$ approximately Gaussian [Cowan p. 34–35]. A quantity whose logarithm is Gaussian is **log-normal**. Whenever an effect acts by multiplication (gains in cascades, fragmentation, growth, the stretching of fluid elements, the luminosity of a galaxy built by many mergers) the log-normal is the default, just as the Gaussian is for additive effects.

If $\ln X = Y \sim \mathcal{N}(\mu, \sigma^2)$, then $X = e^Y$ and $\mathrm{d}y/\mathrm{d}x = 1/x$, so

$$f(x) \overset{(1)}{=} f_Y(\ln x)\left|\frac{\mathrm{d}y}{\mathrm{d}x}\right| \overset{(2)}{=} \frac{1}{\sqrt{2\pi\sigma^2}}\,\frac{1}{x}\,\exp\!\left[-\frac{(\ln x-\mu)^2}{2\sigma^2}\right], \qquad x>0, \tag{2.109}$$

(1) The one-dimensional change of variables.
(2) Insert the Gaussian density for $Y = \ln x$ [Cowan (2.31)], [CasellaBerger (3.3.21)].

The parameters μ and σ^2 are the mean and variance of $\ln X$, not of X. The moments of X come from the Gaussian moment generating function $\mathbb{E}\,e^{rY}$:

$$\begin{aligned}\mathbb{E}\,X^r = \mathbb{E}\,e^{rY} &\overset{(1)}{=} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}}\exp\!\left[ry-\frac{(y-\mu)^2}{2\sigma^2}\right]\mathrm{d}y \\ &\overset{(2)}{=} e^{r\mu+r^2\sigma^2/2}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\!\left[-\frac{(y-\mu-r\sigma^2)^2}{2\sigma^2}\right]\mathrm{d}y \overset{(3)}{=} e^{r\mu+r^2\sigma^2/2}.\end{aligned} \tag{2.110}$$

(1) $X^r = e^{rY}$, integrated against the density of Y.
(2) Complete the square: $ry-\frac{(y-\mu)^2}{2\sigma^2} = -\frac{(y-\mu-r\sigma^2)^2}{2\sigma^2}+r\mu+\frac{r^2\sigma^2}{2}$, as expanding both sides confirms.
(3) The remaining integral is a normalised Gaussian centred at $\mu+r\sigma^2$.

So $\mathbb{E}\,X = e^{\mu+\sigma^2/2}$ and $\operatorname{Var} X = \mathbb{E}\,X^2 - (\mathbb{E}\,X)^2 = e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$ [Cowan (2.32)–(2.33)]. The median is e^{μ}, because ln is monotonic and the median of Y is μ. So mean > median > mode = $e^{\mu-\sigma^2}$: a long tail to the right.

Example: the gain of a ten-stage photomultiplier. The script gives each of 10 dynodes an independent uniform multiplication factor $g_i \sim \mathrm{U}(3,5)$, draws 4×10^5 gains, and compares with the log-normal predicted by the CLT applied to $\ln G$.

Listing 2.19: The gain of a ten-dynode photomultiplier (code/ch02/27_zoo_lognormal_maxwell.py)

```
n_dyn = 10
g = rng.uniform(3.0, 5.0, (NREP, n_dyn))
G = g.prod(axis=1)
lnG = np.log(G)
# CLT prediction for ln G: sum of n_dyn iid ln g_i
# E ln g and Var ln g for g ~ U(3,5), computed by quadrature-free closed form
a, b = 3.0, 5.0
E_ln = (b * np.log(b) - b - (a * np.log(a) - a)) / (b - a)
E_ln2 = (b * (np.log(b) ** 2 - 2 * np.log(b) + 2) - a * (np.log(a) ** 2 - 2 * np.log(a) +
    2)) / (b - a)
mu_ln = n_dyn * E_ln
s_ln = np.sqrt(n_dyn * (E_ln2 - E_ln ** 2))
mean_G_th_ln = np.exp(mu_ln + s_ln ** 2 / 2)       # log-normal mean
mean_G_exact = 4.0 ** n_dyn                          # E prod g_i = (E g)^n exactly
median_G = np.median(G)
```

Lines 36–38 run the machine: multiply ten factors, take the log. Lines 42–45 compute the CLT prediction exactly. For a uniform g, $\mathbb{E}\ln g$ and $\mathbb{E}(\ln g)^2$ are elementary integrals ($\int \ln g\,\mathrm{d}g = g\ln g - g$ and $\int (\ln g)^2 \mathrm{d}g = g[(\ln g)^2 - 2\ln g + 2]$). The mean and variance of $\ln G$ are ten times those of $\ln g$. The prediction is $\mu = 13.757$, $\sigma = 0.4633$; the simulated $\ln G$ has mean 13.758 and standard deviation 0.4629.

The gain itself is strongly skewed: its skewness is 1.333, while $\ln G$ has -0.054. Its median is 0.948×10^6 against $e^\mu = 0.943 \times 10^6$, and its mean is 1.050×10^6, matching the log-normal $e^{\mu+\sigma^2/2} = 1.050 \times 10^6$ and the exact $\mathbb{E}\,G = (\mathbb{E}\,g)^{10} = 4^{10} = 1.049 \times 10^6$ (Figure 2.41).

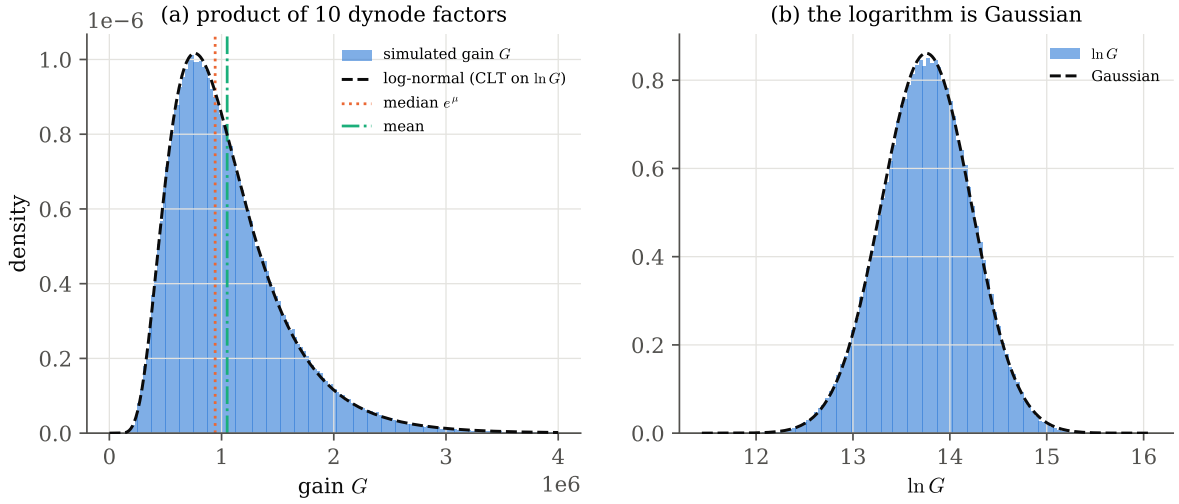


Figure 2.41: (a) The simulated gain of a ten-stage photomultiplier (histogram) and the log-normal predicted by applying the central limit theorem to $\ln G$ (dashed). The median e^μ (dotted) lies well below the mean (dash-dotted). (b) The same gains on a logarithmic axis: a symmetric Gaussian.

Watch out: Averaging logarithms gives the median, not the mean

For a log-normal quantity, $\exp(\overline{\ln x})$ estimates the median e^μ, which is smaller than the mean by the factor $e^{\sigma^2/2}$ (11% for the photomultiplier). Fitting a straight line to $\ln x$ and exponentiating the prediction has the same bias. Decide which of the two you need, and correct by $e^{\sigma^2/2}$ if it is the mean.

2.15.4 Rayleigh and Maxwell–Boltzmann: the length of a Gaussian vector

The velocity components (v_x, v_y, v_z) of an argon atom in a gas at temperature T are independent Gaussians with mean zero and variance $s^2 = k_B T/m$ (this is the MaxEnt result of Section 2.11.5, or the Boltzmann factor $e^{-mv^2/2k_BT}$). The *speed* $v = |\boldsymbol{v}|$ is the length of a three-dimensional Gaussian vector. We derived its distribution in all but name in the geometric route to χ^2: the density of $\boldsymbol{v}$ is $(2\pi s^2)^{-3/2} e^{-v^2/2s^2}$, and the volume of the shell at radius v is $S_2 v^2 \mathrm{d}v = 4\pi v^2 \mathrm{d}v$. Hence

$$f(v) \overset{(1)}{=} 4\pi v^2 \frac{e^{-v^2/2s^2}}{(2\pi s^2)^{3/2}} \overset{(2)}{=} \sqrt{\frac{2}{\pi}}\,\frac{v^2}{s^3}\,e^{-v^2/2s^2}, \qquad v > 0, \tag{2.111}$$

(1) The shell argument of (2.91) with $\nu = 3$.

(2) $4\pi/(2\pi)^{3/2} = \sqrt{2/\pi}$.

This is the Maxwell–Boltzmann speed distribution. Equivalently, $v^2/s^2 \sim \chi^2_3$. Its three landmark speeds, the most probable, the rms and the mean, are

$$v_{\text{mode}} \overset{(1)}{=} \sqrt{2}\,s = \sqrt{\frac{2k_BT}{m}}, \qquad v_{\text{rms}} \overset{(2)}{=} \sqrt{3}\,s,$$
$$\mathbb{E}\,v \overset{(3)}{=} \sqrt{\frac{2}{\pi}}\,\frac{1}{s^3}\int_0^\infty v^3 e^{-v^2/2s^2}\,\mathrm{d}v \overset{(4)}{=} \sqrt{\frac{8}{\pi}}\,s = \sqrt{\frac{8k_BT}{\pi m}}. \tag{2.112}$$

(1) $\frac{\mathrm{d}}{\mathrm{d}v}\left(v^2e^{-v^2/2s^2}\right) = \left(2v - v^3/s^2\right)e^{-v^2/2s^2} = 0$ at $v^2 = 2s^2$.
(2) $\mathbb{E}\,v^2 = s^2\;\mathbb{E}\,\chi^2_3 = 3s^2$: equipartition, $\frac{1}{2}m\,\mathbb{E}\,v^2 = \frac{3}{2}k_BT$.
(3) Definition of the mean.
(4) With $t = v^2/2s^2$, $\int_0^\infty v^3e^{-v^2/2s^2}\mathrm{d}v = 2s^4\int_0^\infty te^{-t}\mathrm{d}t = 2s^4$, and $\sqrt{2/\pi}\cdot 2s = \sqrt{8/\pi}\,s$.

For argon at 300 K the script finds $s = 249.9\,\mathrm{m/s}$, and from 4×10^5 simulated atoms a mean speed of 398.9 m/s (theory 398.7), with the mode at 353.4 and the rms speed at 432.8 (theory 432.8) (Figure 2.42b).

In two dimensions the shell has circumference $2\pi r$, and the same argument gives the **Rayleigh** distribution

$$f(r) = \frac{r}{s^2}\,e^{-r^2/2s^2}, \qquad \mathbb{E}\,r = s\sqrt{\pi/2}, \qquad r^2 \sim s^2\chi^2_2, \tag{2.113}$$

and the squared length r^2 is *exponential* with mean $2s^2$, since $\chi^2_2 = \text{Gamma}(1,2)$. Of all the cases in this book, this one reaches furthest. A complex Gaussian mode $z = x + iy$, with independent real and imaginary parts of equal variance, has a Rayleigh amplitude $|z|$, a uniform phase, and an exponential power $|z|^2$. That describes a Fourier mode $\tilde\delta(\boldsymbol{k})$ of a Gaussian random field (Chapter 7), each $a_{\ell m}$ with $m \neq 0$ on the CMB sky, and each frequency bin of Gaussian detector noise in a gravitational-wave periodogram.

What follows for the scatter of a measured power? The exponential has standard deviation equal to its mean. So a raw periodogram fluctuates by 100% in every bin, however long the data stream. Only averaging reduces the scatter: over bins or segments (Welch's method), or over the $2\ell+1$ values of m, which gives the cosmic variance of Section 2.13.2. The script checks the unit-variance case: $\mathbb{E}\,|z| = 1.2531$ against $\sqrt{\pi/2} = 1.2533$, and $|z|^2$ has mean 1.9986 and standard deviation 2.0012, both 2 for an exponential (Figure 2.42a).

2.15.5 Beta: probabilities of probabilities

The distributions so far describe data. The beta distribution usually describes a *probability*: the efficiency of a trigger, the bias of a coin, the fraction of events that are signal. It lives on $[0,1]$ and has density [CasellaBerger (3.3.16)], [Kruschke (6.3)]

$$f(\theta\,|\,a,b) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}, \qquad B(a,b) = \int_0^1 \theta^{a-1}(1-\theta)^{b-1}\mathrm{d}\theta = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}. \tag{2.114}$$

It is produced by three different machines, and each explains something.

Machine 1: sorting uniform numbers. Draw n independent uniform numbers on $[0,1]$ and let $U_{(k)}$ be the k-th smallest, the k-th **order statistic**. For $U_{(k)}$ to lie in $[u, u+\mathrm{d}u]$, one of the n numbers must fall in that interval, $k-1$ of them below u and $n-k$ above it. The number of ways to choose which is which is the multinomial coefficient $\frac{n!}{(k-1)!\,1!\,(n-k)!}$ of Section 2.3, and the

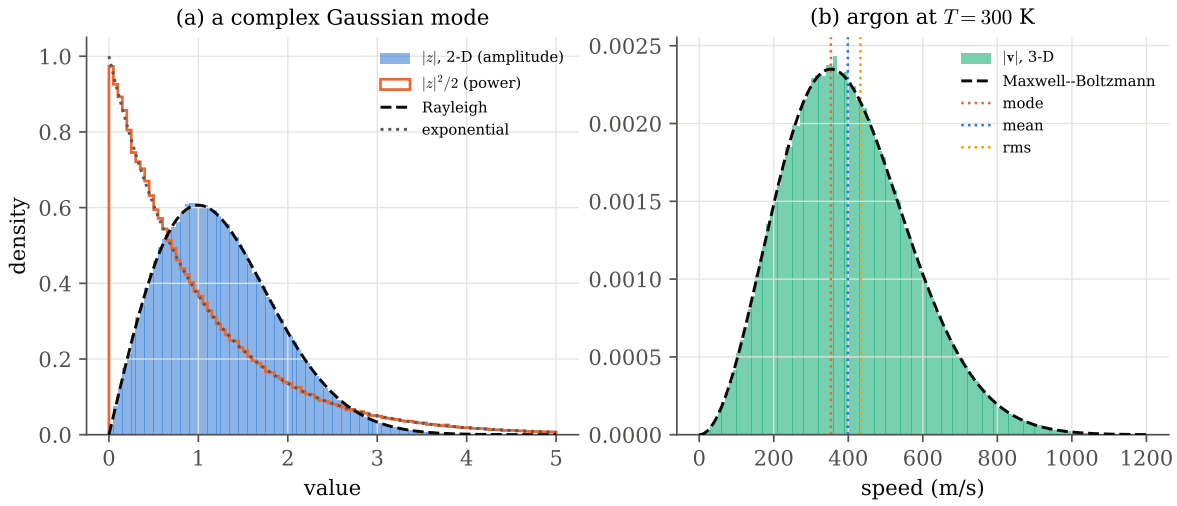


Figure 2.42: (a) The amplitude of a complex Gaussian mode follows the Rayleigh density, and half its squared amplitude, the power, follows a unit exponential. (b) Speeds of simulated argon atoms at 300 K against the Maxwell–Boltzmann law, with the mode, the mean and the rms speed marked in that order.

probabilities of the three cells are u, $\mathrm{d}u$ and $1-u$. So

$$f_{U_{(k)}}(u)\,\mathrm{d}u = \frac{n!}{(k-1)!\,(n-k)!}\,u^{k-1}(1-u)^{n-k}\,\mathrm{d}u, \tag{2.115}$$

which is Beta$(k, n-k+1)$, with $B(k, n-k+1) = \frac{(k-1)!(n-k)!}{n!}$ [CasellaBerger Ex. 5.4.5]. For a general continuous parent with cumulative distribution F and density f, the same count gives $\frac{n!}{(k-1)!(n-k)!}F(x)^{k-1}[1-F(x)]^{n-k}f(x)$ [CasellaBerger Thm 5.4.4]. Casella and Berger prove it by differentiating the binomial probability that at least k of the n values lie below x. The same counting gives the joint density of two order statistics, and from it the distribution of the range $U_{(n)} - U_{(1)} \sim$ Beta$(n-1, 2)$ [CasellaBerger Thm 5.4.6, Ex. 5.4.7].

Order statistics describe the median, quantiles, the fastest of n photon arrival times (which sets the timing of a scintillator), and the largest of n noise fluctuations (the look-elsewhere effect of Chapter 4). With $n = 10$ and $k = 3$ the script finds the mean and variance of $U_{(3)}$ to be 0.2728 and 0.01659, against $\frac{k}{n+1} = 0.2727$ and $\frac{k(n-k+1)}{(n+1)^2(n+2)} = 0.01653$ (Figure 2.44a).

Machine 2: the share of a total. Let $X \sim$ Gamma$(a, 1)$ and $Y \sim$ Gamma$(b, 1)$ be independent. They could be the waiting times for the a-th and the b-th events of two independent unit-rate processes, or two χ^2 variables. What is the distribution of the share $U = X/(X+Y)$? Change variables to $S = X + Y$ and U, so $X = SU$, $Y = S(1-U)$, with Jacobian $\left|\partial(x,y)/\partial(s,u)\right| = s$:

$$\begin{aligned} f(u,s) &\overset{(1)}{=} s\,\frac{(su)^{a-1}e^{-su}}{\Gamma(a)}\,\frac{(s(1-u))^{b-1}e^{-s(1-u)}}{\Gamma(b)} \\ &\overset{(2)}{=} \underbrace{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\,u^{a-1}(1-u)^{b-1}}_{\text{Beta}(a,b)} \times \underbrace{\frac{s^{a+b-1}e^{-s}}{\Gamma(a+b)}}_{\text{Gamma}(a+b,1)}. \end{aligned} \tag{2.116}$$

(1) The joint density of independent X and Y at $x = su$, $y = s(1-u)$, times the Jacobian. The determinant of $\begin{pmatrix} u & s \\ 1-u & -s \end{pmatrix}$ is $-us - s(1-u) = -s$.

(2) Collect powers: $s \cdot s^{a-1}s^{b-1} = s^{a+b-1}$, and $e^{-su}e^{-s(1-u)} = e^{-s}$. Multiply and divide by $\Gamma(a+b)$.

The density factorises, so the share U is Beta(a,b) and is *independent* of the total S. As a by-product, the second factor integrates to one, so the first must too, which proves $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ [CasellaBerger (3.3.17)]. With $a = 2.5$, $b = 4$ the script finds $\mathbb{E}\,U = 0.3846$ against $a/(a+b) = 0.3846$ (Figure 2.44b). The relation of F to the beta distribution in Section 2.14.2 is this machine applied to two χ^2 variables.

Summary numbers and shapes. Recognising kernels again, $\mathbb{E}\,\theta^r = B(a+r,b)/B(a,b)$ [CasellaBerger (3.3.18)], so

$$\mathbb{E}\,\theta = \frac{a}{a+b}, \qquad \operatorname{Var}\theta = \frac{ab}{(a+b)^2(a+b+1)}, \qquad \text{mode} = \frac{a-1}{a+b-2}\ \ (a,b>1). \tag{2.117}$$

The two shape parameters make the family flexible (Figure 2.43). It is uniform for $a = b = 1$, U-shaped for $a, b < 1$, J-shaped when one of them is below one, and single-peaked for $a, b > 1$. It is symmetric when $a = b$, with variance $1/[4(2a+1)]$.

How do we choose a and b? Kruschke's advice [Kruschke p. 128–131] is to think of a and b as "previously seen" successes and failures. The concentration $\kappa = a + b$ measures how much prior data that is, and

$$a = \mu\kappa, \quad b = (1-\mu)\kappa \qquad \text{or} \qquad a = \omega(\kappa-2)+1, \quad b = (1-\omega)(\kappa-2)+1$$

set the mean μ or the mode ω [Kruschke (6.5)–(6.6)]. A prior with mode 0.8 and concentration 12 is Beta$(9,3)$. From a desired mean and standard deviation, $a = \mu\nu$ and $b = (1-\mu)\nu$ with $\nu = \mu(1-\mu)/\sigma^2 - 1$ [Kruschke (6.7)]: mean 0.5 and standard deviation 0.1 give Beta$(12,12)$. The standard deviation must be below $1/\sqrt{12} = 0.289$, that of the uniform, or the formula returns a U-shaped (or impossible) answer.

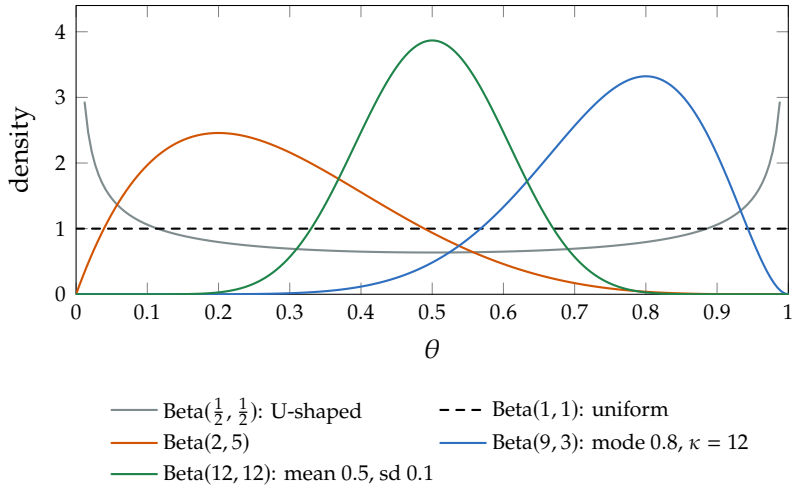


Figure 2.43: Members of the beta family. With both parameters below one the density piles up at the ends; with both equal to one it is flat; above one it has a single peak whose position is set by the ratio of the parameters and whose width shrinks as their sum grows. The blue and green curves are Kruschke's two worked prior specifications.

Working definition 2.20 (beta distribution)

$\theta \sim \text{Beta}(a, b)$, $a, b > 0$, has density (2.114) on $[0, 1]$ and moments (2.117). *Recognise it by:* a probability or a fraction that is itself uncertain. It is the k-th smallest of n uniform numbers (Beta$(k, n - k + 1)$), the share $X/(X + Y)$ of two independent gamma variables with a common scale (Beta(a, b)), and, as shown next, the updated belief about a success probability after $a - 1$ successes and $b - 1$ failures observed from a flat start [with prior Beta(a_0, b_0) and h successes in N trials the result is Beta$(a_0 + h, b_0 + N - h)$].

Machine 3: keep the runs that match. The third machine is inference itself. A coin has unknown bias θ. Before flipping it, our belief about θ is Beta$(2, 2)$: probably near a half, but anything is possible. *The event.* We flip it $N = 10$ times and see $h = 7$ heads. *The question.* Which values of θ could have produced seven heads, and how strongly should we now believe in each? Let the computer answer by brute force, with no formula at all. Draw θ from the prior; flip a coin with that θ ten times; keep θ if and only if the run produced exactly seven heads; repeat.

Listing 2.20: Keep the runs that match: a posterior by rejection (`code/ch02/28_beta_dirichlet.py`)

```
44 # ------------------------------------------------------------- (3) rejection "posterior"
45 a0, b0, N, h = 2.0, 2.0, 10, 7
46 theta = rng.beta(a0, b0, size=2_000_000)
47 heads = rng.binomial(N, theta)
48 kept = theta[heads == h]
49 acc = kept.size / theta.size
50 evidence_th = special.comb(N, h) * special.beta(a0 + h, b0 + N - h) / special.beta(a0, b0)
```

Line 46 draws two million values of θ from the prior. Line 47 runs the *forward* model, the prediction direction of Chapter 1: for each θ, simulate the experiment. Line 48 keeps only those θ that reproduced the data, which is conditioning on the observation, the inference direction. The histogram of the kept values is the answer (Figure 2.44c). *What is it analytically?* Take the candidate values of θ one at a time. A given θ is kept with the binomial probability of the data, $\binom{10}{7}\theta^7(1-\theta)^3$. It was drawn in the first place with the prior density. So the kept values have density proportional to the product of the two,

$$\underbrace{\theta(1-\theta)}_{\text{Beta(2,2) kernel}} \times \underbrace{\theta^7(1-\theta)^3}_{\text{probability of the data}} = \theta^8(1-\theta)^4,$$

the kernel of Beta$(9, 5)$. In general, prior Beta(a_0, b_0) and binomial data give Beta(a_0+h, b_0+N-h). The family is **conjugate** to the binomial: prior and posterior belong to the same family. And "a and b are previous successes and failures" is literally true. The script's kept values have mean 0.6428 and standard deviation 0.1240, against $\frac{9}{14} = 0.6429$ and 0.1237.

The acceptance fraction has a meaning too. It is 0.1123, the probability of seeing seven heads averaged over the prior, $\binom{10}{7}B(9, 5)/B(2, 2) = 0.1119$. That is the *evidence*, the denominator of Bayes' theorem. So this little machine is the whole of Chapter 5 in miniature. Rejection wastes 89% of the draws here. It would waste nearly all of them for a continuous observation or many parameters, which is why Chapter 9 replaces it by Markov chain Monte Carlo. The same logic makes the gamma distribution conjugate to the Poisson rate: a Gamma(α, β) prior times the Poisson likelihood $\lambda^n e^{-\lambda}$ is again a gamma kernel.

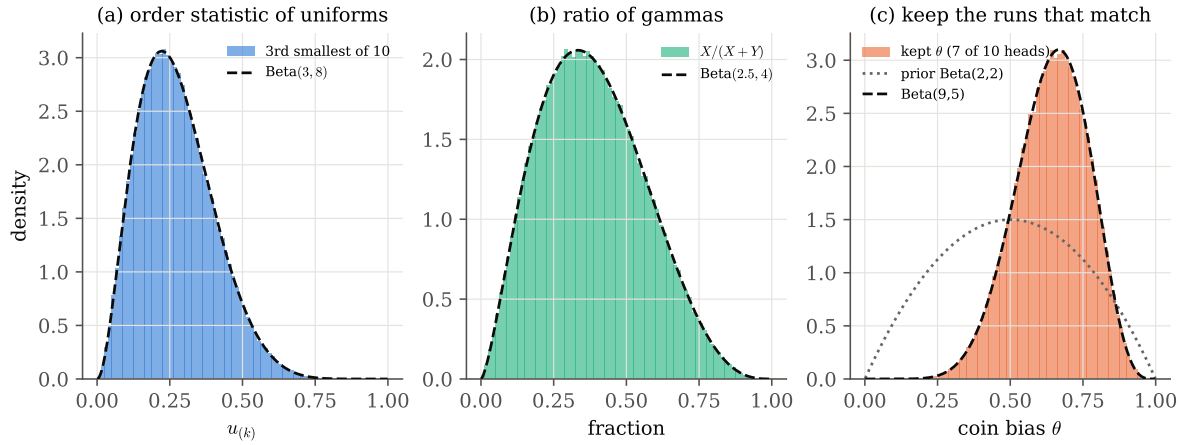


Figure 2.44: Three machines that make a beta distribution. (a) The third smallest of ten uniform numbers. (b) The share $X/(X+Y)$ of two independent gamma variables. (c) Coin biases drawn from a Beta(2, 2) prior (dotted) and kept only when their simulated ten flips gave seven heads: the survivors follow Beta(9, 5).

2.15.6 Dirichlet: several probabilities that sum to one

A histogram with K bins has bin probabilities $p_1, \ldots, p_K$ with $\sum_j p_j = 1$; an election has vote shares; a decay has branching ratios. The distribution for such a vector is the **Dirichlet**, and the gamma machine of (2.116) (Machine 2, the share of a total) builds it. Draw independent $G_j \sim \text{Gamma}(\alpha_j, 1)$, $j = 1, \ldots, K$, and normalise, $p_j = G_j / \sum_l G_l$. The same change of variables, now with K variables, gives the density on the simplex [the set of vectors with all $p_j \geq 0$ and $\sum_j p_j = 1$, a triangle for $K = 3$]

$$f(p_1, \ldots, p_K) = \frac{\Gamma(\alpha_0)}{\prod_j \Gamma(\alpha_j)} \prod_{j=1}^{K} p_j^{\alpha_j - 1}, \qquad \alpha_0 = \sum_j \alpha_j, \tag{2.118}$$

which reduces to the beta for $K = 2$. Its properties follow from the machine without any integration. *Marginals.* One component is $p_1 = G_1/(G_1 + R)$ with $R = \sum_{j \geq 2} G_j \sim \text{Gamma}(\alpha_0 - \alpha_1, 1)$ by (2.107), so by machine 2 each marginal is a beta,

$$p_j \sim \text{Beta}(\alpha_j, \alpha_0 - \alpha_j), \qquad \mathbb{E}\, p_j = \frac{\alpha_j}{\alpha_0}, \qquad \text{Var}\, p_j = \frac{\alpha_j(\alpha_0 - \alpha_j)}{\alpha_0^2(\alpha_0 + 1)}.$$

Covariances. These follow from the constraint $\sum_j p_j = 1$. For $K = 3$, $p_1 + p_2 = 1 - p_3$, so

$$\begin{aligned} 2\,\text{Cov}(p_1, p_2) &\overset{(1)}{=} \text{Var}\, p_3 - \text{Var}\, p_1 - \text{Var}\, p_2 \\ &\overset{(2)}{=} \frac{\alpha_3(\alpha_0 - \alpha_3) - \alpha_1(\alpha_0 - \alpha_1) - \alpha_2(\alpha_0 - \alpha_2)}{\alpha_0^2(\alpha_0 + 1)} \overset{(3)}{=} \frac{-2\alpha_1\alpha_2}{\alpha_0^2(\alpha_0 + 1)}. \end{aligned} \tag{2.119}$$

(1) $\text{Var}(p_1 + p_2) = \text{Var}(1 - p_3) = \text{Var}\, p_3$, and $\text{Var}(p_1 + p_2) = \text{Var}\, p_1 + \text{Var}\, p_2 + 2\,\text{Cov}(p_1, p_2)$.

(2) Insert the beta variances.

(3) Write $s = \alpha_1 + \alpha_2$, so $\alpha_3 = \alpha_0 - s$. The numerator is $(\alpha_0 - s)s - \alpha_1(\alpha_0 - \alpha_1) - \alpha_2(\alpha_0 - \alpha_2) = -s^2 + \alpha_1^2 + \alpha_2^2 = -2\alpha_1\alpha_2$.

So $\text{Cov}(p_i, p_j) = -\alpha_i\alpha_j/[\alpha_0^2(\alpha_0 + 1)]$ for $i \neq j$ (for $K > 3$ lump the other components into one, which is again gamma), the same negative-correlation pattern as the multinomial counts of (2.9): if one share grows, the others must shrink.

For Lambert's three-party election with Dirichlet$(4,2,2)$ [Lambert §8.5.1], the script finds $\mathbb{E}\,p_1 = 0.4999$ (0.5000), $\mathrm{Var}\,p_1 = 0.02767$ (0.02778) and $\mathrm{Cov}(p_1,p_2) = -0.01386$ (-0.01389) (Figure 2.45). Two facts decide when to use it. It is conjugate to the multinomial, exactly as the beta is to the binomial: prior Dirichlet$(\boldsymbol{\alpha})$ and bin counts $\boldsymbol{n}$ give Dirichlet$(\boldsymbol{\alpha}+\boldsymbol{n})$. And, as Lambert warns, its correlations are all fixed by the α's, so it cannot express a belief such as "if party 2 does well, so does party 3"; logistic-normal priors are used when that matters.

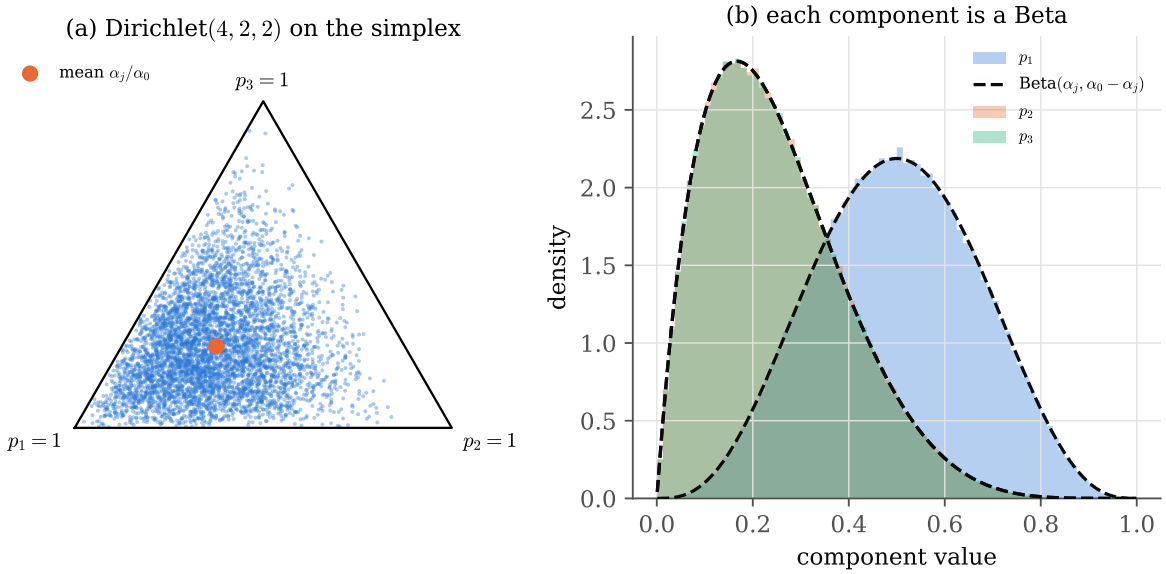


Figure 2.45: (a) Draws from Dirichlet$(4,2,2)$ on the triangle of possible vote shares: each corner is one party winning everything, and the cloud sits nearer the corner of the party with the largest α. (b) Each share on its own follows a beta distribution (dashed curves).

2.15.7 Two relatives from earlier sections

Two more distributions complete the zoo, and both have appeared already. The **Laplace** or double-exponential density $\frac{1}{2\sigma}e^{-|x-\mu|/\sigma}$, with Var $= 2\sigma^2$ [CasellaBerger (3.3.22)–(3.3.23)], was the MaxEnt answer for a known mean absolute deviation. Its machine is the difference of two independent exponential waits with the same mean σ. To check, multiply the characteristic functions: $(1-i\sigma k)^{-1}(1+i\sigma k)^{-1} = (1+\sigma^2k^2)^{-1}$, which is the transform of the Laplace density. It has a cusp at the centre and exponential tails, fatter than Gaussian but with all moments finite. The **Cauchy** (Breit–Wigner) distribution of Section 2.10.5 has three machines: the line shape of an unstable state, the ratio of two independent standard normals [CasellaBerger p. 89], and Student's t_1. Its location parameter is the median, even though it has no mean [CasellaBerger (3.3.19)–(3.3.20)].

One family tree. Stepping back, nearly everything in this part hangs from two roots. From the *Gaussian*: squares give χ^2, lengths give Rayleigh and Maxwell, ratios give t, F and Cauchy, exponentials give the log-normal, linear maps give the multivariate normal. From the *exponential* (the waiting time of a Poisson process): sums give the gamma, shares of sums give the beta and the Dirichlet, powers give the Weibull, differences give the Laplace. The two trees meet at $\chi^2_\nu = \mathrm{Gamma}(\nu/2,2)$, and χ^2_2 is itself an exponential. Section 2.16 collects them into one table.

2.16 Which distribution when

A new set of numbers arrives: the invariant masses of a few thousand dielectron events, the output of a matched filter running over a month of detector noise, the χ^2 of a fit that took a week to set up. Someone asks what distribution the numbers follow. The tempting answer is to make a histogram and pick whichever curve from the zoo looks closest. The better order is the opposite one. Every distribution of this chapter came out of a machine, and the machine is what we can know before seeing a single number. So the question "which distribution?" becomes "which machine made these numbers?". The histogram comes afterwards, to check the answer where the check is most demanding: in the tails.

Choosing by machine is a procedure, and it can be written down. The machines fit into one family tree and one table. The procedure works on all three kinds of data the book keeps returning to: collider counts, detector noise and fitted spectra. And it has one firm warning attached: two distributions that agree in mean and variance can disagree by factors of thousands where discoveries are claimed.

2.16.1 Reading the mechanism

Before fitting any curve, we answer four questions about how the number was made.

1. *Is it a count or a continuous measurement?* Counts go to the counting machines of Section 2.1 to Section 2.8, whose table ends the previous part. Continuous measurements go to this part.
2. *Which operation produced it?* Adding many small pieces, multiplying many factors, squaring and adding, taking the length of a vector, dividing one random quantity by another, waiting for an event, taking a share of a total, or sorting.
3. *Are the ingredients independent, and do they have a finite variance?* If the answer is no, the limit theorems that manufacture Gaussians and χ^2's are switched off, and a Cauchy or a Landau law (Sections 2.10.5 and 2.10.6) may be what arrives.
4. *Which part of the distribution will the conclusion rest on?* An average and its error bar use the centre, where almost any sensible model agrees. A p-value, a discovery claim or a false-alarm rate uses the tail, where models disagree most.

Lambert draws the same procedure as a decision tree for choosing a likelihood [Lambert §8.6, Fig. 8.33], and adds the practical advice to err on the side of robustness [Lambert §8.8]: when unsure, prefer the model with heavier tails (the negative binomial over the Poisson, Student's t over the Gaussian).

Most continuous distributions descend from just two parents, the Gaussian and the exponential. Figure 2.46 arranges the machines as two trees [Lambert §8.3, Fig. 8.1]. Each box names a distribution and, underneath, the operation that makes it from its parent. The Gaussian tree is the world of sums, the exponential tree the world of waiting. The dashed line is the one place where they are literally the same distribution: $\chi^2_\nu = \text{Gamma}(\nu/2, 2)$, so a sum of ν squared Gaussians is the wait for the $(\nu/2)$-th event of a Poisson stream of rate $\frac{1}{2}$. Above both sits the uniform distribution, which by the inverse-transform rule of Section 2.15.1 can be turned into any of them.

The table collects the continuous machines with their summary numbers and the places they turn up in physics. Every mean and variance in it was derived earlier in the chapter, and the section that derived it also gives the density.

Taxonomy: which continuous distribution, when

mechanism	distribution	mean; variance	physics examples
sum of many small independent pieces with finite variance; or only mean and variance known	$\mathcal{N}(\mu, \sigma^2)$	μ; σ^2	measurement and thermal noise, large-λ counts, real and imaginary part of a noise Fourier bin
linear map $A\boldsymbol{Z} + \boldsymbol{\mu}$ of independent standard normals	$\mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$	$\boldsymbol{\mu}$; $\mathbf{C} = AA^{\mathsf{T}}$	correlated parameters, pixels of a Gaussian field, Fisher forecasts
only a range known; $F(X)$ of any continuous X	$\mathrm{U}(a, b)$	$\frac{a+b}{2}$; $\frac{(b-a)^2}{12}$	strip pitch/$\sqrt{12}$, rounding, random phase, p-values under H_0
wait for the first event of a constant-rate stream; power of a complex Gaussian mode	$\mathrm{Exp}(\lambda)$	$\frac{1}{\lambda}$; $\frac{1}{\lambda^2}$	decay times, arrival gaps, periodogram power, $\lvert a_{\ell m}\rvert^2$
sum of α waits; χ^2 with a scale	$\mathrm{Gamma}(\alpha, \beta)$	$\alpha\beta$; $\alpha\beta^2$	time to the n-th photon, prior on a Poisson rate
failure rate a power of age	Weibull, $Y = X^{1/\gamma}$, $X \sim \mathrm{Exp}(1)$	$\Gamma(1 + \frac{1}{\gamma})$; $\Gamma(1 + \frac{2}{\gamma}) - \Gamma(1 + \frac{1}{\gamma})^2$	lifetimes of tubes and electronics, extremes
product of many positive independent factors	log-normal, $\ln X \sim \mathcal{N}(\mu, \sigma^2)$	$e^{\mu+\sigma^2/2}$; $e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)$	photomultiplier gain, log-normal density-field models
sum of ν squared independent standard normals	χ^2_ν	ν; 2ν	goodness of fit, $(n-1)S^2/\sigma^2$, $\widehat{C}_\ell$ and cosmic variance
length of a 2-D / 3-D Gaussian vector (scale s)	Rayleigh / Maxwell	$s\sqrt{\pi/2}$; $(2 - \frac{\pi}{2})s^2$ / $s\sqrt{8/\pi}$; $(3 - \frac{8}{\pi})s^2$	Fourier amplitudes, matched-filter SNR in noise / molecular speeds
Gaussian over an independently estimated scale	Student t_ν	0 $(\nu > 1)$; $\frac{\nu}{\nu-2}$ $(\nu > 2)$	mean with an estimated error bar, likelihood robust to outliers
ratio of two independent χ^2/dof	$F_{p,q}$	$\frac{q}{q-2}$ $(q > 2)$	comparing two resolutions, nested fits
resonance line shape; ratio of two normals	Cauchy (Breit–Wigner)	none; none (median M, half-width γ)	unstable-particle mass, natural line width
resonance seen with Gaussian resolution	Voigt	none; none	measured mass peaks, pressure-broadened lines
energy lost in a thin layer	Landau	none; none	$\mathrm{d}E/\mathrm{d}x$ in silicon and gas
uncertain probability; k-th of n uniforms; share of two gammas	$\mathrm{Beta}(a, b)$	$\frac{a}{a+b}$; $\frac{ab}{(a+b)^2(a+b+1)}$	efficiencies, branching fractions, prior on a success probability
uncertain probability vector	$\mathrm{Dirichlet}(\boldsymbol{\alpha})$	$\frac{\alpha_j}{\alpha_0}$; $\frac{\alpha_j(\alpha_0-\alpha_j)}{\alpha_0^2(\alpha_0+1)}$	bin probabilities of a histogram, branching ratios
difference of two equal exponentials; known mean absolute deviation	$\mathrm{Laplace}(\mu, \sigma)$	μ; $2\sigma^2$	least-absolute-deviation fits

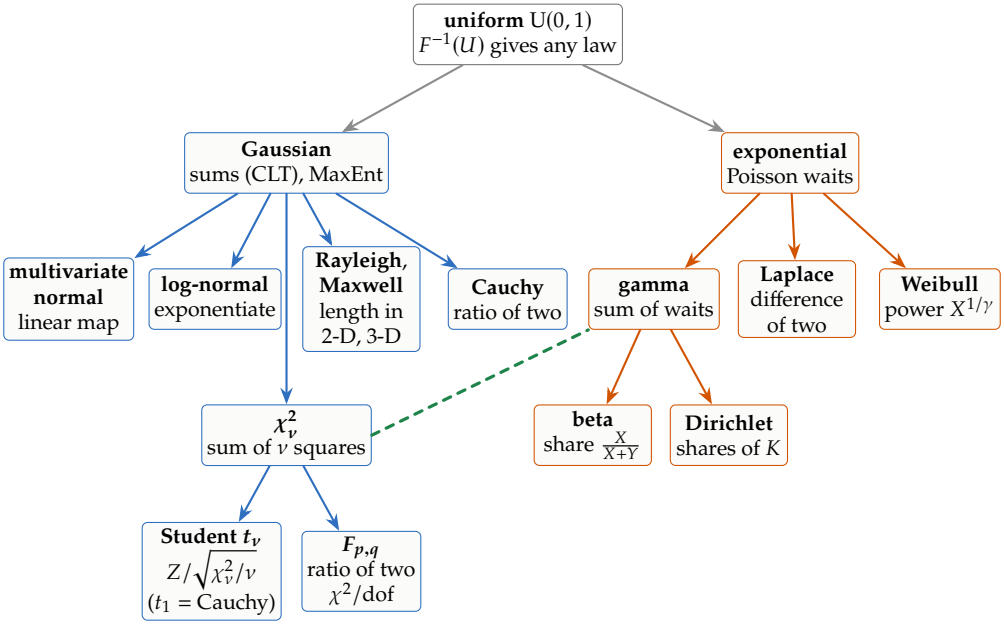


Figure 2.46: Two family trees hold most of the continuous zoo. Blue: everything built from Gaussians, with the building operation written under each name. Orange: everything built from the exponential waiting time of a Poisson process. The green dashed line joins χ^2_ν and the gamma family, which are the same distribution. The uniform at the top feeds both trees, since any law is F^{-1} of a uniform number.

One row of the table is new, the Voigt profile. It is the first of three recognition problems, and it shows how two rows combine.

2.16.2 Three recognitions

Example: a Z peak seen through a detector. In Section 2.10.5 the Z line shape was the Breit–Wigner (2.58), with detector resolution ignored. A real calorimeter adds to each true mass B an independent Gaussian error $G \sim \mathcal{N}(0, \sigma^2)$, with $\sigma = 2.0\,\text{GeV}$ for our example. The measured mass is $m = B + G$. *The question:* does the smearing turn the peak into a Gaussian, and what happens to the tails? *Reading the mechanism:* a *sum* of two independent pieces, so the characteristic function is a product:

$$\phi_m(k) \overset{(1)}{=} \mathbb{E}\big(e^{ikB}\big)\;\mathbb{E}\big(e^{ikG}\big) \overset{(2)}{=} e^{iMk-\gamma|k|}\;e^{-\sigma^2k^2/2}, \qquad \gamma = \Gamma/2. \tag{2.120}$$

(1) $e^{ik(B+G)} = e^{ikB}e^{ikG}$, and the expectation of a product of independent factors is the product of expectations, (2.47).

(2) The Cauchy transform (2.59) with centre M and half-width γ, and the Gaussian transform (2.50) with mean zero.

The inverse transform of (2.120) is the **Voigt profile**, the convolution of the two densities. It has no closed form, but the characteristic function already answers the question: no, the Cauchy tail survives. First, ϕ_m still has the kink $|k|$ at $k = 0$, so it is not differentiable there and m still has no mean: Gaussian smearing cannot cure a Cauchy tail. Second, the tail itself survives unchanged. For $|m - M|$ much larger than both σ and γ,

$$\begin{aligned} p_m(m) &\overset{(1)}{=} \int_{-\infty}^{\infty} p_B(m-g)\,p_G(g)\,\mathrm{d}g \overset{(2)}{=} p_B(m)\int_{-\infty}^{\infty} p_G(g)\,\mathrm{d}g\,\big[1 + O(\sigma^2/(m-M)^2)\big] \\ &\overset{(3)}{=} \frac{1}{\pi}\,\frac{\gamma}{(m-M)^2+\gamma^2}\,\big[1+\ldots\big] \overset{(4)}{=} \frac{\gamma}{\pi\,(m-M)^2}\,\big[1+\ldots\big]. \end{aligned} \tag{2.121}$$

(1) The density of a sum of independent variables is the convolution of their densities: to reach m, the Gaussian must supply g and the resonance $m - g$.
(2) p_G is non-negligible only for $|g| \lesssim 3\sigma$, and over that range the Breit–Wigner, far out in its tail, changes little. Expanding $p_B(m - g)$ to second order in g, the linear term averages to zero because $\mathbb{E}\, G = 0$, and the quadratic term is smaller by a factor of order $\sigma^2/(m - M)^2$.
(3) p_G integrates to one, and (2.58) written with $\gamma = \Gamma/2$.
(4) $(m - M)^2 \gg \gamma^2$.

So the fraction of events more than Δ from the peak is, for $\Delta \gg \sigma, \gamma$,

$$\mathrm{P}(|m - M| > \Delta) \overset{(1)}{=} 2 \int_{\Delta}^{\infty} \frac{\gamma}{\pi u^2}\, \mathrm{d}u \overset{(2)}{=} \frac{2\gamma}{\pi\Delta}, \tag{2.122}$$

(1) The symmetry of the two tails and the approximation (2.121) with $u = |m - M|$.
(2) $\int_{\Delta}^{\infty} u^{-2}\mathrm{d}u = 1/\Delta$.

For $\gamma = 1.2476\,\mathrm{GeV}$ and $\Delta = 10\,\mathrm{GeV}$ this gives 0.0794.

Listing 2.21: A Breit–Wigner resonance smeared by Gaussian resolution (`code/ch02/29_which_distribution.py`)

```
M, Gam, sig = 91.1876, 2.4952, 2.0          # GeV: Z mass, full width, detector resolution
gam = Gam / 2                               # Cauchy half-width
m = M + gam * rng.standard_cauchy(NREP) + rng.normal(0.0, sig, NREP)
Delta = 10.0
out_sim = np.mean(np.abs(m - M) > Delta)
uu = np.linspace(-Delta, Delta, 200_001)
core = np.trapezoid(voigt_profile(uu, sig, gam), uu)
out_voigt = 1.0 - core
out_tail = 2 * gam / (np.pi * Delta)        # Cauchy tail estimate, valid for Delta >> sigma, gamma
# FWHM of the Voigt density, found numerically, and the Gaussian with the same FWHM
grid = np.linspace(0, 10, 200_001)
vp = voigt_profile(grid, sig, gam)
fwhm = 2 * grid[np.argmin(np.abs(vp - vp[0] / 2))]
sig_eq = fwhm / (2 * np.sqrt(2 * np.log(2)))
out_gauss = 2 * stats.norm.sf(Delta / sig_eq)
```

The script builds the machine literally, and the Gaussian fails by a factor of several hundred in the wings. Line 39 adds a Cauchy draw of half-width γ and a Gaussian draw of width σ to the Z mass, for each of 2×10^5 events. Line 41 counts the fraction of events beyond Δ. Lines 42–44 integrate the exact Voigt density (`scipy.special.voigt_profile`, whose arguments are the Gaussian σ and the Cauchy half-width γ) over the core. Lines 47–51 find the full width at half maximum of the Voigt profile and build the Gaussian with the same width, the curve one would get by fitting a Gaussian to the peak by eye.

The three estimates of the tail agree: the simulated fraction beyond 10 GeV is 0.0820, the exact Voigt value is 0.0825, and the tail estimate (2.122) is 0.0794. The Voigt peak has a full width at half maximum of 6.18 GeV, and a Gaussian with that width ($\sigma_{\mathrm{eq}} = 2.63\,\mathrm{GeV}$) puts only 1.4×10^{-4} of its probability beyond 10 GeV. On a linear plot the two curves are hard to tell apart. On the logarithmic plot of Figure 2.47b the difference is a factor of several hundred. An analysis that modelled the peak as a Gaussian would find about 8% of the Z events in its "background" sidebands.

Example: false alarms in a gravitational-wave search. A matched filter correlates the detector output with a template waveform whose phase is unknown. It returns two numbers at each time, the correlation with the cosine-phase and with the sine-phase template, normalised so

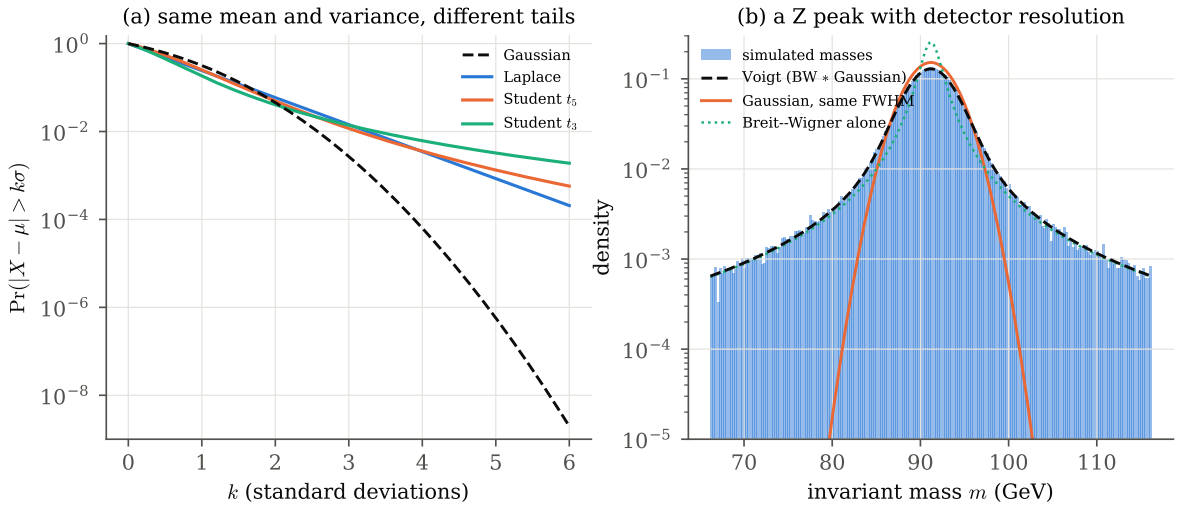


Figure 2.47: (a) Two-sided tail probability against distance from the mean, in units of the standard deviation, for four symmetric distributions scaled to the same mean and variance. They agree within a factor of two out to about two standard deviations and then separate by orders of magnitude. (b) Simulated Z masses with a Breit–Wigner width and Gaussian detector resolution, on a logarithmic scale. The Voigt density (dashed) follows them into the wings; the Gaussian with the same half-maximum width (solid) matches only the core; far from the peak the Voigt rejoins the pure Breit–Wigner (dotted).

that in pure Gaussian noise each is a standard normal and the two are independent. The signal-to-noise ratio is $\rho = \sqrt{x^2 + y^2}$, with x and y the two correlations. *The question:* how often does noise alone push ρ above a threshold? *Reading the mechanism: the length of a two-dimensional Gaussian vector.* So ρ is Rayleigh with $s = 1$, and ρ^2 is χ^2_2, the exponential with mean 2 (Section 2.15.4). The chance that noise alone crosses a threshold ρ_* is

$$\mathrm{P}(\rho > \rho_*) \overset{(1)}{=} \mathrm{P}(\rho^2 > \rho_*^2) \overset{(2)}{=} \int_{\rho_*^2}^{\infty} \tfrac{1}{2} e^{-u/2}\,\mathrm{d}u \overset{(3)}{=} e^{-\rho_*^2/2}. \tag{2.123}$$

(1) $\rho \geq 0$, and squaring is increasing on $[0, \infty)$.
(2) The χ^2_ν density (2.92) with $\nu = 2$ is $u^0 e^{-u/2}/(2\,\Gamma(1)) = \tfrac{1}{2} e^{-u/2}$.
(3) $\int_a^\infty \tfrac{1}{2} e^{-u/2}\mathrm{d}u = e^{-a/2}$.

The script checks the formula at $\rho_* = 3$: the fraction of 2×10^5 noise draws above threshold is 0.0112, against $e^{-9/2} = 0.0111$ (script lines 54–59). At the conventional detection threshold $\rho_* = 8$ the formula gives 1.27×10^{-14} per independent template and time sample. That looks reassuringly small. But a search examines very many templates at every sample of months of data, so the answer depends on the χ^2_2 law holding fourteen decades down its tail. Real detector noise contains short non-Gaussian bursts ("glitches"), which is exactly the kind of fat tail of Figure 2.47a.

So searches do not trust (2.123) for their final significance. They measure the background empirically. They shift one detector's data in time against another's, by more than the light travel time, and count how often noise alone produces a coincident trigger. The formula *verifies* the pipeline on simulated Gaussian noise; the time slides are the research instrument on real data. The same Rayleigh and exponential laws describe the amplitude and the power of a noise frequency bin, which returns in the gravitational-wave Fisher forecasts of Chapter 10, where the noise enters through its power spectral density.

Example: is $\chi^2 = 130$ for 100 degrees of freedom a bad fit? It is suspicious but not damning, and the Gaussian shortcut gets its tail wrong by 40%. A fit of a model with P adjustable parameters to N Gaussian measurements returns a minimum χ^2. For a model linear in its parameters, the minimum is a quadratic form in the residuals with $N-P$ independent directions left over (the argument of Sections 2.13.3 and 2.13.4, made precise in Chapter 3). Reading the mechanism: *a sum of $\nu = N - P$ squared standard normals*, so $\chi^2_{\min} \sim \chi^2_\nu$. *The question:* how often would a correct model give $\chi^2 \geq 130$? With $\nu = 100$ it is tempting to use the central limit theorem, since χ^2_{100} is a sum of 100 independent χ^2_1 variables, and write $\chi^2 \approx \mathcal{N}(\nu, 2\nu)$:

$$z = \frac{\chi^2 - \nu}{\sqrt{2\nu}} = \frac{130 - 100}{\sqrt{200}} = 2.12, \qquad \mathrm{P}(Z > z) = 0.0169.$$

The exact tail, from `scipy.stats.chi2.sf(130, 100)`, is 0.0235, about 40% larger. *Why?* The parent χ^2_1 is skewed, and the CLT removes skewness only slowly (Section 2.10.3). To measure it, we need the cumulants κ_n (the Taylor coefficients of $\ln\phi$). The cumulants of χ^2_1 follow from its characteristic function $(1-2ik)^{-1/2}$ (2.93): $\ln\phi = -\frac{1}{2}\ln(1-2ik) = \sum_{n\geq 1}\frac{(2ik)^n}{2n}$, and matching with $\ln\phi = \sum_n \kappa_n (ik)^n/n!$ gives $\kappa_n = 2^{n-1}(n-1)!$, so $\kappa_2 = 2$ and $\kappa_3 = 8$. Then

$$\gamma_1\big(\chi^2_\nu\big) \overset{(1)}{=} \frac{\kappa_3(\chi^2_\nu)}{\kappa_2(\chi^2_\nu)^{3/2}} \overset{(2)}{=} \frac{\nu\,\kappa_3(\chi^2_1)}{\big[\nu\,\kappa_2(\chi^2_1)\big]^{3/2}} \overset{(3)}{=} \frac{8\nu}{(2\nu)^{3/2}} = \sqrt{\frac{8}{\nu}}. \tag{2.124}$$

(1) Skewness is the third cumulant in units of the standard deviation cubed, with the cumulants of (2.54).
(2) Cumulants of independent variables add, and χ^2_ν is the sum of ν independent χ^2_1.
(3) $\kappa_3(\chi^2_1) = 8$ and $\kappa_2(\chi^2_1) = 2$.

At $\nu = 100$ the skewness is still 0.28. So the right tail is longer than the Gaussian's, and the Gaussian approximation underestimates how often a correct model gives a large χ^2. The verdict itself is mild. A correct model would give $\chi^2 \geq 130$ in about 2% of repetitions, so the fit is suspicious but not damning, and Chapter 4 turns this reasoning into a test. The practical rule is simpler: never use $\mathcal{N}(\nu, 2\nu)$ for a tail probability when the exact χ^2_ν survival function is one line of code.

2.16.3 The tails decide

The three examples share a lesson: the core of each distribution was well described by a Gaussian, and the question was about the tail. Figure 2.47a puts numbers on it. It shows four symmetric distributions, each scaled to mean zero and variance one: the Gaussian, the Laplace, and Student's t_5 and t_3 (for t_ν with scale c the variance is $c^2\nu/(\nu-2)$ by (2.103), so $c^2 = 3/5$ and $1/3$). They are indistinguishable at one standard deviation and cross near two. At three standard deviations the two-sided tail probabilities are 0.0027 for the Gaussian but 0.0144 (Laplace), 0.0117 (t_5) and 0.0138 (t_3), four to five times more. At five standard deviations the Gaussian gives 5.7×10^{-7} and t_3 gives 3.2×10^{-3}, a ratio of 5650.

Key idea: Choose by mechanism, check in the tails

A distribution is chosen by recognising the machine that made the numbers (sum, product, squares, length, ratio, wait, share). Its mean and variance are then fixed by the machine, but agreement in mean and variance says almost nothing beyond about two standard deviations. Any conclusion that uses a tail probability must justify the tail separately, by the mechanism or by an empirical measurement of the background.

The physics of the tails is usually different from the physics of the core. Cowan makes the point for the two distributions with divergent moments, the Breit–Wigner and the Landau. A real resonance mass is bounded by the masses of its decay products and by the collision energy, and a real energy loss by the energy of the incoming particle, so "the models break down in the tails of the distributions" [Cowan pp. 38–39]. Near the peak the idealised machine is accurate; far from it, other mechanisms take over (kinematic limits, radiation, pile-up, detector glitches, non-Gaussian foregrounds).

So particle physicists quote a "5σ" discovery as a tail probability of 2.9×10^{-7} computed from the *distribution of the test statistic*, obtained by simulation or by an asymptotic theorem (Chapter 4), and never from a Gaussian assumption about the data alone. The same caution applies to a posterior interval far from the peak of a posterior (Chapter 5), and to the Fisher forecast of Chapter 10, which is Gaussian by construction and must be checked against a sampled posterior in the tails (Chapter 9).

Watch out: A good fit to the peak is not a test of the tails

A histogram with most of its entries within $\pm 2\sigma$ fits a Gaussian, a Laplace or a t_5 almost equally well: the χ^2 of the fit is dominated by the populated bins, and the tail bins hold too few entries to discriminate. The choice between them must come from the mechanism, or from a dedicated sample that populates the tails (a control region, a time slide, a large simulation).

With a mechanism in hand, the summary numbers of the table give a first check that costs nothing: compare the sample mean and variance with the predicted ones, the Fano factor for counts (Section 2.8), and, when enough data exist, the sample skewness with the predicted γ_1. A mismatch in these moments rules the mechanism out; a match leaves the tails still to be checked.

Recap: Part summary: continuous distributions

- **Gaussian, three ways.** (i) The binomial near its peak, by Stirling's formula and a second-order expansion of $\ln p_k$ (de Moivre–Laplace), with error $O(1/n)$ at the peak. (ii) The central limit theorem: the characteristic function of a standardised sum tends to $e^{-k^2/2}$ whenever the parent has finite variance, with error $O(n^{-1/2})$ set by the skewness (Berry–Esseen). (iii) Maximum entropy: the least committal density with a given mean and variance.
- **When the CLT fails or is slow.** Cauchy (Breit–Wigner): no mean, and the average of n draws is as wide as one draw. Landau: finite range in reality, but so skewed that the mean of 25 layers is still far from Gaussian.
- **Multivariate normal.** $\boldsymbol{X} = \boldsymbol{\mu} + A\boldsymbol{Z}$ with $\mathbf{C} = AA^{\mathsf{T}}$; contours are ellipses, and the region $\Delta\chi^2 \le 2.30$ (two parameters) holds 68.3%; marginals drop rows and columns of $\mathbf{C}$; conditionals shift the mean by $\mathbf{C}_{ba}\mathbf{C}_{aa}^{-1}(\boldsymbol{x}_a - \boldsymbol{\mu}_a)$ and shrink the covariance to the Schur complement.
- **χ^2_ν.** Sum of ν squared standard normals, $\chi^2_\nu = \text{Gamma}(\nu/2, 2)$, mean ν, variance 2ν; $(\boldsymbol{X}-\boldsymbol{\mu})^{\mathsf{T}}\mathbf{C}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}) \sim \chi^2_k$; $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$, independent of $\bar{X}$, because estimating the mean uses up one direction.
- **t and F.** Student $t_\nu = Z/\sqrt{\chi^2_\nu/\nu}$: the studentised mean, heavier tails, variance $\nu/(\nu-2)$, a scale mixture of Gaussians; $F_{p,q}$ is a ratio of two scaled χ^2 and $t_q^2 = F_{1,q}$.

- **The zoo.** Uniform (only a range; parent of all through F^{-1}), gamma and Weibull (waits), log-normal (products), Rayleigh and Maxwell (lengths of Gaussian vectors; a complex Gaussian mode has exponential power), beta and Dirichlet (uncertain probabilities, conjugate to binomial and multinomial counts), Laplace (differences of waits).
- **Which one when.** Recognise the mechanism first, then check moments, and check tails separately: laws with equal variance can differ by factors of thousands at 5σ.

Chapter summary

Every common probability distribution comes from a machine that generates it. Counting machines give the discrete laws: one trial, many trials, histogram bins, rare events, a random stream in time, the waits between its events, and the ways real detectors break these assumptions. Measuring machines give the continuous laws: the Gaussian three times over, then the multivariate normal, the χ^2, Student's t and the F ratio built from it, and the rest of the continuous zoo built from the Gaussian and the exponential. All of it runs in the direction of *prediction*: fix a hypothesis and ask what data it produces and how often. On the pipeline of Chapter 0, that is the step which says how a statistic S scatters given the parameters θ, $P(S \mid \theta)$, for the simplest statistics (counts, sums, sample means and variances, χ^2's, power-spectrum estimates). Every later chapter either reads such a $P(S \mid \theta)$ backwards as a likelihood or builds a new one by simulation. The two tables list each concept, what it means operationally (how to compute, recognise or test it), and where it is used next.

Counting distributions (first part)

concept	operational meaning	used later in
distribution as a mechanism	state the machine and its assumptions, derive the law, then simulate the machine (not the formula) and compare histograms	every chapter
Bernoulli	one yes/no trial, $p^x(1-p)^{1-x}$; N trials give $p^z(1-p)^{N-z}$, depending only on the number of successes	Chapters 3 and 5
binomial	successes in n independent trials with the same p: count the $\binom{n}{k}$ paths; mean np, variance $np(1-p)$	Chapters 3 to 5
multinomial	a histogram with a fixed total: binomial marginals, covariance $-Np_ip_j$ between bins	Chapters 3 and 4
Poisson	$n \to \infty$, $np = \lambda$ limit of the binomial; mean = variance = λ; closed under sums and thinning	Chapters 3 to 6
Poisson process	constant rate, no memory; $\mathrm{d}P_n/\mathrm{d}t = \lambda(P_{n-1} - P_n)$; with a Poisson total, histogram bins become independent Poisson counts	Chapters 3 and 6
exponential, gamma (Erlang)	wait for the first / n-th event; memoryless; means $1/\lambda$, n/λ	Chapters 5 and 6
geometric, negative binomial	trials to the first / r-th success; also a Poisson with gamma-distributed rate (overdispersion)	Chapters 3 and 5
Fano factor	variance over mean of counts; ≈ 1 Poisson, > 1 fluctuating rate or clumping, < 1 dead time; judge against $\sqrt{2/(k-1)}$	Chapters 4 and 8
dead time	recorded rate $n/(1+n\tau)$ or $ne^{-n\tau}$; correcting it removes bias but inflates the variance by $1 + n\tau$	Chapter 6
large-count limit	all Poisson cumulants equal λ, skewness $1/\sqrt{\lambda}$: Gaussian near the peak, exact Poisson in the tails	Chapters 3 and 4

Continuous distributions (second part)

concept	operational meaning	used later in
Gaussian $\mathcal{N}(\mu, \sigma^2)$	μ the centre, σ the width; probabilities from Φ; 68.3/95.4/99.7% within $1/2/3\sigma$	every chapter
Stirling, Laplace's method	expand the logarithm of a peaked integrand to second order about its maximum	Chapters 5 and 10
de Moivre–Laplace	binomial near its peak $\approx$ Gaussian with mean np, variance $np(1-p)$; error largest in the tails	Chapters 3 and 4
characteristic function	$\mathbb{E}\, e^{ikX}$, always exists; sums multiply it; cumulants are the Taylor coefficients of its logarithm	Chapters 6 and 7
central limit theorem	standardised sums of finite-variance variables tend to $\mathcal{N}(0,1)$; error $\sim \gamma_1/\sqrt{n}$ (Berry–Esseen); check by simulation	Chapters 3, 4 and 6
Cauchy, Landau	no mean: averaging does not narrow a Cauchy; use the median or a truncated mean; Landau sums converge very slowly	Chapters 3 and 6
maximum entropy	least committal distribution given constraints: $p \propto e^{-\sum_j \lambda_j f_j}$; mean $\to$ exponential, variance $\to$ Gaussian	Chapter 5
multivariate normal	$\boldsymbol{\mu} + A\boldsymbol{Z}$ with $\mathbf{C} = AA^{\mathsf{T}}$ (Cholesky); ellipses with $\Delta\chi^2 = 2.30, 6.18$ in 2-D; marginal = delete rows and columns; conditional = Schur complement	Chapters 7, 9 and 10
χ^2_ν	sum of ν squared standard normals; mean ν, variance 2ν; quadratic forms $\boldsymbol{r}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{r}$; exact tail by `chi2.sf`	Chapters 3, 4, 7 and 8
sample variance, $n-1$	$(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$, independent of $\bar{X}$ for Gaussian data; one direction is used by the mean	Chapters 3 and 4
Student t	mean over an estimated error bar; heavier tails, $t_\nu \to \mathcal{N}$ as $\nu \to \infty$; robust likelihood	Chapters 4 and 5
F ratio	ratio of two independent scaled χ^2; compare variances or nested fits; $t^2 = F_{1,q}$	Chapter 4
uniform, inverse transform	$F(X)$ is uniform; $F^{-1}(U)$ samples any law; p-values are uniform under H_0	Chapters 4 and 6
log-normal	products of positive factors; mean $e^{\mu+\sigma^2/2}$ above the median e^{μ}	Chapters 5 and 7
Rayleigh, Maxwell, exponential power	length of a 2-D / 3-D Gaussian vector; a complex Gaussian mode has Rayleigh amplitude, uniform phase, exponential power	Chapters 7, 8 and 10
beta, Dirichlet	uncertain probability (vector); conjugate to binomial (multinomial) counts; "keep the runs that match" gives the posterior	Chapters 5 and 9
which distribution when	recognise the mechanism, check the moments, justify the tails separately	Chapters 4, 10 and 12

Reading the literature

- **Main line.** [Wasserman §§2.3–2.4, 2.10, 5.4] gives the discrete and continuous families, the multivariate normal with its marginals and conditionals, and the central limit theorem with the Berry–Esseen bound, compactly and with short examples; the MGF proof of the CLT is in [Wasserman §5.7.2].
- **The clearest first reading.** [Kruschke Ch. 6]: the beta distribution as a description of belief about a probability, its parameters as previous successes and failures, and how to set them from a mean and a spread.

- **The physicist's version.** [Cowan Ch. 2] for each distribution with a physics setting (including the Breit–Wigner and Landau laws and their divergent moments), and [Cowan Ch. 10] for characteristic functions, the limits binomial → Poisson → Gaussian, the χ^2 from squares, and the practical validity of the CLT.
- **Careful derivations.** [CasellaBerger §§3.2–3.3, 4.5–4.6, 5.3–5.4]: the families with full proofs of their moments, the bivariate and multivariate normal, the independence of $\bar{X}$ and S^2, the t and F densities by Jacobians, and order statistics.
- **Maximum entropy.** [Sivia Ch. 5]: the monkey argument for the entropy, the loaded die, the kangaroo problem, and the exponential, Gaussian, Laplace and Poisson as MaxEnt assignments.
- **Choosing a likelihood and a prior.** [Lambert Ch. 8]: a catalogue of sampling and prior distributions with a story for each, their interrelations, the robust versions (negative binomial, Student t), and a decision tree for choosing a likelihood.
- **Pictures.** [SeeingTheory pp. 38–39]: an interactive demonstration of the central limit theorem.
- **Going further.** The Gaussian approximation to the $\widehat{C}_\ell$ likelihood and its failure at low ℓ, where the $\chi^2_{2\ell+1}$ skewness matters, is treated in [16]; Chapters 7 and 8 build on it.

Chapter 3

Estimators, Error Propagation and the Likelihood

Where this chapter is going

A peak fit gives us an estimate of a particle's mass, just as an average gives us an estimate from repeated readings. To assign a statistical error bar to either number, we have to ask how far the estimate would move if we repeated the measurement under the same conditions.

In the language of Chapter 0, the rule for computing the estimate defines a statistic S. A physical model with parameters θ then predicts how S varies across repetitions. Its sampling distribution $P(S \mid \theta)$ connects the error bar we report for one measurement to the variation we would expect across many such measurements.

The key step is to treat a summary of the data as a **random variable** in its own right. If we imagine repeating the experiment many times, the summary takes a different value each time, so it has a distribution. That distribution has a centre (does it sit on the true value, or is it biased?), a width (the standard error) and correlations with other summaries. From this one step come the answers to several practical questions: why widths shrink like $1/\sqrt{N}$ (the law of large numbers and the central limit theorem), why the sample variance divides by $N - 1$, how noisy a covariance matrix estimated from simulations is, how errors pass through a formula, and how to combine measurements of different quality. The same probability of the data, read *backwards* as a function of the parameters, is the likelihood, the subject of the second half of the chapter.

Many estimators, few ideas. The chapter meets some thirty estimators: the sample mean and the median, twice the mean and the largest reading of a uniform box, the sample variance with three different divisors, covariances and correlations, weighted means, least-squares fits, fits to Poisson counts, posterior means and modes. It is easy to take them for thirty separate tricks. They are not. Each one comes from answering two physical questions, and then applying one of four principles. The questions: how does the theory say the observable quantities change with the parameter θ, and what do we know about the noise on them? The principles: match a predicted average; maximise the likelihood; take the unbiased rule with the least variance; take the rule with the least expected loss. Everything else is implementation. The same principle applied to a different noise law gives a different formula, and several formulas are reached by more than one principle: the weighted mean is both the least-variance combination and the Gaussian maximum-likelihood estimate, and when two different roads end at the same formula, that is good evidence it is the right one. Figure 3.2 lays out the whole tree.

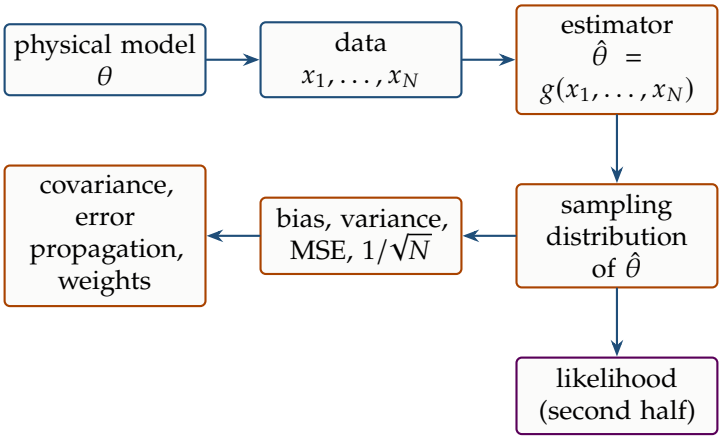


Figure 3.1: The route through the ideas. A model with true parameter θ produces data. A recipe (the estimator) turns data into a number. Imagining many repetitions of the experiment turns that number into a random variable with a sampling distribution (orange boxes). Its centre and width give the error bar, and the same machinery propagates and combines error bars. The last box, the likelihood, reads the probability of the data as a function of θ and is the subject of the second part.

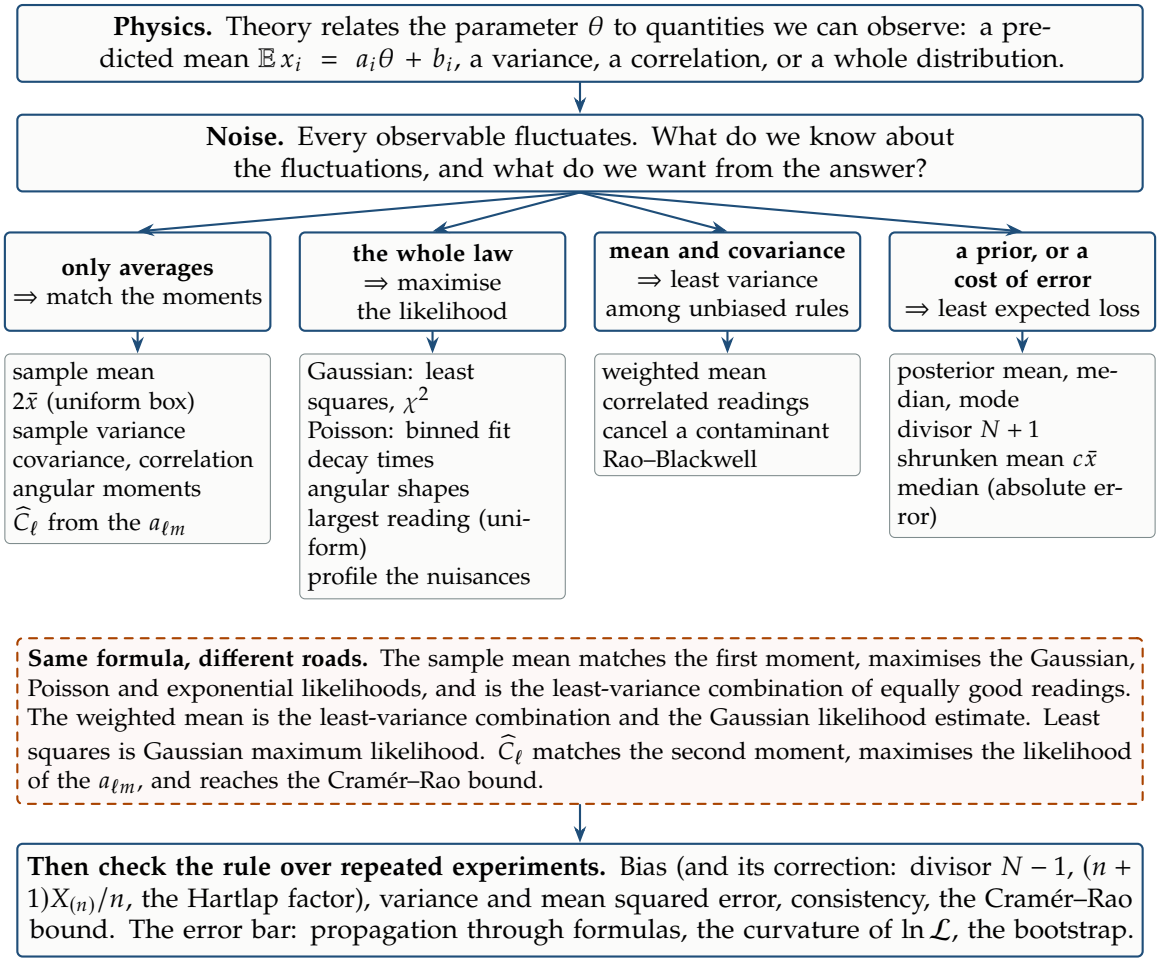


Figure 3.2: The estimators of this chapter as one tree. Two physical questions (top) decide which of four principles applies (middle row); each principle, fed a particular noise law, produces the formulas listed below it. The dashed box records formulas that several principles reach together. Every rule, however it was built, is then judged by the same criteria over imagined repetitions of the experiment (bottom).

3.1 One number from many readings: estimators and their sampling distribution

A student times a pendulum of length $L = 1\,\mathrm{m}$ ten times and converts each period into a value of the gravitational acceleration. The ten numbers are

$$9.79,\ 9.83,\ 9.86,\ 9.80,\ 9.77,\ 9.84,\ 9.81,\ 9.78,\ 9.85,\ 9.82\ \ \mathrm{m\,s^{-2}}.$$

No two agree. What is "the answer"? The average, 9.815? The middle value? The first reading? And whatever we choose, how far from the true g can it be? A particle physicist faces the same question with twenty muon decay times and the lifetime τ_μ, and a cosmologist with one CMB map and the amplitude of fluctuations.

The way forward is to stop asking for "the answer" and choose a *recipe* instead: a fixed rule that turns any list of readings into one number. A rule can be applied to data we did not get. So we can ask how its output would scatter if the experiment were repeated. That scatter is the error bar.

The shape of every problem that follows. In each of these examples there is a quantity we care about and can never look at directly: the true g, the muon lifetime, an amplitude of the early Universe. Call it s. What we hold instead are readings $x_1, \ldots, x_N$, and each of them carries some information about s mixed with things we do not want: noise, a contaminating signal, a calibration offset. Schematically,

$$x_i = (\text{a piece that responds to } s) + (\text{everything else}), \qquad \hat{s} = F(x_1, \ldots, x_N). \tag{3.1}$$

The recipe is the function F. Choosing it, given what we know about how each reading responds to s and about the size and structure of everything else, is the whole subject of estimation. This chapter first learns to judge a given F by the scatter of its output; its second half, the likelihood, and Chapter 5, the posterior, are two systematic ways of choosing F in the first place.

The experiment as a random number generator. Think of the whole experiment, ten readings and all, as one throw of a very complicated die. Nature fixes the true g and the noise level. Each throw produces a new list of ten numbers, and our recipe turns each list into one number. Throw the die many times and those numbers form a histogram. This imagined histogram plays the role of the *ensemble* of statistical mechanics. We only ever see one microstate, our one data set. The error bars we quote, however, are properties of the whole ensemble.

Before any symbol.

1. What did we observe? A list of numbers $x_1, \ldots, x_N$.
2. What could have produced it? A family of probability distributions, labelled by the unknown quantity we care about (the *statistical model*).
3. What recipe turns the list into one number? (The *estimator*.)
4. If nature reran the experiment with the same true value, how would the recipe's output be distributed? (The *sampling distribution*.)
5. What does that distribution say about our one number? Its centre tells us whether the recipe aims at the right place. Its width is the error bar.

Step 4 is a *prediction* in the sense of Chapter 0: we start from a hypothesised truth and ask what outcomes it produces. Every error bar built from a sampling distribution runs in this forward direction, from truth to data. The reverse question, from the data we saw back to the truths that could have produced them, is the likelihood of the second half of the chapter and the Bayesian inference of Chapter 5.

3.1.1 Statistical models: the list of possible truths

Statistical inference is "using data to infer the distribution that generated the data" [Wasserman p. 87]. So before inferring anything, we must write down which distributions could have generated it.

Working definition 3.1 (statistical model)

A **statistical model** $\mathfrak{F}$ is a set of probability distributions [the candidate explanations of how the data were produced]. It is **parametric** if its members can be labelled by finitely many numbers,

$$\mathfrak{F} = \{\, f(x;\theta) \;:\; \theta \in \Theta \,\}, \tag{3.2}$$

where θ is the **parameter** [a fixed but unknown number, or vector, that picks one member] and Θ the **parameter space** [all values θ is allowed to take]. If $\theta = (\theta_1, \theta_2)$ and we only care about θ_1, then θ_2 is a **nuisance parameter** [needed to describe the data, not wanted in the answer]. A model that cannot be labelled by finitely many numbers is **nonparametric**.

The Gaussian model for the pendulum is the two-parameter family

$$\mathfrak{F} = \left\{ f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\!\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \;:\; \mu \in \mathbb{R},\ \sigma > 0 \right\}. \tag{3.3}$$

Here μ is the centre of the readings and σ their noise level. The semicolon in $f(x;\mu,\sigma)$ separates the value of the random variable, x, from the labels of the distribution, μ and σ. If all we want is $g = \mu$, then σ is a nuisance parameter: we need it to describe the scatter, but it is not part of the answer.

A coin flip (or "did this muon decay inside the detector?") is the one-parameter Bernoulli model $f(x;p) = p^x(1-p)^{1-x}$, $x \in \{0,1\}$, where p is the probability of a 1 (see Chapter 2).

The set of *all* distribution functions, $\mathfrak{F}_{\text{ALL}}$, is nonparametric: no finite list of numbers labels its members. Even there we can ask about single numbers computed from the distribution function F, such as the mean $T(F) = \int x\,\mathrm{d}F(x)$, the variance $T(F) = \int x^2 \mathrm{d}F(x) - \left(\int x\,\mathrm{d}F(x)\right)^2$ or the median $T(F) = F^{-1}(1/2)$. Any such function of the whole distribution is called a **statistical functional** [Wasserman p. 89]. Estimating a whole density f is harder: it needs a smoothness assumption, for instance $\int (f'')^2 \mathrm{d}x < \infty$, which says the candidate densities are "not too wiggly". Density estimation comes back with histograms and kernels.

Data = model + noise. A very common model has the form

$$Y = r(X) + \epsilon, \qquad \mathbb{E}(\epsilon) = 0, \tag{3.4}$$

where X is the setting we control (a pendulum length, a frequency), Y the reading we get, and $r(x) = \mathbb{E}(Y \mid X = x)$ the **regression function**, the average reading at setting x. Every measurement written as "the model prediction plus noise" has this form. The condition $\mathbb{E}(\epsilon) = 0$ costs nothing: we can always *define* the noise as the gap between the reading and its

average, $\epsilon := Y - r(X)$, and then its mean is zero automatically:

$$\begin{aligned}\mathbb{E}(\epsilon) &\overset{(1)}{=} \mathbb{E}\big[\mathbb{E}(\epsilon \mid X)\big] \\ &\overset{(2)}{=} \mathbb{E}\big[\mathbb{E}(Y - r(X) \mid X)\big] \\ &\overset{(3)}{=} \mathbb{E}\big[\mathbb{E}(Y \mid X) - r(X)\big] \\ &\overset{(4)}{=} \mathbb{E}\big[r(X) - r(X)\big] = 0. \end{aligned} \tag{3.5}$$

(1) Law of total expectation: average first over Y at fixed X, then over X.
(2) Insert the definition of ϵ.
(3) Once X is fixed, $r(X)$ is a constant and comes out of the conditional average.
(4) By definition $r(X) = \mathbb{E}(Y \mid X)$.

So "zero-mean noise" is not an extra assumption. What *is* an assumption is the *shape* of the noise distribution (Gaussian, Poisson, ...), and that assumption is what the likelihood in the second half of this chapter is built from.

3.1.2 Statistic, estimator, estimate

Lab talk uses three words loosely that we must keep apart: the recipe, the random number it will produce before the data are in, and the value it produces once they are.

Working definition 3.2 (statistic, estimator, estimate)

Let $X_1, \dots, X_N$ be a **sample** [N independent readings, each drawn from the same $f(x;\theta)$]. A **statistic** is any function of the sample that contains no unknown quantity. An **estimator** of θ is a statistic used to guess θ,

$$\hat{\theta}_N = g(X_1, \dots, X_N). \tag{3.6}$$

The **estimate** is the number $g(x_1, \dots, x_N)$ obtained by feeding in the particular data we observed.

The distinction matters. The estimator is a fixed *function*, like "add them up and divide by N". Because its inputs are random, its output is a random variable. The estimate is one realisation of that random variable, the one our universe happened to give us. For the pendulum, the sample mean is an estimator of g, and $9.815\,\mathrm{m\,s^{-2}}$ is an estimate [Cowan §5.1, p. 64].

Where the theory ends and the statistics begins. The pendulum shows the division of labour. Theory relates the quantity we want to quantities we can observe: a pendulum of length L swings with period $T = 2\pi\sqrt{L/g}$, so $g = 4\pi^2 L/T^2$. That relation is physics, and it is exact. If every timing were exact too, one swing would give g and nothing would be left to do. But each timing comes with its own fluctuation (a reaction time, a draught, a stopwatch tick), so ten swings give ten different values of g, and the relation alone cannot say which one to believe. Statistics begins at exactly this point. Its questions are how to combine the noisy readings into one number, the choice of the estimator (3.6), and how far that number can be from the truth, the sampling distribution. Every estimator in this chapter has the same two parts: a physical relation between θ and the observables, which theory supplies, and a way of dealing with the noise on those observables, which is what the chapter is about.

The readings are independent and identically distributed (iid): each comes from the same $f(x;\theta)$, and none influences another. Independence means the joint density of the whole list is the product of the single densities,

$$f_{\text{sample}}(x_1,\ldots,x_N;\theta) = f(x_1;\theta)\,f(x_2;\theta)\cdots f(x_N;\theta) = \prod_{i=1}^{N} f(x_i;\theta), \tag{3.7}$$

so the entire experiment is a single draw of one N-dimensional random vector $(X_1,\ldots,X_N)$ [Cowan (5.1)].

3.1.3 The sampling distribution and the standard error

An estimator is a random variable, so it has a distribution. The width of that distribution is the error bar.

Working definition 3.3 (sampling distribution, standard error)

The **sampling distribution** of $\hat{\theta}_N$ is its probability distribution over imagined repetitions of the whole experiment [each repetition draws a fresh sample of size N from the same $f(x;\theta)$, with θ held fixed]. Its standard deviation is the **standard error**,

$$\text{se}(\hat{\theta}_N) = \sqrt{\text{Var}_\theta(\hat{\theta}_N)}. \tag{3.8}$$

Usually se depends on the unknown truth. Replacing the truth by estimates gives the **estimated standard error** $\widehat{\text{se}}$, the number we actually quote.

The subscript in $\mathbb{E}_\theta$, Var_θ, P_θ means "computed with the density $f(x;\theta)$ at this one fixed θ". It does *not* mean an average over values of θ [Wasserman p. 89]. Written out, the mean of the sampling distribution can be computed in two ways [Cowan (5.3)]: as a one-dimensional integral over the values of $\hat{\theta}$, or as an N-dimensional integral over all possible data sets:

$$\mathbb{E}_\theta[\hat{\theta}] = \int \hat{\theta}\; g(\hat{\theta};\theta)\,\mathrm{d}\hat{\theta} = \int\cdots\int \hat{\theta}(x_1,\ldots,x_N)\,f(x_1;\theta)\cdots f(x_N;\theta)\,\mathrm{d}x_1\cdots\mathrm{d}x_N, \tag{3.9}$$

where $g(\hat{\theta};\theta)$ is the sampling density. The second form is often the easier one, because it never needs g. It needs only the recipe $\hat{\theta}(x_1,\ldots,x_N)$ and the model density f.

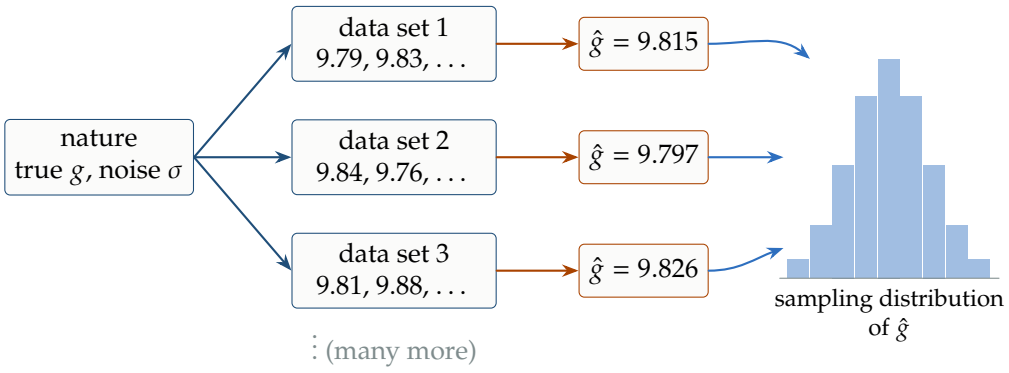


Figure 3.3: Imagined repetitions. Nature, with its true value held fixed, produces a fresh data set on each run of the experiment. The same recipe turns each data set into an estimate, and the estimates pile up into a histogram, the sampling distribution. We observe only one data set, but the error bar is the width of this histogram.

Example: The fraction of successes. Let $X_1, \dots, X_N$ be Bernoulli(p): $X_i = 1$ if, say, the i-th muon decays inside the fiducial volume. Take $\hat{p} = N^{-1}\sum_i X_i$. We need $\mathbb{E}\,X_i = 1\cdot p + 0\cdot(1-p) = p$ and $\operatorname{Var} X_i = \mathbb{E}\,X_i^2 - p^2 = p - p^2 = p(1-p)$ (because $X_i^2 = X_i$). Then

$$\mathbb{E}_p(\hat{p}) \overset{(1)}{=} \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_p(X_i) \overset{(2)}{=} \frac{1}{N}\,Np = p,$$

$$\operatorname{Var}_p(\hat{p}) \overset{(3)}{=} \frac{1}{N^2}\operatorname{Var}_p\Big(\sum_{i=1}^{N} X_i\Big) \overset{(4)}{=} \frac{1}{N^2}\sum_{i=1}^{N}\operatorname{Var}_p(X_i) \overset{(5)}{=} \frac{p(1-p)}{N}.$$

(1) Expectation is linear: $\mathbb{E}(aU + bV) = a\,\mathbb{E}\,U + b\,\mathbb{E}\,V$, for any random variables.
(2) Each of the N terms equals p.
(3) $\operatorname{Var}(cU) = c^2 \operatorname{Var} U$ for a constant c.
(4) For independent variables the covariances vanish, so the variance of a sum is the sum of the variances (derived in Eq. (3.20) below).
(5) Each term equals $p(1-p)$.

So $\hat{p}$ is centred on the truth, and $\mathrm{se} = \sqrt{p(1-p)/N}$. We do not know p, so we quote $\widehat{\mathrm{se}} = \sqrt{\hat{p}(1-\hat{p})/N}$ [Wasserman p. 91, Ex. 6.8]. With $N = 400$ and $\hat{p} = 0.30$ this is $\sqrt{0.3\times 0.7/400} = 0.023$.

Example: Estimating π with darts. Throw darts uniformly at a square of side L with an inscribed circle. A dart lands in the circle with probability $p = \pi(L/2)^2/L^2 = \pi/4$ [SeeingTheory pp. 16–17]. With $X_i = 1$ for a hit, $\hat{\pi} = 4\hat{p}$ is an estimator of π. From Section 3.1.3,

$$\mathbb{E}(\hat{\pi}) \overset{(1)}{=} 4\;\mathbb{E}(\hat{p}) = 4\cdot\frac{\pi}{4} = \pi,$$

$$\mathrm{se}(\hat{\pi}) \overset{(2)}{=} 4\sqrt{\frac{p(1-p)}{n}} \overset{(3)}{=} 4\sqrt{\frac{(\pi/4)(1-\pi/4)}{n}} = \frac{1.642}{\sqrt{n}}.$$

(1) Linearity, and $\mathbb{E}\,\hat{p} = p = \pi/4$.
(2) $\operatorname{Var}(4\hat{p}) = 16\operatorname{Var}\hat{p}$, so the standard deviation is multiplied by 4.
(3) Insert $p = \pi/4$: $(\pi/4)(1-\pi/4) = 0.7854\times 0.2146 = 0.1685$, and $4\sqrt{0.1685} = 1.642$.

With $n = 100$ darts the simulation in Section 3.1.4 gives a mean of 3.142 and a spread of 0.1642 over 200 000 repetitions, against the derived 0.1642. A hundred darts pin π down only to about ± 0.16. This is our first sighting of the Monte Carlo error $\propto 1/\sqrt{n}$ of Chapter 6.

Example: The muon lifetime from twenty decays. Decay times of a muon at rest are exponential, $f(t;\tau) = \tau^{-1}e^{-t/\tau}$, with $\mathbb{E}\,t = \tau$ and $\operatorname{Var} t = \tau^2$ (Chapter 2). The estimator $\hat{\tau} = N^{-1}\sum_i t_i$ has, exactly as in Section 3.1.3, $\mathbb{E}\,\hat{\tau} = \tau$ and $\operatorname{Var}\hat{\tau} = \tau^2/N$. But here we can do better than the mean and variance: we know the whole sampling distribution. A sum of N independent exponentials of mean τ is a Gamma (Erlang) variable with shape N and scale τ (Chapter 2). Dividing by N divides the scale by N:

$$g(\hat{\tau};\tau) = \frac{\hat{\tau}^{\,N-1}e^{-N\hat{\tau}/\tau}}{\Gamma(N)\,(\tau/N)^N}, \qquad \hat{\tau} > 0. \tag{3.10}$$

This density is skewed to the right: its skewness is $2/\sqrt{N}$, i.e. 0.4472 for $N = 20$ (the simulation gives 0.4475). So the error bar $\tau/\sqrt{N}$ does not tell the whole story, because a skewed distribution treats the two sides differently. For $N = 20$, integrating (3.10) shows that $\hat{\tau}$ falls below the true τ slightly more often than above it (53% against 47%), but large upward excursions are the more common ones: $\hat{\tau}$ exceeds τ by two standard errors with probability 3.3%, and falls short by two standard errors with probability only 1.0%.

3.1.4 Seeing it in Python

Listing 3.1: Four recipes applied to the same simulated experiments (code/ch03/01_sampling_distribution.py)

```
rng = rng_for("ch03", "01_sampling_distribution")
setup()
M = 200_000                       # number of imagined repetitions of the experiment

# ---------- (A) ten Gaussian readings of g ----------
g_true, sigma, N = 9.81, 0.05, 10
x = g_true + sigma * rng.standard_normal((M, N))      # each row = one experiment
est = {
    "sample mean": x.mean(axis=1),
    "sample median": np.median(x, axis=1),
    "mid-range": 0.5 * (x.min(axis=1) + x.max(axis=1)),
    "first reading": x[:, 0],
}
sd = {k: v.std(ddof=1) for k, v in est.items()}
```

The script creates a matrix `x` of shape $(M, N) = (200\,000, 10)$: each row is one complete pendulum experiment with true $g = 9.81$ and $\sigma = 0.05\,\mathrm{m\,s^{-2}}$. The dictionary `est` applies four different recipes to every row: the sample mean, the sample median, the mid-range $(\min + \max)/2$, and the lazy "first reading only". The last line measures the spread of each recipe over the M repetitions. That spread *is* the standard error, obtained without any formula.

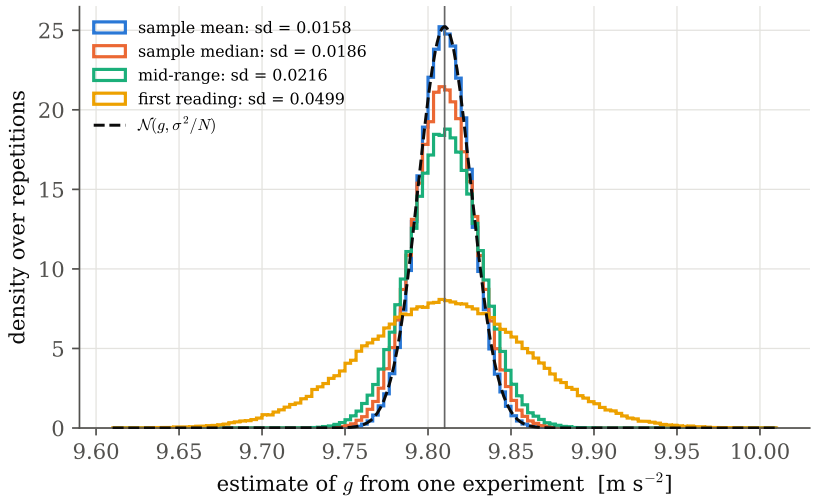


Figure 3.4: Sampling distributions of four estimators of g, each built from the same 200 000 simulated ten-reading experiments. All four are centred on the true value (vertical line). The sample mean is the narrowest and sits on the derived curve $\mathcal{N}(g, \sigma^2/N)$. Using only the first reading gives a distribution as wide as the noise itself.

The numbers: the sample mean scatters by 0.01583 against the derived $\sigma/\sqrt{N} = 0.01581$; the median by 0.01859 (its variance is 1.379 times larger, tending to $\pi/2 \approx 1.57$ for large N with Gaussian noise); the mid-range by 0.02155; the first reading by $0.04993 \approx \sigma$. All four recipes are "right on average". They differ in how much they scatter, and that is what makes one of them better. Section 3.2 makes "better" precise.

The simulated mean and spread of $\hat{\tau}$ are 2.198 and 0.4914 μs, against the derived $\tau = 2.197$ and $\tau/\sqrt{20} = 0.491$ μs.

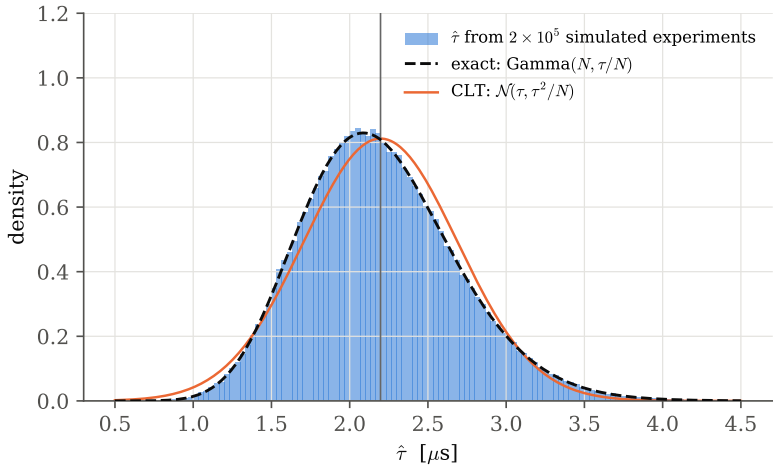


Figure 3.5: Sampling distribution of the lifetime estimate from twenty decays. The histogram of simulated $\hat\tau$ follows the exact Gamma density (3.10) (dashed), including its longer right tail. The Gaussian with the same mean and width (solid) is close but visibly too symmetric.

What this bought us. An estimator is a recipe, so it has a sampling distribution. The standard error is the width of that distribution, not the width of the data: with ten readings of noise σ the sample mean scatters by $\sigma/\sqrt{10}$, a single reading by σ.

word	what it is	pendulum example
parameter θ	fixed, unknown number labelling the true distribution	the true g
statistic	any function of the data with no unknowns	$\max_i x_i$, $\sum x_i^2$
estimator $\hat\theta$	a statistic used to guess θ; a random variable	$\bar X = N^{-1}\sum X_i$
estimate	the estimator evaluated on our data; one number	9.815
standard error	width of the estimator's sampling distribution	$\sigma/\sqrt{N} = 0.016$

Watch out: What is random and what is fixed

In the frequentist picture, θ is fixed and $\hat\theta$ is random. A statement such as "$g = 9.815\pm0.016$" is shorthand for a statement about the *recipe*: applied to repeated experiments, it lands within 0.016 of the truth about two times in three. It is not a probability distribution for g. Probability distributions *for* parameters are the Bayesian answer of Chapter 5, and the precise meaning of an interval is the subject of Chapter 4.

A first look at intervals and tests. The sampling distribution also gives the two other basic tools of statistics, intervals and tests. Suppose $\hat\theta_N$ is approximately Gaussian with estimated standard error $\widehat{\mathrm{se}}$ (the central limit theorem of Section 3.3 explains why this is so common). Then the interval $C_N = (\hat\theta_N - z_{\alpha/2}\widehat{\mathrm{se}},\ \hat\theta_N + z_{\alpha/2}\widehat{\mathrm{se}})$ with $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$ (Φ is the standard normal distribution function) traps the fixed θ in a fraction $1-\alpha$ of repetitions as $N \to \infty$ [Wasserman p. 94, Thm 6.16]. For $\alpha = 0.05$, $z_{\alpha/2} = 1.96 \approx 2$: "estimate ± two standard errors". For the Bernoulli case the interval is $\hat p \pm z_{\alpha/2}\sqrt{\hat p(1-\hat p)/N}$ [Wasserman Ex. 6.17].

A **hypothesis test** uses the same sampling distribution the other way round. We start from a default hypothesis, such as "the coin is fair", $p = 1/2$. Under it we know how $T = |\hat p - 1/2|$ is distributed. If the observed T is larger than the default would usually produce, we reject the default [Wasserman p. 95, Ex. 6.18]. Both ideas are developed properly in Chapter 4.

Verification and research. Here the Monte Carlo *verifies* a result we can derive, $\mathrm{se}(\bar{X}) = \sigma/\sqrt{N}$. The same few lines, "simulate the experiment many times, apply the recipe, histogram the output", are the *research instrument* when no formula exists. The sampling distribution of a CMB power spectrum estimated from a masked, noisy, beam-smoothed map (Chapter 8), or of a topological summary of a field such as its Euler characteristic (Chapter 12), is obtained in exactly this way, with a simulator of the sky in place of `rng.standard_normal`.

3.2 Aiming and scattering: bias, variance and mean squared error

Charged particles cross a silicon strip detector and leave hits whose positions along one strip are uniformly spread between 0 and the strip's unknown active length θ. We record $n = 10$ hit positions and consider three ways to infer the length. Since the mean position for a uniform distribution lies halfway along the strip, one estimate doubles the sample average. Another uses the largest observed position, which is a lower bound on the length. The third adjusts this maximum upwards to allow for the gap between the last of ten hits and the end of the strip.

Which recipe should we use? Each one is a guess at θ, and each guess misses by some amount that changes from experiment to experiment. The fair comparison is the average squared miss, the mean squared error. It splits into two parts every experimentalist already knows: a systematic offset and a statistical scatter.

Darts at a target. Picture the sampling distribution of Section 3.1 as a cluster of darts around a bull's-eye at the true value. Two things can go wrong. The cluster can be centred in the wrong place (a thrower with a consistent pull to the left), or it can be wide (a shaky hand). The first is *bias*, the second is *variance*. A good estimator needs both to be small, and it may be worth accepting a little of one to remove a lot of the other.

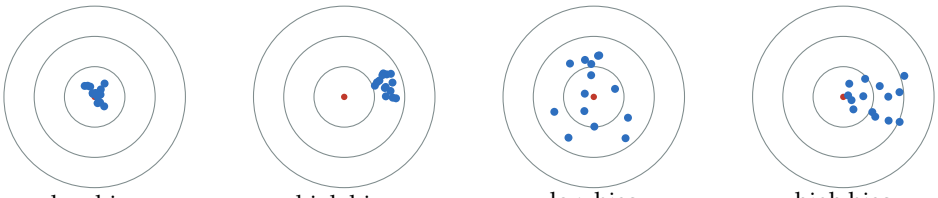


Figure 3.6: Four throwers, fourteen darts each. The red dot is the true value of the parameter and every blue dot is the estimate from one repetition of the experiment. Moving the whole cluster off the centre is bias. Spreading it out is variance. The mean squared error is the average squared distance of the darts from the red dot, and it grows with both.

The questions, in words.

1. Over many repetitions, where is the recipe's output centred? Is that the true value? (Bias.)
2. How widely does the output scatter about its own centre? (Variance, or its square root, the standard error.)
3. On average, how far is the output from the *truth*, counting both effects? (Mean squared error.)
4. As we take more data, does the output settle down on the truth? (Consistency.)

3.2.1 Bias, variance and their sum

Two things can be wrong with an estimator: its centre can be off, and its spread can be large. Each has a name, and the two combine into one measure of total error.

Working definition 3.4 (bias, unbiased, mean squared error)

The **bias** of an estimator $\hat{\theta}_n$ is

$$\text{bias}(\hat{\theta}_n) = \mathbb{E}_\theta(\hat{\theta}_n) - \theta \tag{3.11}$$

[the offset of the centre of the sampling distribution from the truth, at this fixed θ]. The estimator is **unbiased** if the bias is zero for *every* θ in Θ. The **mean squared error** is

$$\text{MSE}(\hat{\theta}_n) = \mathbb{E}_\theta\left[(\hat{\theta}_n - \theta)^2\right] \tag{3.12}$$

[the average over repetitions of the squared miss distance].

Both are functions of the unknown θ, because the average is taken with $f(x;\theta)$. The two ideas fit together exactly: the MSE is the variance plus the squared bias. To show it, call the centre of the sampling distribution $\bar{\theta}_n := \mathbb{E}_\theta(\hat{\theta}_n)$. It is a fixed number, not a random variable. Then

$$\begin{aligned}
\text{MSE}(\hat{\theta}_n) &\overset{(1)}{=} \mathbb{E}_\theta\left[(\hat{\theta}_n - \bar{\theta}_n + \bar{\theta}_n - \theta)^2\right] \\
&\overset{(2)}{=} \mathbb{E}_\theta\left[(\hat{\theta}_n - \bar{\theta}_n)^2\right] + 2(\bar{\theta}_n - \theta)\,\mathbb{E}_\theta\left[\hat{\theta}_n - \bar{\theta}_n\right] + (\bar{\theta}_n - \theta)^2 \\
&\overset{(3)}{=} \mathbb{E}_\theta\left[(\hat{\theta}_n - \bar{\theta}_n)^2\right] + (\bar{\theta}_n - \theta)^2 \\
&\overset{(4)}{=} \text{Var}_\theta(\hat{\theta}_n) + \text{bias}^2(\hat{\theta}_n).
\end{aligned} \tag{3.13}$$

(1) Add and subtract the constant $\bar{\theta}_n$ inside the square.
(2) Expand $(a+b)^2 = a^2 + 2ab + b^2$ with $a = \hat{\theta}_n - \bar{\theta}_n$ (random) and $b = \bar{\theta}_n - \theta$ (constant). Constants come out of the expectation, and $\mathbb{E}(b^2) = b^2$.
(3) $\mathbb{E}_\theta[\hat{\theta}_n - \bar{\theta}_n] = \bar{\theta}_n - \bar{\theta}_n = 0$ by the definition of $\bar{\theta}_n$, so the cross term vanishes.
(4) The first term is the variance by definition. The second is the squared bias (3.11).

Key idea: MSE = bias^2 + variance

The mean squared error of any estimator is the square of its systematic offset plus the square of its statistical scatter, MSE = bias^2 + se^2. For an unbiased estimator the MSE is just the variance.

Why is the result true, beyond the algebra? It is Pythagoras. The miss $\hat{\theta}_n - \theta$ is a fixed offset b (the bias) plus a random part a whose average is zero. Squaring gives $a^2 + 2ab + b^2$. The cross term averages to zero, because b is a constant and a averages to zero. So the two parts add in square, like the two perpendicular sides of a right triangle. The same identity is the parallel-axis theorem of mechanics: the moment of inertia about the true value equals the moment about the centre of mass plus the total mass times the squared distance between the two.

Statistics and the laboratory.

statistics	laboratory
bias	systematic error (a miscalibrated scale)
standard error	statistical error (shrinks with more data)
$\text{MSE} = \text{bias}^2 + \text{se}^2$	"add the errors in quadrature"
unbiased estimator	a method with no systematic offset, for every true value
consistent estimator	a method that homes in on the truth as data accumulate

The variance of the estimator is the *statistical* error. A bias, or a mistake in the model, is a *systematic* error, and more data do not remove it [Cowan §5.1, (5.5)]. One caution about the quadrature row. Physicists add a systematic in quadrature because they treat its unknown size as random, drawn from an ensemble of possible miscalibrations. The bias in (3.13) is not random. It is a definite, if unknown, number, and it enters squared for a purely algebraic reason: the cross term vanished.

3.2.2 Three recipes for the strip length

Back to the silicon strip from the start of this section. A strip of unknown active length θ records $n = 10$ hits, and each hit position X_i is equally likely to be anywhere along it, $X_i \sim \text{Uniform}(0, \theta)$, with density $1/\theta$ on $[0, \theta]$. The three colleagues each proposed a way of turning the ten positions into one number for θ, that is, an estimator $\hat{\theta} = g(X_1, \dots, X_n)$ in the sense of Section 3.1.2. It is customary to call a candidate estimator T (for a statistic of the data), and we number the three candidates in the order the colleagues proposed them. Two ingredients appear in them: the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, and the largest of the ten positions, written $X_{(n)} = \max_i X_i$ (the subscript in parentheses means "the n-th in order of size"). The moments of a single hit we need are (from Chapter 2)

$$\mathbb{E}\, X_i = \frac{\theta}{2}, \qquad \mathbb{E}\, X_i^2 = \int_0^\theta \frac{x^2}{\theta}\,\mathrm{d}x = \frac{\theta^2}{3}, \qquad \text{Var}\, X_i = \frac{\theta^2}{3} - \frac{\theta^2}{4} = \frac{\theta^2}{12}. \tag{3.14}$$

The three recipes, written as formulas, are

$T_1 = 2\bar{X}$	double the average position (a uniform hit sits at $\theta/2$ on average),
$T_2 = X_{(n)}$	take the largest position (no hit can lie beyond the end),
$T_3 = \frac{n+1}{n} X_{(n)}$	nudge the largest position up (it almost never reaches the end).

For each recipe we ask the questions listed above: where is it centred, how widely does it scatter, and what is its total error?

Example: Doubling the mean.

$$\mathbb{E}\, T_1 \overset{(1)}{=} 2 \cdot \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\, X_i \overset{(2)}{=} 2 \cdot \frac{\theta}{2} = \theta,$$

$$\text{Var}\, T_1 \overset{(3)}{=} 4\ \text{Var}\, \bar{X} \overset{(4)}{=} 4 \cdot \frac{\theta^2}{12\, n} = \frac{\theta^2}{3n}.$$

(1) Linearity of expectation.
(2) Each $\mathbb{E}\, X_i = \theta/2$ by (3.14).
(3) $\text{Var}(cU) = c^2\, \text{Var}\, U$ with $c = 2$.
(4) $\text{Var}\, \bar{X} = \text{Var}\, X_i / n$ for independent readings (the same steps as in Section 3.1.3, and derived generally in (3.20)).

So T_1 is unbiased and $\text{MSE}(T_1) = \theta^2/(3n)$. For $n = 10$ and $\theta = 1$ this is 0.0333.

Example: The largest hit. We first need the sampling distribution of the maximum. The maximum is below t exactly when *every* hit is below t. For $0 \le t \le \theta$,

$$\mathrm{P}(X_{(n)} \le t) \overset{(1)}{=} \mathrm{P}(X_1 \le t,\ X_2 \le t,\ \ldots,\ X_n \le t) \overset{(2)}{=} \prod_{i=1}^{n} \mathrm{P}(X_i \le t) \overset{(3)}{=} \left(\frac{t}{\theta}\right)^n.$$

(1) "The largest is at most t" is the same event as "all are at most t".
(2) The hits are independent, so the joint probability factorises.
(3) For a uniform variable $\mathrm{P}(X_i \le t) = t/\theta$.

Differentiating, the density is $g(t) = n\,t^{n-1}/\theta^n$ on $[0, \theta]$. It piles up near the end of the strip, as the picture of ten hits suggests. Its moments are

$$\mathbb{E}\,X_{(n)} \overset{(1)}{=} \int_0^\theta t \cdot \frac{n\,t^{n-1}}{\theta^n}\mathrm{d}t \overset{(2)}{=} \frac{n}{\theta^n}\,\frac{\theta^{n+1}}{n+1} = \frac{n}{n+1}\,\theta,$$

$$\mathbb{E}\,X_{(n)}^2 \overset{(3)}{=} \frac{n}{\theta^n}\,\frac{\theta^{n+2}}{n+2} = \frac{n}{n+2}\,\theta^2,$$

$$\mathrm{Var}\,X_{(n)} \overset{(4)}{=} \frac{n}{n+2}\,\theta^2 - \frac{n^2}{(n+1)^2}\,\theta^2 \overset{(5)}{=} \frac{n\,\theta^2}{(n+1)^2(n+2)}.$$

(1) Definition of the mean of a continuous variable with density g.
(2) $\int_0^\theta t^n \mathrm{d}t = \theta^{n+1}/(n+1)$.
(3) The same integral with one more power of t: $\int_0^\theta t^{n+1}\mathrm{d}t = \theta^{n+2}/(n+2)$.
(4) Var $= \mathbb{E}\,X^2 - (\mathbb{E}\,X)^2$.
(5) Common denominator: $n(n+1)^2 - n^2(n+2) = n\big[(n+1)^2 - n(n+2)\big] = n$.

So $\mathrm{bias}(T_2) = \frac{n}{n+1}\theta - \theta = -\frac{\theta}{n+1}$: the largest hit always falls short of the end. By (3.13),

$$\mathrm{MSE}(T_2) \overset{(1)}{=} \frac{\theta^2}{(n+1)^2} + \frac{n\,\theta^2}{(n+1)^2(n+2)} \overset{(2)}{=} \frac{\theta^2\left[(n+2)+n\right]}{(n+1)^2(n+2)} = \frac{2\,\theta^2}{(n+1)(n+2)}.$$

(1) Squared bias plus variance.
(2) Put both terms over the denominator $(n+1)^2(n+2)$. Then $2n+2 = 2(n+1)$ cancels one factor of $n+1$.

For $n = 10$: bias $= -1/11 = -0.09091$ and MSE $= 2/132 = 0.01515$ (with $\theta = 1$). Even though it is biased, T_2 has less than half the MSE of the unbiased T_1.

Example: Correcting the bias. The bias of T_2 is known exactly, as a fraction of θ, so we can divide it out:

$$\mathbb{E}\,T_3 \overset{(1)}{=} \frac{n+1}{n}\,\mathbb{E}\,X_{(n)} \overset{(2)}{=} \frac{n+1}{n}\cdot\frac{n}{n+1}\,\theta = \theta,$$

$$\mathrm{Var}\,T_3 \overset{(3)}{=} \frac{(n+1)^2}{n^2}\cdot\frac{n\,\theta^2}{(n+1)^2(n+2)} = \frac{\theta^2}{n(n+2)}.$$

(1) Linearity, with the constant $(n+1)/n$.
(2) The mean of the maximum from Section 3.2.2.
(3) $\mathrm{Var}(cU) = c^2\,\mathrm{Var}\,U$ with $c = (n+1)/n$, and the variance of the maximum.

Now $\mathrm{MSE}(T_3) = \theta^2/(n(n+2)) = 1/120 = 0.008333$ for $n = 10$: unbiased *and* four times better than T_1. For large n, $\mathrm{MSE}(T_1) \propto 1/n$ while $\mathrm{MSE}(T_3) \propto 1/n^2$. The maximum uses the one feature of the data (the sharp edge) that carries the most information about θ. The mean washes that edge out.

Seeing it in Python. Listing 3.2 simulates $M = 200\,000$ experiments of ten hits and applies all three recipes to each.

Listing 3.2: The three recipes and their derived bias and variance (code/ch03/02_bias_variance_uniform.py)

```
def estimators(x):
    n = x.shape[1]
    mx = x.max(axis=1)
    return {"T1": 2 * x.mean(axis=1), "T2": mx, "T3": (n + 1) / n * mx}


def theory(n):
    """bias, variance, MSE of T1, T2, T3 (derived in the text), theta = 1."""
    return {
        "T1": (0.0, theta**2 / (3 * n)),
        "T2": (-theta / (n + 1), n * theta**2 / ((n + 1) ** 2 * (n + 2))),
        "T3": (0.0, theta**2 / (n * (n + 2))),
    }
```

The function `estimators` takes a matrix whose rows are experiments and returns the three estimates for each row. The function `theory` encodes the formulas just derived, so the script can put the two side by side. The simulated biases are 2.896×10^{-4}, -0.09081 and 1.080×10^{-4} (derived: 0, -0.09091, 0). The simulated MSEs are 0.03342, 0.01517 and 0.008373 (derived: 0.03333, 0.01515, 0.008333). Agreement to the third digit is what $M = 200\,000$ repetitions buy.

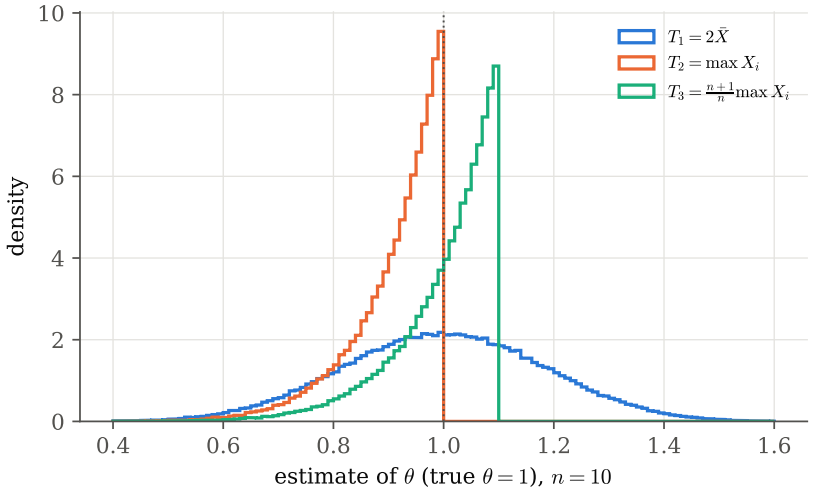


Figure 3.7: Sampling distributions of the three strip-length recipes with ten hits each. Doubling the mean gives a broad, symmetric bump centred on the truth. The largest hit gives a sharp distribution that ends exactly at the truth and so sits entirely to its left. Stretching it by 11/10 moves it onto the truth and keeps it narrow.

3.2.3 The MSE depends on the truth

For the strip, T_3 beat T_1 at every θ. That is unusual. The MSE is a function of the unknown truth, so one recipe can win for some true values and lose for others. The simplest case shows how.

Example: Shrinking the mean towards zero. Let $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ and consider the family of recipes $T_c = c\,\bar{X}$ with a constant $0 < c \le 1$. Then

$$\mathrm{MSE}(T_c) \overset{(1)}{=} \left(\mathbb{E}\,T_c - \mu\right)^2 + \mathrm{Var}\,T_c \overset{(2)}{=} (c-1)^2\mu^2 + c^2\,\frac{\sigma^2}{n}.$$

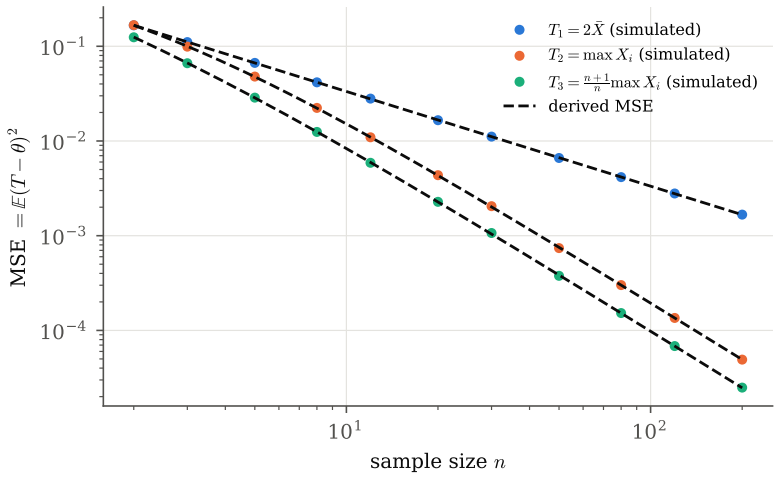


Figure 3.8: Mean squared error against the number of hits, on log–log axes. The points are simulations and the dashed curves the derived formulas. Doubling the mean falls with slope -1. Both recipes built on the largest hit fall with slope -2, and the corrected one stays below the uncorrected one at every n.

(1) The decomposition (3.13).
(2) $\mathbb{E}\,T_c = c\mu$ and $\operatorname{Var} T_c = c^2 \operatorname{Var} \bar{X} = c^2\sigma^2/n$.

Setting the derivative with respect to c to zero, $2(c-1)\mu^2 + 2c\,\sigma^2/n = 0$, gives the best constant $c^* = \mu^2/(\mu^2 + \sigma^2/n)$. It depends on the unknown μ, so it is not a recipe we can use. Fix $c < 1$ instead and compare with $\bar{X}$, which is T_1 here and has MSE σ^2/n. Shrinking lowers the variance term by the factor c^2 but adds the bias term $(c-1)^2\mu^2$, which grows with μ. So shrinking wins when μ is near zero and loses when μ is far from zero (Figure 3.9). With $\sigma^2/n = 1$ and $c = 0.8$ the curves cross where $0.04\,\mu^2 + 0.64 = 1$, i.e. at $|\mu| = 3$.

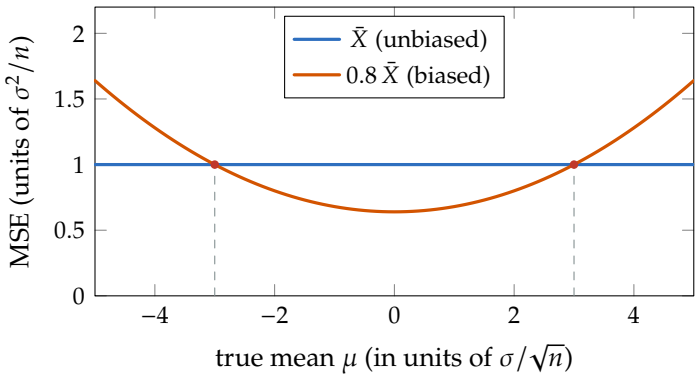


Figure 3.9: Mean squared error of two recipes for a Gaussian mean, as a function of the true value. The sample mean has the same MSE everywhere. The shrunken recipe wins inside $|\mu| < 3$ and loses outside, and the two curves meet at the red dots. Neither is better for every possible truth.

There is usually no "best" estimator. Because the MSE is a function of θ, two estimators are usually not ordered: one wins for some true values and loses for others. The simplest way to get a meaningful comparison is to restrict the class, for instance to unbiased estimators, and look for the smallest variance inside it. That programme leads to the Cramér–Rao bound and the Fisher information (second half of Chapter 3 and Chapter 10).

Watch out: Unbiased is not the same as good

An unbiased estimator can be much worse than a biased one. T_1 in the strip example is unbiased and has four times the MSE of T_3 and twice that of the biased T_2. Conversely, a biased estimator is not "wrong": Wasserman notes that unbiasedness "used to receive much attention but these days is considered less important" [Wasserman p. 90]. What matters is the size of the total error, and that it shrinks as data accumulate.

Example: Poisson counts. A counter records $X_1, \dots, X_n \sim \text{Poisson}(\lambda)$ events in n equal time bins, and we estimate the rate by $\hat{\lambda} = \bar{X}$. From Chapter 2, $\mathbb{E}\,X_i = \text{Var}\,X_i = \lambda$. Then

$$\text{bias}(\hat{\lambda}) \overset{(1)}{=} \frac{1}{n}\sum_i \mathbb{E}\,X_i - \lambda = 0,$$

$$\text{se}(\hat{\lambda}) \overset{(2)}{=} \sqrt{\frac{\text{Var}\,X_i}{n}} = \sqrt{\frac{\lambda}{n}},$$

$$\text{MSE}(\hat{\lambda}) \overset{(3)}{=} 0^2 + \frac{\lambda}{n} = \frac{\lambda}{n}.$$

(1) Linearity and $\mathbb{E}\,X_i = \lambda$.
(2) Variance of a mean of independent readings, and $\text{Var}\,X_i = \lambda$ for a Poisson variable.
(3) Decomposition (3.13) with zero bias.

The estimated standard error is $\sqrt{\hat{\lambda}/n}$. For $n = 1$ this is the familiar "$\sqrt{N}$ error on N counts" [Wasserman Exercise 6.1].

3.2.4 Consistency: does the recipe home in on the truth?

A minimal demand on any recipe is that more data make it better, eventually without limit. To say this precisely we must say what it means for a sequence of random numbers to approach a constant. One random number can still land anywhere, so "approach" has to be a statement about probabilities. The simplest version is below (the other versions are in Section 3.3.2).

Convergence in probability and consistency. A sequence of random variables Y_n **converges in probability** to a constant c, written $Y_n \xrightarrow{P} c$, if for every $\epsilon > 0$

$$\text{P}\big(|Y_n - c| > \epsilon\big) \longrightarrow 0 \qquad (n \to \infty) \tag{3.15}$$

[any band around c, however narrow, eventually contains Y_n in almost all repetitions]. An estimator is **consistent** if $\hat{\theta}_n \xrightarrow{P} \theta$ for every $\theta \in \Theta$.

To prove consistency we need a bound on the probability of a large miss. One elementary inequality supplies it.

Markov and Chebyshev inequalities. If $Y \geq 0$ and $t > 0$, then $\text{P}(Y \geq t) \leq \mathbb{E}\,Y/t$ (Markov). Applied to $Y = (U - a)^2$ with $t = \epsilon^2$ it gives, for any random variable U and any constant a,

$$\text{P}\big(|U - a| \geq \epsilon\big) \leq \frac{\mathbb{E}\big[(U-a)^2\big]}{\epsilon^2}. \tag{3.16}$$

With $a = \mathbb{E}\,U$ the numerator is $\text{Var}\,U$ (Chebyshev).

Markov's inequality is true for a reason we can see: a non-negative quantity cannot have a big chunk of probability at large values without dragging its average up. Written out, with $\mathbb{1}\{\cdot\}$ the indicator that is 1 when the condition holds and 0 otherwise,

$$\mathbb{E}\,Y \overset{(1)}{\geq} \mathbb{E}\big[Y\,\mathbb{1}\{Y \geq t\}\big] \overset{(2)}{\geq} \mathbb{E}\big[t\,\mathbb{1}\{Y \geq t\}\big] \overset{(3)}{=} t\;\mathrm{P}(Y \geq t). \tag{3.17}$$

(1) $Y \geq 0$, so throwing away the part of the average where $Y < t$ cannot increase it.
(2) Where the indicator is 1, $Y \geq t$. Where it is 0, both sides are 0.
(3) The average of an indicator is the probability of its event.

Dividing by t gives Markov. Now take $U = \hat{\theta}_n$ and $a = \theta$ in (3.16): the numerator is exactly the MSE, so

$$\mathrm{P}\big(|\hat{\theta}_n - \theta| \geq \epsilon\big) \leq \frac{\mathrm{MSE}(\hat{\theta}_n)}{\epsilon^2} = \frac{\mathrm{bias}^2(\hat{\theta}_n) + \mathrm{se}^2(\hat{\theta}_n)}{\epsilon^2}. \tag{3.18}$$

A sufficient condition for consistency. If $\mathrm{bias}(\hat{\theta}_n) \to 0$ and $\mathrm{se}(\hat{\theta}_n) \to 0$ as $n \to \infty$, then $\hat{\theta}_n$ is consistent. The reason is (3.18): its right side then goes to zero for every fixed ϵ, and it bounds the probability of a miss larger than ϵ.

Example: Which recipes are consistent?

- $\hat{p} = \bar{X}$ for coin flips: bias 0, se $= \sqrt{p(1-p)/n} \to 0$. Consistent [Wasserman Ex. 6.11].
- $T_2 = X_{(n)}$ for the strip: bias $-\theta/(n+1) \to 0$, variance $\sim \theta^2/n^2 \to 0$. *Biased but consistent.*
- "Use the first reading only" for the pendulum: bias 0 but se $= \sigma$ for every n. The inequality tells us nothing, and indeed $\mathrm{P}(|X_1 - g| > \epsilon)$ is the same fixed number for all n. *Unbiased but not consistent.*

Bias and consistency are therefore independent properties. One is about the centre of the sampling distribution at fixed n, the other about the whole distribution as $n \to \infty$.

Try it. Show that $T_c = c_n \bar{X}$ with $c_n = n/(n+1)$ is biased for every finite n and yet consistent for a Gaussian mean. (Compute its bias and variance and let $n \to \infty$.)

What this bought us. Choose recipes by their total error, MSE = bias^2 + se^2. Bias alone is a poor guide. A recipe whose bias and standard error both vanish as the data grow is consistent, by Chebyshev's inequality.

Verification and research. For the strip we could derive every number, and the simulation only checked the algebra. In real analyses the recipe is often a long chain (a fit inside a selection inside a reconstruction) and the bias has no formula. The standard remedy is the same simulation used as an instrument: inject a known truth into simulated data, run the full chain many times, and measure the bias and scatter of the output. This "injection–recovery" test is how the bias of a CMB power spectrum estimator on a masked sky is found and removed (Chapter 8).

3.3 Why error bars shrink like $1/\sqrt{N}$: the law of large numbers and the central limit theorem

Every accelerator physicist knows the rule of thumb: to halve the error on a cross-section, collect four times the luminosity. A Monte Carlo integration that gives two correct digits with 10^4 samples needs 10^6 for three. The rule has exceptions that bite. The average of energies drawn from a Breit–Wigner resonance does not improve at all as we add events.

Two theorems account for the rule and for its exceptions. The law of large numbers says that an average settles on the true mean. The central limit theorem says how it is distributed on the way: as a Gaussian of width $\sigma/\sqrt{N}$, provided the readings have a finite variance σ^2. The Breit–Wigner has none, and that is why it escapes the rule.

A random walk, rescaled. The sum of N independent readings is a random walk of N steps. Its distance from the expected position grows like $\sqrt{N}$, the familiar result from diffusion ($\langle x^2\rangle \propto t$). The mean is the sum divided by N, so its scatter is $\sqrt{N}/N = 1/\sqrt{N}$. And the shape of a diffusing cloud forgets the shape of the individual step: after many steps it is a Gaussian, whatever the kick distribution was. That is the central limit theorem, the same statement as the approach of a diffusion kernel to a heat-equation Gaussian.

What we want to know.

1. How does the scatter of an average depend on the number of readings, and on correlations between them?
2. In what sense does a random average "tend to" the true mean? (There are several inequivalent senses.)
3. What does the histogram of averages look like for large N, and how large is large?
4. What breaks, and why, when the parent has no finite variance?

3.3.1 The variance of a sum and of a mean

The $1/\sqrt{N}$ comes from one identity: the variance of a sum is the sum of all the variances and covariances of its terms. Let $X_1, \dots, X_N$ have means μ_i, variances σ_i^2 and covariances $\operatorname{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)]$, with $\operatorname{Cov}(X_i, X_i) = \sigma_i^2$. For the sum,

$$\operatorname{Var}\Big(\sum_{i=1}^{N} X_i\Big) \overset{(1)}{=} \mathbb{E}\Big[\Big(\sum_{i=1}^{N}(X_i-\mu_i)\Big)^2\Big] \overset{(2)}{=} \mathbb{E}\Big[\sum_{i=1}^{N}\sum_{j=1}^{N}(X_i-\mu_i)(X_j-\mu_j)\Big]$$
$$\overset{(3)}{=} \sum_{i=1}^{N}\sum_{j=1}^{N}\operatorname{Cov}(X_i,X_j) \overset{(4)}{=} \sum_{i=1}^{N}\sigma_i^2 + \sum_{i\neq j}\operatorname{Cov}(X_i,X_j). \tag{3.19}$$

(1) By linearity the mean of the sum is $\sum \mu_i$. The variance is the mean squared deviation from it.
(2) The square of a sum is the double sum of all products.
(3) Linearity of expectation, term by term, and the definition of covariance.
(4) Split the double sum into the diagonal $i = j$ and the rest.

If the readings are independent, every covariance with $i \neq j$ vanishes. If they also share one variance σ^2, the mean $\bar{X} = N^{-1}\sum X_i$ has

$$\mathbb{E}\,\bar{X} = \mu, \qquad \operatorname{Var}\bar{X} = \frac{1}{N^2}\,N\sigma^2 = \frac{\sigma^2}{N}, \qquad \operatorname{se}(\bar{X}) = \frac{\sigma}{\sqrt{N}} \tag{3.20}$$

[Wasserman p. 52, Thm. 3.17], [Cowan (5.8), (5.13)]. This is the $1/\sqrt{N}$. It needs nothing about the parent except a finite σ.

Example: Averaging a die. For a fair die, $\mathbb{E}\,X = 3.5$ and $\mathbb{E}\,X^2 = (1{+}4{+}9{+}16{+}25{+}36)/6 = 91/6$, so $\operatorname{Var} X = 91/6 - 49/4 = 35/12 = 2.917$. The average of 50 rolls therefore has $\operatorname{Var}\bar{X} = 2.917/50 = 0.0583$ and se $= 0.24$ [SeeingTheory pp. 38–39]. The individual rolls scatter over 1–6, their average over roughly 3.26–3.74 (two thirds of the time).

Example: Correlated readings do not average away. Suppose every pair of readings has the same correlation ρ, so $\operatorname{Cov}(X_i, X_j) = \rho\sigma^2$ for $i \neq j$. This happens when a common drift, such as a temperature change in the pendulum lab, moves all readings together. There are $N(N-1)$ ordered pairs with $i \neq j$, so from (3.19)

$$\operatorname{Var}\bar{X} \overset{(1)}{=} \frac{1}{N^2}\Big[N\sigma^2 + N(N-1)\rho\sigma^2\Big] \overset{(2)}{=} \frac{\sigma^2}{N}\big[1 + (N-1)\rho\big] \xrightarrow{N\to\infty} \rho\,\sigma^2.$$

(1) Divide the variance of the sum by N^2. Diagonal terms give $N\sigma^2$, off-diagonal ones $N(N-1)\rho\sigma^2$.
(2) Factor out σ^2/N.

With $\rho = 0.1$ the error of the mean can never fall below $\sqrt{0.1}\,\sigma = 0.32\,\sigma$, however many readings we take. The ratio $N_{\text{eff}} = \sigma^2/\operatorname{Var}\bar{X} = N/[1 + (N-1)\rho]$ is the number of *effectively independent* readings. The same quantity decides how useful a correlated Markov chain is (Chapter 9).

3.3.2 What "tends to" means for random variables

For numbers, $a_n \to a$ has one meaning. For random variables there are several. A random variable is not one number: it is a rule that assigns a value to every outcome of the experiment, and that rule has a distribution. A sequence can approach its limit value by value, or only in its distribution, and these are different things. Two simple sequences show the difference [Wasserman p. 71]. If $X_n \sim \mathcal{N}(0,1)$ for every n, independently, the X_n never settle anywhere, yet every one has the same distribution. If $X_n \sim \mathcal{N}(0, 1/n)$, the distributions squeeze onto 0 (Figure 3.10), and we would like to say $X_n \to 0$.

Working definition 3.5 (three modes of convergence)

Let X_n be random variables with distribution functions F_n, and X a random variable with distribution function F.

- $X_n \xrightarrow{P} X$ (**in probability**) if $\mathrm{P}(|X_n - X| > \epsilon) \to 0$ for every $\epsilon > 0$ [the chance of a miss bigger than ϵ dies out].
- $X_n \rightsquigarrow X$ (**in distribution**) if $F_n(t) \to F(t)$ at every t where F is continuous [only the histograms converge, not the values themselves].
- $X_n \xrightarrow{\text{qm}} X$ (**in quadratic mean**) if $\mathbb{E}(X_n - X)^2 \to 0$ [the mean squared miss dies out].

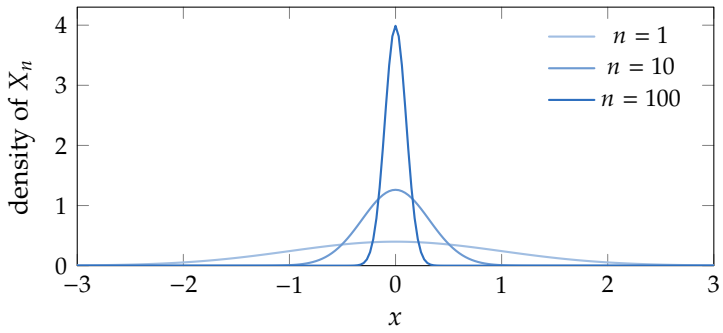


Figure 3.10: The densities of $X_n \sim \mathcal{N}(0, 1/n)$ for $n = 1, 10, 100$. As the width shrinks like $1/\sqrt{n}$, the area remains one and the probability outside any fixed interval containing 0 in its interior tends to zero.

Example: $X_n \sim \mathcal{N}(0, 1/n)$ converges to 0 in every sense. In quadratic mean, $\mathbb{E}\, X_n^2 = \operatorname{Var} X_n = 1/n \to 0$. In probability, by Chebyshev (3.16), $\mathrm{P}(|X_n| > \epsilon) \le 1/(n\epsilon^2) \to 0$. In distribution, write $Z = \sqrt{n} X_n \sim \mathcal{N}(0, 1)$ with distribution function Φ:

$$F_n(t) = \mathrm{P}(X_n \le t) \overset{(1)}{=} \mathrm{P}(\sqrt{n} X_n \le \sqrt{n}\, t) \overset{(2)}{=} \Phi(\sqrt{n}\, t) \xrightarrow{n\to\infty} \begin{cases} 0, & t < 0, \\ 1/2, & t = 0, \\ 1, & t > 0. \end{cases}$$

(1) Multiply both sides of the inequality by $\sqrt{n} > 0$.
(2) $\sqrt{n} X_n$ is standard normal.

The limit $X = 0$ (a point mass) has $F(t) = 0$ for $t < 0$ and $F(t) = 1$ for $t \ge 0$. The two agree everywhere except at $t = 0$, where $F_n(0) = 1/2 \ne F(0) = 1$. That one disagreement is not a failure of X_n to approach 0. At $t = 0$ the limit F jumps from 0 to 1, and each X_n, being symmetric about 0, keeps exactly half its probability below the jump. This is why the definition asks for convergence only where F is continuous: the jump points of F are ignored [Wasserman Ex. 5.3].

How the modes are related. Quadratic mean implies in probability, which implies in distribution. The converse implications fail in general. There is one important partial converse: convergence in distribution *to a constant* implies convergence in probability to that constant [Wasserman Thm. 5.4].

$$X_n \xrightarrow{\text{qm}} X \quad \Longrightarrow \quad X_n \xrightarrow{P} X \quad \Longrightarrow \quad X_n \rightsquigarrow X$$

Figure 3.11: Implications between the modes of convergence. The solid arrows always hold. The dashed red arrows, pointing back, fail in general. The dashed arrow from distribution back to probability does hold in one case: when the limit is a single number, which is the case for consistency.

Why each arrow holds, and why the reverse ones fail:

- *qm* $\Rightarrow$ *P.* Chebyshev with $U = X_n$, $a = X$: $\mathrm{P}(|X_n - X| > \epsilon) \le \mathbb{E}(X_n - X)^2/\epsilon^2 \to 0$. This is the argument of (3.18).
- *P* $\not\Rightarrow$ *qm.* Let $U \sim \text{Uniform}(0, 1)$ and $X_n = \sqrt{n}$ if $U < 1/n$, else 0. Then $\mathrm{P}(|X_n| > \epsilon) = 1/n \to 0$, but $\mathbb{E}\, X_n^2 = n \cdot (1/n) = 1$ for every n. A rare but large value keeps the mean square alive. The same device with $X_n = n^2$ on probability $1/n$ gives $X_n \xrightarrow{P} 0$ while $\mathbb{E}\, X_n = n \to \infty$: convergence in probability says nothing about averages [Wasserman p. 75].

- $P \Rightarrow D$. If X_n is very probably within ϵ of X, then $\mathrm{P}(X_n \le t)$ is squeezed between $\mathrm{P}(X \le t - \epsilon)$ and $\mathrm{P}(X \le t + \epsilon)$ up to a vanishing error. At a continuity point both bounds tend to $F(t)$ as $\epsilon \to 0$.
- $D \not\Rightarrow P$. Let $X \sim \mathcal{N}(0,1)$ and $X_n = -X$. Every X_n has the distribution $\mathcal{N}(0,1)$, so $X_n \rightsquigarrow X$ trivially, but $|X_n - X| = 2|X|$ never shrinks.

Convergence in probability and in distribution survive the usual operations [Wasserman Thm. 5.5]: if $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$ then $X_n + Y_n \xrightarrow{P} X + Y$, $X_n Y_n \xrightarrow{P} XY$, and $h(X_n) \xrightarrow{P} h(X)$ for continuous h. For convergence in distribution the analogous rule is **Slutsky's theorem**: if $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow c$, a *constant*, then $X_n + Y_n \rightsquigarrow X + c$ and $X_n Y_n \rightsquigarrow cX$. (Without the constant it fails: X_n and Y_n could each be $\mathcal{N}(0,1)$ but perfectly anticorrelated, so that $X_n + Y_n = 0$.)

There are two further modes: *almost sure* convergence, $\mathrm{P}(X_n \to X) = 1$, which asks for convergence along each single sequence of outcomes, and convergence in mean, $\mathbb{E}\,|X_n - X| \to 0$. Only the first is needed, once, for the strong law below.

3.3.3 The law of large numbers

Weak law of large numbers. If $X_1, X_2, \ldots$ are iid with mean μ, then $\bar{X}_N \xrightarrow{P} \mu$. The distribution of the sample mean piles up on the true mean.

With a finite variance the proof is one line of Chebyshev (3.16) with $U = \bar{X}_N$, $a = \mu$:

$$\mathrm{P}\big(|\bar{X}_N - \mu| > \epsilon\big) \overset{(1)}{\le} \frac{\operatorname{Var} \bar{X}_N}{\epsilon^2} \overset{(2)}{=} \frac{\sigma^2}{N\epsilon^2} \xrightarrow{N\to\infty} 0. \tag{3.21}$$

(1) Chebyshev's inequality, with $\mathbb{E}\,\bar{X}_N = \mu$ so that the numerator is the variance.
(2) Variance of the mean (3.20).

The theorem holds even without a finite variance, provided $\mathbb{E}\,|X| < \infty$ [Wasserman Thm. 5.6]. The *strong* law says more: with probability one, the running average of a single infinite sequence of readings converges to μ [Wasserman Thm. 5.18]. That is the version that justifies computing an average along one long Monte Carlo run or one long Markov chain.

The law of large numbers is also a statement about the meaning of probability. Let $X_i = \mathbb{1}\{\text{event } A \text{ on trial } i\}$. Then $\bar{X}_N$ is the relative frequency of A and $\mu = \mathrm{P}(A)$, so relative frequencies converge to probabilities. The "frequentist" reading of probability rests on this.

Example: How many tosses? Chebyshev against the truth. For a fair coin, how large must n be so that $\mathrm{P}(0.4 \le \bar{X}_n \le 0.6) \ge 0.7$? With $\mu = 1/2$ and $\operatorname{Var} \bar{X}_n = p(1-p)/n = 1/(4n)$,

$$\mathrm{P}(0.4 \le \bar{X}_n \le 0.6) \overset{(1)}{=} 1 - \mathrm{P}\big(|\bar{X}_n - \tfrac{1}{2}| > 0.1\big) \overset{(2)}{\ge} 1 - \frac{1}{4n\,(0.1)^2} = 1 - \frac{25}{n}.$$

(1) The complement of "within 0.1 of 1/2".
(2) Chebyshev with $\operatorname{Var} \bar{X}_n = 1/(4n)$ and $\epsilon = 0.1$.

The bound exceeds 0.7 once $25/n \le 0.3$, i.e. $n \ge 84$ [Wasserman Ex. 5.7]. Chebyshev is a guarantee for *every* distribution with this variance, so it is pessimistic for any particular one. The exact binomial probability at $n = 84$ is 0.937. Using the Gaussian approximation of the next subsection, $\mathrm{P}(|\bar{X}_n - \tfrac{1}{2}| \le 0.1) \approx 2\Phi(0.1\sqrt{n}/0.5) - 1 \ge 0.7$ requires $0.2\sqrt{n} \ge z_{0.85} = 1.036$, i.e. $n \ge 27$, where the exact probability is 0.7522.

A sharper guarantee, and a first interval. Chebyshev uses only the variance, so it must hold for every distribution. For Bernoulli readings (each X_i is 0 or 1, with $\mathrm{P}(X_i = 1) = p$) we know more, and **Hoeffding's inequality** uses it: $\mathrm{P}(|\bar{X}_n - p| > \epsilon) \le 2e^{-2n\epsilon^2}$ [Wasserman Thm. 4.5, (4.4)]. This bound falls exponentially in n, not as $1/n$. With $n = 100$ and $\epsilon = 0.2$ it gives $2e^{-8} = 0.00067$, where Chebyshev gives only $1/(4 \cdot 100 \cdot 0.04) = 0.0625$ [Wasserman Ex. 4.6]. (For the moderate $\epsilon = 0.1$ of the coin example, Chebyshev happens to be the stronger of the two.)

The bound also gives an interval. Set the right side equal to a small probability α and solve for ϵ: $\epsilon_n = \sqrt{\ln(2/\alpha)/(2n)}$. Then the interval $\hat{p} \pm \epsilon_n$ contains the true p with probability at least $1 - \alpha$, for *every* p and every n, with no large-n approximation [Wasserman Ex. 6.15]. For $\alpha = 0.05$ and $n = 100$, $\epsilon_n = \sqrt{3.689/200} = 0.136$.

What does "with probability at least $1 - \alpha$" refer to? Not to p, which is fixed. It refers to the interval recipe, applied over repeated experiments. A striking case makes the point: Wasserman's Ex. 6.14 builds a legitimate 75% interval that, on some data sets, is known with certainty to contain θ. So the sentence "with probability 0.95, p is contained in the interval", said of $\bar{X} \pm 1.96\, S/\sqrt{n}$ [SeeingTheory p. 42], is loose: the 0.95 belongs to the recipe, not to p. The full story of intervals is in Chapter 4.

3.3.4 The central limit theorem

The law of large numbers says where the sample mean goes. It does not say how $\bar{X}_N$ is distributed on the way, which is what we need for error bars. The central limit theorem does.

Key idea: Central limit theorem

Let $X_1, \dots, X_N$ be iid with mean μ and finite variance σ^2. Then

$$Z_N := \frac{\bar{X}_N - \mu}{\sqrt{\mathrm{Var}\,\bar{X}_N}} = \frac{\sqrt{N}(\bar{X}_N - \mu)}{\sigma} \rightsquigarrow \mathcal{N}(0,1), \tag{3.22}$$

that is, $\mathrm{P}(Z_N \le z) \to \Phi(z)$ for every z. Informally, $\bar{X}_N \approx \mathcal{N}(\mu, \sigma^2/N)$ whatever the parent distribution.

The theorem approximates probabilities, not values: "it's the probability statements that we are approximating, not the random variable itself" [Wasserman p. 77]. No single $\bar{X}_N$ becomes a Gaussian number; what becomes Gaussian is the chance that $\bar{X}_N$ falls in a given range.

In practice we do not know σ. The theorem still holds with σ replaced by the sample standard deviation S_N of Section 3.4. The reason: $S_N \xrightarrow{P} \sigma$, and Slutsky's theorem lets us swap a factor that tends to a constant [Wasserman Thm. 5.10]. For random vectors, $\sqrt{N}(\bar{\boldsymbol{X}}_N - \boldsymbol{\mu}) \rightsquigarrow \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ the covariance matrix of one reading [Wasserman Thm. 5.12].

Why it is true: cumulants add. The cleanest physicist's argument uses the **cumulant generating function** $K_Y(t) = \ln \mathbb{E}\, e^{tY} = \sum_{k \ge 1} \kappa_k t^k / k!$, the logarithm of the moment generating function. Its Taylor coefficients κ_k are the **cumulants** [κ_1 is the mean, κ_2 the variance, $\kappa_3/\kappa_2^{3/2}$ the skewness]. For independent U, V, $\mathbb{E}\, e^{t(U+V)} = \mathbb{E}\, e^{tU}\ \mathbb{E}\, e^{tV}$, so $K_{U+V} = K_U + K_V$: cumulants of independent variables add, like free energies of independent subsystems. Let $Y_i = (X_i - \mu)/\sigma$, which has $\kappa_1 = 0$, $\kappa_2 = 1$, and write $Z_N = N^{-1/2} \sum_i Y_i$. Then

$$K_{Z_N}(t) \overset{(1)}{=} \sum_{i=1}^{N} K_Y\big(t/\sqrt{N}\big) \overset{(2)}{=} N \sum_{k \ge 2} \frac{\kappa_k}{k!} \frac{t^k}{N^{k/2}}$$

$$\overset{(3)}{=} \frac{t^2}{2} + \frac{\kappa_3}{6\sqrt{N}}\,t^3 + \frac{\kappa_4}{24\,N}\,t^4 + \cdots \xrightarrow{N\to\infty} \frac{t^2}{2}. \tag{3.23}$$

(1) Cumulants of independent variables add, and scaling $Y \to cY$ turns $K_Y(t)$ into $K_Y(ct)$, here with $c = 1/\sqrt{N}$.
(2) Insert the series. The $k = 1$ term is absent because $\kappa_1 = 0$.
(3) The $k = 2$ term is $N \cdot \frac{1}{2}t^2/N$. The k-th term carries $N^{1-k/2}$, which vanishes for every $k \geq 3$.

The limit $t^2/2$ is the cumulant generating function of $\mathcal{N}(0,1)$, whose only nonzero cumulant is the variance. Every other cumulant of the standardised mean dies, the skewness fastest, like $\kappa_3/\sqrt{N}$. For an exponential parent $\kappa_3 = 2$, so the skewness of the mean is $2/\sqrt{N}$, exactly the value we met for the muon lifetime in Section 3.1.3. (The rigorous proof uses the characteristic function $\mathbb{E}\,e^{itY}$, which exists for every distribution, and the same expansion to second order [Wasserman Lemma 5.20].)

How large is large? The expansion shows that the error of the Gaussian approximation is controlled by the skewness divided by $\sqrt{N}$. The **Berry–Esseen inequality** makes this quantitative [Wasserman Thm. 5.11, (5.4)]:

$$\sup_z \left|\mathrm{P}(Z_N \leq z) - \Phi(z)\right| \leq \frac{C\;\mathbb{E}\,|X_1 - \mu|^3}{\sqrt{N}\,\sigma^3}, \tag{3.24}$$

with $C = 33/4$ in Wasserman. (That constant is very generous. Much smaller values, below one half, have since been proved.) Symmetric parents converge fast. Skewed ones, like exponential lifetimes or Poisson counts with small means, need more readings.

Example: Pile-up in a collider. The number of proton–proton interactions per bunch crossing is Poisson with mean $\lambda = 5$. Over $n = 125$ crossings, what is the probability that the average number of interactions is below 5.5? With $\mu = \sigma^2 = \lambda = 5$,

$$\mathrm{P}(\bar{X}_n < 5.5) \overset{(1)}{=} \mathrm{P}\Big(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} < \frac{\sqrt{n}(5.5-\mu)}{\sigma}\Big) \overset{(2)}{=} \mathrm{P}\Big(Z < \frac{\sqrt{125}\times 0.5}{\sqrt{5}}\Big) \approx \Phi(2.5) = 0.9938.$$

(1) Subtract μ, multiply by $\sqrt{n}/\sigma > 0$: the event is unchanged.
(2) $\sqrt{125}/\sqrt{5} = \sqrt{25} = 5$, times 0.5 is 2.5. By the CLT the left side is approximately standard normal.

(The same numbers describe a count of errors in computer programs [Wasserman Ex. 5.9].)

Example: When there is no variance: the Breit–Wigner. A resonance of mass m and width Γ has the Breit–Wigner (Cauchy) line shape, which in units $x = (E-m)/(\Gamma/2)$ is $f(x) = 1/[\pi(1+x^2)]$. Its tails fall like x^{-2}, so $\int |x| f\,\mathrm{d}x$ diverges: it has no mean and no variance [Cowan §2.8, §5.2]. Its characteristic function is $\phi(t) = \mathbb{E}\,e^{itX} = e^{-|t|}$. For the average of N independent draws,

$$\mathbb{E}\;e^{it\bar{X}_N} \overset{(1)}{=} \prod_{i=1}^{N} \mathbb{E}\;e^{i(t/N)X_i} \overset{(2)}{=} \left(e^{-|t|/N}\right)^N = e^{-|t|}.$$

(1) The exponential of a sum is a product, and independence lets the average factorise.
(2) Each factor is the characteristic function at t/N.

The average of N Cauchy variables has *exactly* the same distribution as one of them. Averaging does not help at all. The half-width of the central 50% of $\bar{X}_N$ is 1 (the quartiles of the Cauchy are ± 1) for every N. The fix is to use a different estimator. The sample median of Cauchy data *is* consistent. For large N its standard error is $1/(2f(0)\sqrt{N})$, the general large-N result for a median at a point where the density is $f(0)$; with $f(0) = 1/\pi$ this is $\pi/(2\sqrt{N})$.

3.3.5 Seeing it in Python

Listing 3.3: Variance of the sample mean for three parents, and the Cauchy exception (code/ch03/03_one_over_sqrtN.py)

```
parents = {
    "Uniform(0,1)": (lambda s: rng.uniform(0, 1, s), 1 / 12),
    "Exponential(1)": (lambda s: rng.exponential(1.0, s), 1.0),
    "Poisson(3)": (lambda s: rng.poisson(3.0, s).astype(float), 3.0),
}
var_of_mean = {k: [] for k in parents}
iqr = {"Exponential(1)": [], "Cauchy": []}
for N in Ns:
    for k, (draw, _) in parents.items():
        m = draw((M, N)).mean(axis=1)
        var_of_mean[k].append(m.var(ddof=1))
        if k == "Exponential(1)":
            q = np.percentile(m, [25, 75]); iqr[k].append(0.5 * (q[1] - q[0]))
    mc = rng.standard_cauchy((M, N)).mean(axis=1)
    q = np.percentile(mc, [25, 75]); iqr["Cauchy"].append(0.5 * (q[1] - q[0]))

slopes = {k: np.polyfit(np.log(Ns), np.log(v), 1)[0] for k, v in var_of_mean.items()}
```

For each $N = 1, 2, 4, \ldots, 1024$ the script draws $M = 20\,000$ experiments of N readings from a uniform, an exponential and a Poisson parent, averages each row, and records the variance of the M averages. It also records the half-width of the central 50% of the averages for exponential and Cauchy parents (a width that exists even when the variance does not). The fitted log–log slopes of $\mathrm{Var}\,\bar{X}_N$ against N are -1.001, -0.9973 and -1.002 (derived: -1). The Cauchy half-width is 1.011 at $N = 1$ and 0.9998 at $N = 1024$ (derived: 1 for all N).

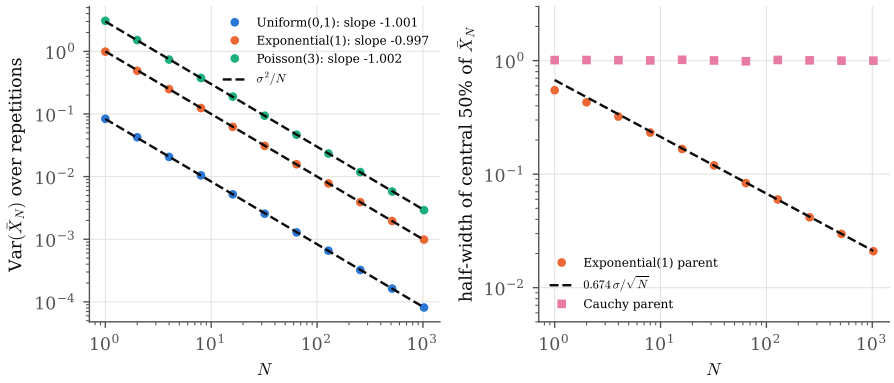


Figure 3.12: Left: the variance of the sample mean falls as σ^2/N (dashed lines) for three very different parents, a straight line of slope -1 on log–log axes. Right: the width of the distribution of the mean falls as $1/\sqrt{N}$ for exponential readings but stays flat for Cauchy readings, where averaging a thousand values is no better than taking one.

Asymptotically normal. An estimator is **asymptotically normal** if $(\hat{\theta}_N - \theta)/\mathrm{se} \rightsquigarrow \mathcal{N}(0,1)$ [for large N its sampling distribution is a Gaussian centred on the truth with width equal to

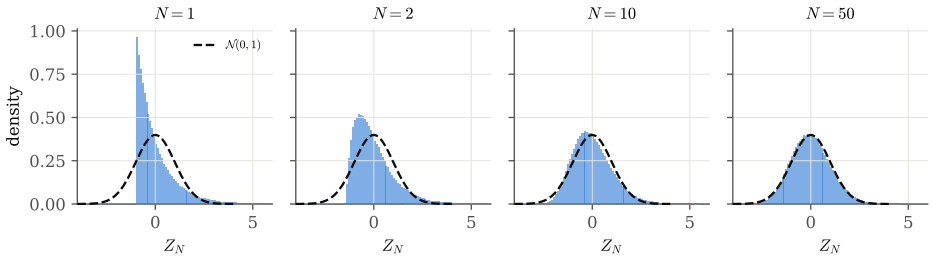


Figure 3.13: The central limit theorem at work on exponential decay times. Each panel is the histogram of the standardised mean Z_N of N exponential readings, over 200 000 repetitions, with the standard normal curve dashed. At $N = 1$ it is the exponential itself, cut off at $Z = -1$. By $N = 10$ it is close to Gaussian with a visible right skew, and by $N = 50$ the residual skewness $2/\sqrt{50} = 0.28$ is hard to see.

the standard error] [Wasserman (6.8)].

Most estimators we will meet are asymptotically normal: sample means by the CLT, smooth functions of means by the delta method of Section 3.6, and maximum-likelihood estimators by the argument of the second half of Chapter 3. That is why "estimate ± standard error" means roughly a 68% interval so often.

Watch out: The CLT is about the mean, not about the data

The CLT does not make the readings Gaussian, and it does not make every statistic Gaussian. The histogram of individual muon decay times stays exponential forever. It is the histogram of their *average* that becomes Gaussian. And it needs a finite variance: for Cauchy-like tails (a resonance, a ratio of noisy quantities, Section 3.6.4) it fails completely.

What this bought us. Independent errors add like the steps of a random walk, so the scatter of a mean is $\sigma/\sqrt{N}$ and its distribution approaches a Gaussian, with a residual skewness that falls like $1/\sqrt{N}$. Correlations put a floor under the error, and a parent without a variance defeats averaging altogether.

Verification and research. The log–log plot verifies σ^2/N where we know σ. In a Monte Carlo computation (Chapter 6) we run the same check backwards: the slope of the measured variance against N tells us whether the samples are really independent (slope -1) or correlated (a flatter slope, as for an MCMC chain in Chapter 9), and the plateau of a correlated run measures its N_{eff}.

3.4 Estimating the scatter itself: the sample variance and its error

The pendulum error bar $\sigma/\sqrt{N}$ contains the noise level σ, which nobody told us. We must estimate it from the same ten readings. A cosmologist faces the same task at every multipole: the angular power spectrum C_ℓ *is* a variance, the variance of the $2\ell + 1$ spherical-harmonic coefficients $a_{\ell m}$, and estimating it from one sky is estimating a variance from a finite sample. So is quoting the error bar of a Monte Carlo average (Chapter 6).

Two facts come out of this. The estimator of a variance that is right on average divides by $N - 1$, not by N. And that estimate has its own uncertainty, the "error on the error", of which the cosmic-variance formula is a special case.

Equipartition with the centre of mass removed. We cannot measure the spread of the readings about the unknown truth μ, only about their own average $\bar{X}$. Deviations from $\bar{X}$ are systematically a little smaller than deviations from μ, because the average sits in the middle of the data by construction. How much smaller? Think of N particles of a gas, each velocity component carrying energy $\frac{1}{2}kT$ on average. If we measure velocities relative to the centre of mass, one degree of freedom is gone and the total relative kinetic energy is $(N-1)\frac{1}{2}kT$. In the same way, the N deviations $X_i - \bar{X}$ sum to zero, only $N-1$ of them are free, and each carries σ^2 on average.

What we want to know.

1. What is the average of $\sum_i (X_i - \bar{X})^2$ over repetitions?
2. What divisor makes the estimate of σ^2 right on average, and is that the best divisor?
3. How much does the estimate of σ^2 itself scatter?
4. What changes when the true mean is known, as for the $a_{\ell m}$ of the CMB?

3.4.1 The $N-1$

Let $X_1, \dots, X_N$ be iid with mean μ and variance σ^2. The question is the average, over repetitions, of the **sum of squared deviations** $\mathrm{SS} = \sum_i (X_i - \bar{X})^2$. We know the average of the squared deviations from μ; it is σ^2 each. So we first relate the two sums. Write $X_i - \mu = (X_i - \bar{X}) + (\bar{X} - \mu)$ and square:

$$\sum_{i=1}^{N} (X_i - \mu)^2 \overset{(1)}{=} \sum_i (X_i - \bar{X})^2 + 2(\bar{X} - \mu) \sum_i (X_i - \bar{X}) + N(\bar{X} - \mu)^2$$
$$\overset{(2)}{=} \mathrm{SS} + N(\bar{X} - \mu)^2. \tag{3.25}$$

(1) Expand the square. The factor $\bar{X} - \mu$ does not depend on i and comes out of the sum.
(2) $\sum_i (X_i - \bar{X}) = N\bar{X} - N\bar{X} = 0$: deviations from the average sum to zero.

This is again the parallel-axis theorem: spread about the truth equals spread about the centre plus N times the squared offset of the centre. Now average over repetitions:

$$\mathbb{E}\,\mathrm{SS} \overset{(1)}{=} \sum_{i=1}^{N} \mathbb{E}(X_i - \mu)^2 - N\,\mathbb{E}(\bar{X} - \mu)^2 \overset{(2)}{=} N\sigma^2 - N \cdot \frac{\sigma^2}{N} = (N-1)\,\sigma^2. \tag{3.26}$$

(1) Rearrange (3.25) and use linearity of expectation.
(2) $\mathbb{E}(X_i - \mu)^2 = \sigma^2$ for each i, and $\mathbb{E}(\bar{X} - \mu)^2 = \mathrm{Var}\,\bar{X} = \sigma^2/N$ from (3.20).

The missing σ^2 is exactly the variance of the average: the data are spread about $\bar{X}$, and $\bar{X}$ is itself off from μ by about $\sigma/\sqrt{N}$.

Working definition 3.6 (sample variance)

The **sample variance** is

$$S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})^2, \tag{3.27}$$

and by (3.26) it is unbiased, $\mathbb{E}\,S^2 = \sigma^2$ [Wasserman p. 52, Thm. 3.17]. The divisor $N-1$ is the

number of **degrees of freedom** [independent directions left in the data after one direction was used to fix $\bar{X}$]. If the true mean μ is *known*, the estimator $S_0^2 = N^{-1}\sum_i (X_i - \mu)^2$ is unbiased with divisor N, since then nothing was used up [Cowan (5.9)–(5.10)].

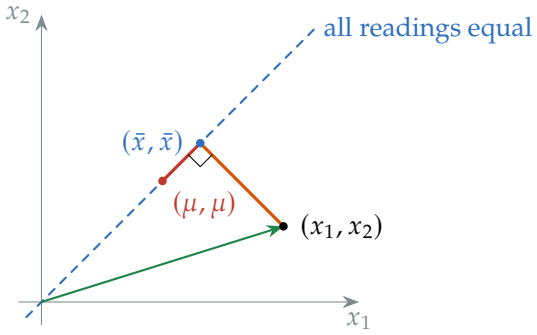


Figure 3.14: Two readings as one point in the plane. The dashed diagonal holds the data sets in which both readings are equal. The average $(\bar{x}, \bar{x})$ is the foot of the perpendicular from the data point onto the diagonal. The orange residual, of squared length SS, lives in the one direction perpendicular to the diagonal. The red segment is the offset of the average from the truth. The right angle is the identity (3.25): distance^2 to the truth = orange^2 + red^2.

The geometry in Figure 3.14 holds for any N. The data are a vector in $\mathbb{R}^N$. The average is its projection on the line spanned by $(1, 1, \ldots, 1)$, and the residual vector $(X_i - \bar{X})_i$ lives in the $(N-1)$-dimensional subspace perpendicular to it. If the noise is isotropic (independent, equal variance), each of those $N-1$ directions carries variance σ^2, which is (3.26) without algebra. For a Gaussian parent we get the whole distribution, not just the average. Rotate to a new orthonormal basis whose first axis is the diagonal. Independent Gaussian noise of equal variance looks the same in every such basis, so the new coordinates are again independent Gaussians of variance σ^2. The first carries $\bar{X}$; the other $N-1$ carry the residual. So SS/σ^2 is a sum of $N-1$ squares of independent standard normals, that is

$$\frac{(N-1)S^2}{\sigma^2} \sim \chi^2_{N-1}, \tag{3.28}$$

independently of $\bar{X}$ (the χ^2 distribution is in Chapter 2).

Example: The pendulum's noise level. For the ten readings of Section 3.1, $\bar{x} = 9.815$ and the deviations are $-0.025, 0.015, 0.045, -0.015, -0.045, 0.025, -0.005, -0.035, 0.035, 0.005$. Their squares sum to $2(0.025^2 + 0.015^2 + 0.045^2 + 0.035^2 + 0.005^2) = 2 \times 0.004125 = 0.00825$. Hence

$$s^2 \overset{(1)}{=} \frac{0.00825}{9} = 9.17 \times 10^{-4}, \qquad s = 0.0303, \qquad \widehat{\mathrm{se}}(\bar{x}) \overset{(2)}{=} \frac{s}{\sqrt{10}} = 0.0096\ \mathrm{m\,s^{-2}}.$$

(1) Divide the sum of squared deviations by $N - 1 = 9$.
(2) The estimated standard error of the mean replaces σ by s in (3.20).

The result is $g = 9.815 \pm 0.010\ \mathrm{m\,s^{-2}}$. How much should we trust the "±0.010"? That is the question of Section 3.4.3.

Example: S^2 is consistent. Expanding the square in (3.27), $\sum_i (X_i - \bar{X})^2 = \sum_i X_i^2 - N\bar{X}^2$, so

$$S^2 \overset{(1)}{=} \frac{N}{N-1}\Big[\frac{1}{N}\sum_{i=1}^{N} X_i^2 - \bar{X}^2\Big] \xrightarrow{P} 1 \cdot \big[\mathbb{E}\,X^2 - \mu^2\big] \overset{(2)}{=} \sigma^2.$$

(1) Factor out $N/(N-1)$. Inside the bracket, the first term is the mean of the iid variables X_i^2, which tends in probability to $\mathbb{E}\,X^2$ by the WLLN. $\bar{X} \xrightarrow{P} \mu$ and squaring is continuous, so $\bar{X}^2 \xrightarrow{P} \mu^2$. Sums and products of sequences that converge in probability converge to the sums and products of the limits, and $N/(N-1) \to 1$.
(2) $\mathbb{E}\,X^2 - \mu^2 = \sigma^2$.

[CasellaBerger Ex. 5.5.3]; [Wasserman Exercise 5.1]. The same argument applied with the continuous map $\sqrt{\cdot}$ gives $S \xrightarrow{P} \sigma$.

Watch out: S is a biased estimator of σ

Unbiasedness does not survive nonlinear maps. Because $\operatorname{Var} S = \mathbb{E}\,S^2 - (\mathbb{E}\,S)^2 \geq 0$, we have $(\mathbb{E}\,S)^2 \leq \mathbb{E}\,S^2 = \sigma^2$, so $\mathbb{E}\,S \leq \sigma$, with equality only if S never fluctuates (this is Jensen's inequality for the concave square root). For Gaussian data, $\mathbb{E}\,S = c_4(N)\,\sigma$ with $c_4(N) = \sqrt{2/(N-1)}\;\Gamma(N/2)/\Gamma\big((N-1)/2\big)$, which follows from the mean of the square root of a χ^2_{N-1} variable. For $N = 5$, $c_4 = 0.94$; the simulation gives $\mathbb{E}\,S/\sigma = 0.94$. S is biased for every finite N but consistent [CasellaBerger Ex. 5.5.5].

3.4.2 Is $N-1$ the best divisor?

Unbiased is not the same as best (Section 3.2). For Gaussian data we can compute the MSE of $W_d = \mathrm{SS}/d$ for any divisor d, because by (3.28) $\mathrm{SS} = \sigma^2\chi^2_{N-1}$ and a χ^2_ν variable has mean ν and variance 2ν.

Example: The divisor that minimises the MSE.

$$\mathrm{MSE}(W_d) \overset{(1)}{=} \left(\frac{(N-1)\sigma^2}{d} - \sigma^2\right)^2 + \frac{2(N-1)\sigma^4}{d^2} \overset{(2)}{=} \sigma^4\Big[\big((N-1)u - 1\big)^2 + 2(N-1)u^2\Big], \quad u = \frac{1}{d}.$$

(1) Squared bias plus variance (3.13), with $\mathbb{E}\,W_d = (N-1)\sigma^2/d$ and $\operatorname{Var} W_d = 2(N-1)\sigma^4/d^2$.
(2) Factor out σ^4 and write $u = 1/d$.

Setting the u-derivative to zero, $2(N-1)\big[(N-1)u - 1\big] + 4(N-1)u = 0$, i.e. $(N+1)u = 1$. The minimum is at $d = N+1$. For $N = 10$ the three candidates give

$$d = N-1:\ \tfrac{2}{9} = 0.222, \qquad d = N:\ 0.01 + 0.18 = 0.190, \qquad d = N+1:\ \tfrac{4}{121} + \tfrac{18}{121} = 0.182$$

(in units of σ^4). The simulation gives 0.2229, 0.1907 and 0.1826.

So the unbiased S^2 does not have the smallest MSE. Even $d = N$ beats it [CasellaBerger Exs. 7.3.3–7.3.4]: the biased maximum-likelihood estimator $\hat{\sigma}^2 = \mathrm{SS}/N$ has MSE $(2N-1)\sigma^4/N^2$, smaller than the $2\sigma^4/(N-1)$ of S^2. A larger divisor shrinks the estimate towards zero, trading a small downward bias for a reduction in variance.

Why, then, do we usually quote S^2? For two reasons. First, estimates are often *averaged* later, and when they are, biases add up while scatter averages away; an unbiased estimate is then what we want. Second, the optimal divisor depends on the parent: $N+1$ is special to Gaussian data.

3.4.3 The variance of the variance

How much does S^2 scatter? The answer involves the fourth central moment $\mu_4 = \mathbb{E}(X-\mu)^4$. Without loss of generality set $\mu = 0$ (shifting every reading by a constant changes neither S^2 nor

the central moments). Using $\sum_i (X_i - \bar{X})^2 = \sum_i X_i^2 - \frac{1}{N}\sum_{i,j} X_i X_j$, split S^2 into a "diagonal" and an "off-diagonal" part:

$$S^2 = \underbrace{\frac{1}{N}\sum_i X_i^2}_{A} - \underbrace{\frac{1}{N(N-1)}\sum_{i\neq j} X_i X_j}_{B}. \tag{3.29}$$

(The coefficient of X_i^2 is $\frac{1}{N-1}(1-\frac{1}{N}) = \frac{1}{N}$, and that of X_iX_j, $i \neq j$, is $\frac{1}{N-1}\cdot\frac{1}{N}$.) The variance of a difference is $\operatorname{Var}(A-B) = \operatorname{Var}A + \operatorname{Var}B - 2\operatorname{Cov}(A,B)$, so we need three pieces.

$$\operatorname{Var}A \overset{(1)}{=} \frac{1}{N^2}\, N\, \operatorname{Var}(X^2) \overset{(2)}{=} \frac{\mu_4 - \sigma^4}{N}, \tag{3.30}$$

$$\mathbb{E}\,B^2 \overset{(3)}{=} \frac{1}{N^2(N-1)^2}\sum_{i\neq j}\sum_{k\neq l}\mathbb{E}[X_iX_jX_kX_l] \overset{(4)}{=} \frac{2N(N-1)\sigma^4}{N^2(N-1)^2} = \frac{2\sigma^4}{N(N-1)}, \tag{3.31}$$

$$\operatorname{Cov}(A,B) \overset{(5)}{=} \frac{1}{N^2(N-1)}\sum_k\sum_{i\neq j}\mathbb{E}[X_k^2X_iX_j] \overset{(6)}{=} 0. \tag{3.32}$$

(1) The X_i^2 are independent, so the variance of their sum is the sum of their variances.
(2) $\operatorname{Var}(X^2) = \mathbb{E}\,X^4 - (\mathbb{E}\,X^2)^2 = \mu_4 - \sigma^4$ (recall $\mu = 0$).
(3) $\mathbb{E}\,B = 0$ because each term has $\mathbb{E}\,X_iX_j = \mathbb{E}\,X_i\;\mathbb{E}\,X_j = 0$. So $\operatorname{Var}B = \mathbb{E}\,B^2$, which is the double sum over pairs of pairs.
(4) A product $X_iX_jX_kX_l$ with $i \neq j$ has zero mean unless every index appears at least twice, which forces $\{k,l\} = \{i,j\}$. There are two such choices, $(k,l) = (i,j)$ or (j,i), each giving $\mathbb{E}\,X_i^2X_j^2 = \sigma^4$. There are $N(N-1)$ ordered pairs (i,j).
(5) $\mathbb{E}\,B = 0$, so $\operatorname{Cov}(A,B) = \mathbb{E}[AB]$. Multiply out the two sums.
(6) Since $i \neq j$, at least one of i, j differs from k and from the other index, so it appears once and its factor $\mathbb{E}\,X = 0$ kills the term.

Putting the pieces together,

$$\begin{aligned}\operatorname{Var}S^2 &\overset{(1)}{=} \operatorname{Var}A + \operatorname{Var}B - 2\operatorname{Cov}(A,B) \overset{(2)}{=} \frac{\mu_4-\sigma^4}{N} + \frac{2\sigma^4}{N(N-1)} \\ &\overset{(3)}{=} \frac{1}{N}\left[\mu_4 - \frac{N-3}{N-1}\,\sigma^4\right].\end{aligned} \tag{3.33}$$

(1) Variance of a difference.
(2) Insert (3.30)–(3.32).
(3) $-\sigma^4 + \frac{2\sigma^4}{N-1} = -\sigma^4\,\frac{N-1-2}{N-1}$.

This is [Cowan (5.14)]. For a Gaussian parent, $\mu_4 = 3\sigma^4$ (from Isserlis' theorem, or by direct integration: $\mathbb{E}\,Z^4 = 3$ for a standard normal), and

$$\operatorname{Var}S^2\big|_{\text{Gauss}} = \frac{\sigma^4}{N}\,\frac{3(N-1)-(N-3)}{N-1} = \frac{2\sigma^4}{N-1}, \tag{3.34}$$

in agreement with $\operatorname{Var}\chi^2_{N-1} = 2(N-1)$ in (3.28).

The error on the error. For Gaussian data the relative error of S^2 is $\sqrt{2/(N-1)}$, and that of S is half as large (a relative error halves under a square root; this is the propagation rule of Section 3.6):

$$\frac{\operatorname{se}(S)}{\sigma} \approx \frac{1}{\sqrt{2(N-1)}}. \tag{3.35}$$

With ten readings an error bar is itself uncertain by about 24%. With a hundred, by 7%. Quoting an error bar to three significant figures from ten points is meaningless.

For a heavy-tailed parent the variance of S^2 is larger, because μ_4 weights the tails heavily. An exponential parent of mean τ has $\sigma = \tau$ and $\mu_4 = 9\tau^4 = 9\sigma^4$, three times the Gaussian value. To quote the uncertainty of S^2 from data, replace μ_4 by the sample moment $m_4 = \frac{1}{N}\sum_i (X_i - \bar{X})^4$ [Cowan (5.15)].

Example: Cosmic variance. In a statistically isotropic Gaussian sky, the $2\ell+1$ coefficients $a_{\ell m}$ at multipole ℓ behave (for this count) like $2\ell+1$ independent real Gaussian variables of mean zero and variance C_ℓ (the precise complex bookkeeping is in Chapter 7). The natural estimator $\widehat{C}_\ell = \frac{1}{2\ell+1}\sum_m |a_{\ell m}|^2$ is the known-mean estimator S_0^2 with $N = 2\ell+1$. With known mean zero, $\sum_m a_{\ell m}^2 / C_\ell$ is a sum of $2\ell+1$ squared standard normals, i.e. $\chi^2_{2\ell+1}$, so

$$\mathbb{E}\,\widehat{C}_\ell \overset{(1)}{=} \frac{C_\ell}{2\ell+1}(2\ell+1) = C_\ell, \qquad \operatorname{Var}\widehat{C}_\ell \overset{(2)}{=} \frac{C_\ell^2}{(2\ell+1)^2}\cdot 2(2\ell+1) = \frac{2C_\ell^2}{2\ell+1}.$$

(1) A χ^2_ν variable has mean ν.
(2) It has variance 2ν, and $\operatorname{Var}(cU) = c^2 \operatorname{Var} U$ with $c = C_\ell/(2\ell+1)$.

This is **cosmic variance**: the relative error $\sqrt{2/(2\ell+1)}$ is 63% at the quadrupole $\ell = 2$ and cannot be reduced, because we have only one sky. It is the known-mean version of (3.34), with N in place of $N-1$ because no degree of freedom is spent on estimating a mean.

3.4.4 Seeing it in Python

Listing 3.4: The two divisors and the scatter of S^2 (`code/ch03/04_sample_variance.py`)

```
mse = {"N-1": [], "N": [], "N+1": []}
for N in Ns:
    x = rng.standard_normal((M, N))                  # sigma = 1
    ss = ((x - x.mean(axis=1, keepdims=True)) ** 2).sum(axis=1)
    mean_biased.append((ss / N).mean())
    mean_unbiased.append((ss / (N - 1)).mean())
    var_S2_gauss.append((ss / (N - 1)).var(ddof=1))
    for lab, d in (("N-1", N - 1), ("N", N), ("N+1", N + 1)):
        mse[lab].append(np.mean((ss / d - 1.0) ** 2))
    y = rng.exponential(1.0, (M, N))                 # sigma = 1, mu_4 = 9
    ssy = ((y - y.mean(axis=1, keepdims=True)) ** 2).sum(axis=1)
```

Each row of `x` is one experiment of N standard normal readings ($\sigma = 1$). The script forms SS row by row, divides by N and by $N-1$, and averages over the $M = 100\,000$ rows. It also records the variance of S^2 across rows and the MSE of the three divisors. The last three lines repeat the variance measurement for an exponential parent. For $N = 5$ the average of SS/N is 0.796 (derived: $4/5 = 0.8$) and of SS/$(N-1)$ is 0.995 (derived: 1). At $N = 10$, $\operatorname{Var} S^2 = 0.2229$ for Gaussian readings (derived 0.2222) and 0.8255 for exponential ones (derived from (3.33) with $\mu_4 = 9$: 0.8222).

What this bought us. Divide by $N-1$ because one degree of freedom went into $\bar{X}$. The resulting S^2 is unbiased but has its own scatter, $\operatorname{Var} S^2 = \frac{1}{N}[\mu_4 - \frac{N-3}{N-1}\sigma^4]$, which for Gaussian data means a relative error of $1/\sqrt{2(N-1)}$ on S. With a known mean the divisor is N, and the Gaussian scatter $2\sigma^4/N$ is cosmic variance.

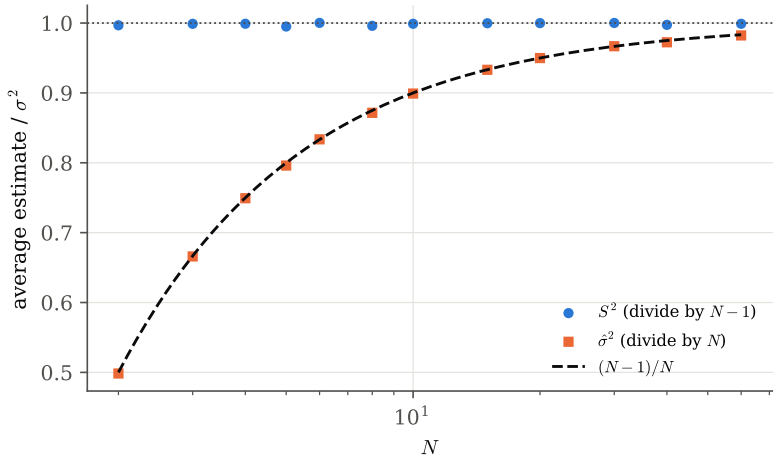


Figure 3.15: Average of two variance estimators over 100 000 Gaussian experiments, as a function of the number of readings. Dividing by $N-1$ is right on average at every N. Dividing by N falls short by the factor $(N-1)/N$ (dashed), badly for $N=2$ and negligibly beyond a few tens.

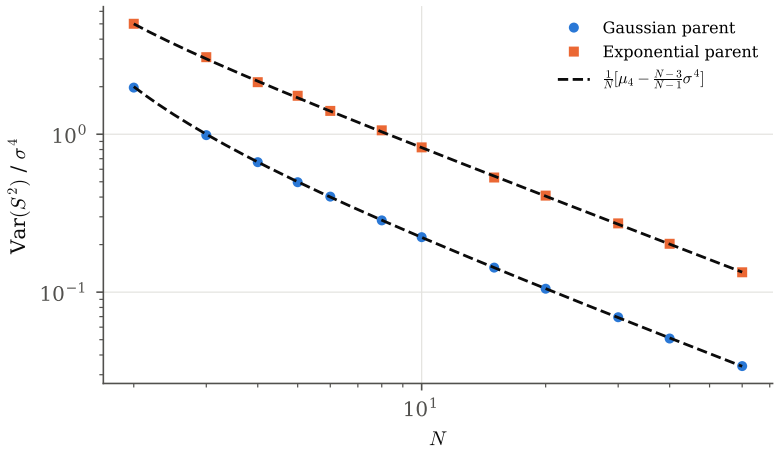


Figure 3.16: Scatter of the sample variance over repetitions. Both parents follow the derived formula (3.33) (dashed), falling roughly as $1/N$. The exponential parent, with its long tail, has a variance of the variance about four times larger than the Gaussian at the same N.

Verification and research. Here the simulation verifies (3.33). The formula is needed later as an instrument: the error bar of any Monte Carlo estimate of a variance, and hence of every simulated covariance matrix (Section 3.5), is set by it. For a masked sky the $a_{\ell m}$ are no longer independent, the effective number of modes drops below $2\ell+1$, often to roughly $f_{\mathrm{sky}}(2\ell+1)$, and the true variance of the power spectrum estimate is found by simulation (Chapter 8).

3.5 Estimating covariance and correlation matrices, and how noisy they are

A CMB or galaxy-survey analysis compresses a map into $p = 20$ band powers and compares them with theory through a $\chi^2 = (\boldsymbol{d}-\boldsymbol{m})^{\mathsf{T}}\mathbf{C}^{-1}(\boldsymbol{d}-\boldsymbol{m})$. The covariance matrix $\mathbf{C}$ of the band powers has no formula once masks and noise enter, so we run n simulations and estimate it from the scatter of the simulated band powers. Supercomputer time limits us to, say, $n = 50$. The resulting matrix shows correlations between bins that are not there, and its inverse is too

large. In the same way a detector with twenty readout channels shows "cross-talk" between channels that is just noise.

The sample variance, extended to a matrix, tells us how noisy each element is, and so how large the false correlations can be. A change of variable gives a reliable error bar for a real correlation. And a single factor removes the bias of the inverse.

A tilted cloud, seen through few points. A correlation coefficient measures the tilt of a cloud of points in a scatter plot. With a thousand points the tilt is obvious. With ten points drawn from a perfectly round cloud, the ten happen to line up a little by chance, and a tilt appears. A 20×20 correlation matrix has 190 tilts to get wrong, so some of them will be large by chance. It is the same reason why a few stars in the sky seem to form constellations.

The questions, in words.

1. How do we estimate, without bias, the covariance of two quantities from paired readings?
2. If two quantities are truly unrelated, how large a sample correlation can chance produce?
3. For a real correlation, what is the sampling distribution of its estimate, and how do we put an error bar on it?
4. What happens when we invert a noisy covariance matrix?

3.5.1 The sample covariance matrix

Recall from Chapter 2 the covariance $\Sigma_{ab} = \mathrm{Cov}(X_a, X_b) = \mathbb{E}[(X_a - \mu_a)(X_b - \mu_b)]$ of two components of a random vector $\boldsymbol{X} = (X_1, \dots, X_p)$, and the correlation $\rho_{ab} = \Sigma_{ab}/\sqrt{\Sigma_{aa}\Sigma_{bb}}$, which lies in $[-1, 1]$.

Working definition 3.7 (sample covariance and correlation)

Given n independent readings $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ of a p-component vector [each $\boldsymbol{x}_j$ is one simulation, or one event, giving all p numbers at once], with average $\bar{\boldsymbol{x}} = n^{-1}\sum_j \boldsymbol{x}_j$, the **sample covariance matrix** is

$$\hat{\mathbf{C}} = \frac{1}{n-1}\sum_{j=1}^{n}(\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})^{\mathsf{T}}, \qquad \hat{C}_{ab} = \frac{1}{n-1}\sum_{j=1}^{n}(x_{aj} - \bar{x}_a)(x_{bj} - \bar{x}_b), \tag{3.36}$$

and the **sample correlation** is

$$r_{ab} = \frac{\hat{C}_{ab}}{\sqrt{\hat{C}_{aa}\hat{C}_{bb}}} = \frac{\sum_j (x_{aj} - \bar{x}_a)(x_{bj} - \bar{x}_b)}{\sqrt{\sum_j (x_{aj} - \bar{x}_a)^2 \, \sum_j (x_{bj} - \bar{x}_b)^2}}. \tag{3.37}$$

In (3.37) the factors $1/(n-1)$ cancel between numerator and denominator, which gives the second form; by the Cauchy–Schwarz inequality it always satisfies $|r| \le 1$. The printed formula [Wasserman (14.6)] drops the $1/(n-1)$ from the numerator only, which would let $|r|$ exceed 1. The form above is the correct one and agrees with [Cowan (5.12)].

The diagonal of $\hat{\mathbf{C}}$ holds the sample variances of Section 3.4. The off-diagonal elements are unbiased by the same argument. Write $\boldsymbol{y}_j = \boldsymbol{x}_j - \boldsymbol{\mu}$, so that $\boldsymbol{x}_j - \bar{\boldsymbol{x}} = \boldsymbol{y}_j - \bar{\boldsymbol{y}}$, and use

$\sum_j (\boldsymbol{y}_j - \bar{\boldsymbol{y}})(\boldsymbol{y}_j - \bar{\boldsymbol{y}})^\mathsf{T} = \sum_j \boldsymbol{y}_j \boldsymbol{y}_j^\mathsf{T} - n\, \bar{\boldsymbol{y}}\bar{\boldsymbol{y}}^\mathsf{T}$ (the matrix version of (3.25)):

$$\mathbb{E}\Big[\sum_j (\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})^\mathsf{T}\Big] \overset{(1)}{=} \sum_j \mathbb{E}[\boldsymbol{y}_j \boldsymbol{y}_j^\mathsf{T}] - n\, \mathbb{E}[\bar{\boldsymbol{y}}\bar{\boldsymbol{y}}^\mathsf{T}] \overset{(2)}{=} n\boldsymbol{\Sigma} - n \cdot \frac{\boldsymbol{\Sigma}}{n} = (n-1)\,\boldsymbol{\Sigma}. \tag{3.38}$$

(1) Linearity of expectation applied element by element.
(2) $\mathbb{E}[\boldsymbol{y}_j \boldsymbol{y}_j^\mathsf{T}] = \boldsymbol{\Sigma}$ by definition. For the mean, $\mathbb{E}[\bar{y}_a \bar{y}_b] = n^{-2} \sum_{j,k} \mathbb{E}[y_{aj} y_{bk}] = n^{-2} \cdot n\, \Sigma_{ab}$, because readings $j \neq k$ are independent with zero mean.

So $\mathbb{E}\,\hat{\mathbf{C}} = \boldsymbol{\Sigma}$ [Wasserman p. 233, (14.4)]. A handy form for computing is $\hat{C}_{ab} = \frac{n}{n-1}\big(\overline{x_a x_b} - \bar{x}_a \bar{x}_b\big)$, where the bar means the average over readings [Cowan (5.11)].

Example: A 2×2 covariance by hand. Four paired readings $(x, y) = (1,2), (2,1), (3,4), (4,3)$. Both averages are 2.5. The deviations are $\delta x = (-1.5, -0.5, 0.5, 1.5)$ and $\delta y = (-0.5, -1.5, 1.5, 0.5)$. Then

$$\hat{C}_{xx} \overset{(1)}{=} \tfrac{1}{3}(2.25 + 0.25 + 0.25 + 2.25) = \tfrac{5}{3}, \qquad \hat{C}_{yy} \overset{(2)}{=} \tfrac{5}{3},$$
$$\hat{C}_{xy} \overset{(3)}{=} \tfrac{1}{3}(0.75 + 0.75 + 0.75 + 0.75) = 1, \qquad r \overset{(4)}{=} \frac{1}{\sqrt{(5/3)(5/3)}} = 0.6.$$

(1) Squares of the x deviations, divided by $n - 1 = 3$.
(2) The y deviations are a permutation of the x ones, so the same sum.
(3) Products $(-1.5)(-0.5)$, $(-0.5)(-1.5)$, $(0.5)(1.5)$, $(1.5)(0.5)$, each 0.75.
(4) Definition (3.37).

Four points already give a respectable-looking $r = 0.6$. Is it real? To answer, we need to know how large r gets when there is no correlation at all.

3.5.2 Correlations from pure noise

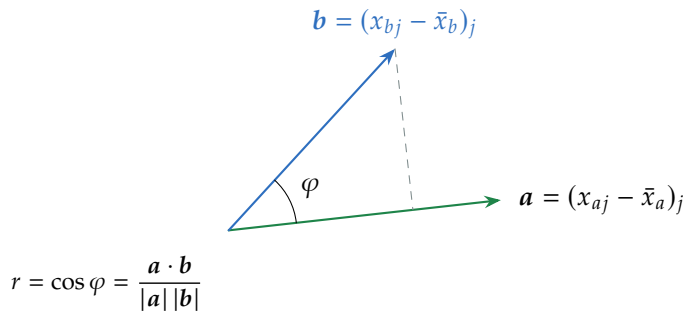


Figure 3.17: The sample correlation as an angle. The deviations of the n readings of each quantity form two vectors in $\mathbb{R}^n$, both perpendicular to $(1, \ldots, 1)$ and so living in an $(n-1)$-dimensional space. The sample correlation is the cosine of the angle between them. Uncorrelated noise points $\boldsymbol{b}$ in a random direction relative to $\boldsymbol{a}$.

Formula (3.37) is a dot product divided by two lengths, so it is the cosine of the angle between the two deviation vectors in Figure 3.17. This picture gives the spread of r for unrelated quantities without any integrals.

Example: The scatter of r when $\rho = 0$. Let the two quantities be independent Gaussians. As in Figure 3.14, the deviation vectors live in the $(m = n - 1)$-dimensional subspace perpendicular

to $(1, \dots, 1)$, and in it the Gaussian noise vector $\boldsymbol{b}$ is isotropic: its direction $\boldsymbol{u} = \boldsymbol{b}/|\boldsymbol{b}|$ is uniformly distributed on the unit sphere, independently of $\boldsymbol{a}$. Rotate coordinates so that $\boldsymbol{a}$ points along the first axis. Then $r = u_1$ and

$$\mathbb{E}\, r \overset{(1)}{=} 0, \qquad \mathbb{E}\, r^2 = \mathbb{E}\, u_1^2 \overset{(2)}{=} \frac{1}{m} \sum_{k=1}^{m} \mathbb{E}\, u_k^2 \overset{(3)}{=} \frac{1}{m}\, \mathbb{E}\Big[\sum_{k=1}^{m} u_k^2\Big] \overset{(4)}{=} \frac{1}{m} = \frac{1}{n-1}.$$

(1) $\boldsymbol{u}$ and $-\boldsymbol{u}$ are equally likely.
(2) By symmetry every component of a uniform unit vector has the same mean square.
(3) Linearity of expectation.
(4) A unit vector has $\sum_k u_k^2 = 1$ exactly.

So for unrelated quantities r scatters about zero with standard deviation $1/\sqrt{n-1}$, and for large n it is close to Gaussian. With $n = 4$ as in Section 3.5.1, $1/\sqrt{3} = 0.58$: our $r = 0.6$ is one standard deviation from zero and means nothing.

Key idea: Spurious correlations

If p quantities are truly independent, each of the $p(p-1)/2$ off-diagonal sample correlations scatters about zero with standard deviation about $1/\sqrt{n-1}$. Among many pairs, roughly 5% exceed $1.96/\sqrt{n-1}$ in magnitude by chance alone.

With $p = 20$ and $n = 50$ there are 190 pairs. The prediction is a spread of $1/\sqrt{49} = 0.1429$ and about ten pairs beyond $|r| = 0.28$. The simulation (Figure 3.18) gives a spread of 0.1484 in one realisation and a largest $|r|$ of 0.4416. Searching a large matrix for the most correlated pair and then quoting its significance as if it were the only pair examined is the *look-elsewhere effect* of Chapter 4.

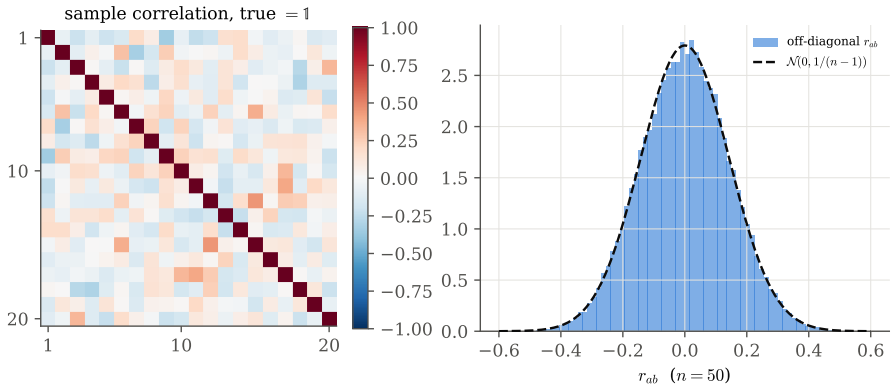


Figure 3.18: Left: the sample correlation matrix of twenty independent quantities from fifty readings. The true matrix is the identity. Every coloured off-diagonal square is noise, and some reach ±0.4. Right: the off-diagonal elements from 400 such matrices, with the Gaussian of standard deviation $1/\sqrt{n-1}$ (dashed).

3.5.3 A real correlation, its error bar, and Fisher's z

For a bivariate Gaussian with correlation $\rho \neq 0$, the sample correlation is biased and skewed, with, to leading order in $1/n$ [Cowan (5.16)–(5.17)],

$$\mathbb{E}\, r \approx \rho - \frac{\rho(1-\rho^2)}{2n}, \qquad \operatorname{Var} r \approx \frac{(1-\rho^2)^2}{n}. \tag{3.39}$$

The variance shrinks as $|\rho| \to 1$, because r cannot exceed 1: the sampling distribution is squashed against the wall and develops a long tail towards zero. An error bar $r \pm \sqrt{\operatorname{Var} r}$ is then misleading, since it depends on the unknown ρ and may poke above 1.

Fisher's remedy is a change of variable that makes the width independent of ρ. We look for $z = h(r)$ with $\operatorname{Var} z$ constant. To first order in the fluctuation (the propagation rule derived in Section 3.6), $\operatorname{Var} h(r) \approx h'(\rho)^2 \operatorname{Var} r$, so we need $h'(\rho) \propto 1/(1-\rho^2)$:

$$h(\rho) \overset{(1)}{=} \int_0^\rho \frac{\mathrm{d}s}{1-s^2} \overset{(2)}{=} \frac{1}{2}\int_0^\rho \Big(\frac{1}{1+s} + \frac{1}{1-s}\Big)\mathrm{d}s \overset{(3)}{=} \frac{1}{2}\ln\frac{1+\rho}{1-\rho} = \operatorname{artanh}\rho. \tag{3.40}$$

(1) Choose the integration constant so that $h(0) = 0$.
(2) Partial fractions: $\frac{1}{1-s^2} = \frac{1}{2}\big(\frac{1}{1+s} + \frac{1}{1-s}\big)$.
(3) Integrate: $\ln(1+s) - \ln(1-s)$, evaluated from 0 to ρ.

Then $\operatorname{Var} z \approx (1-\rho^2)^{-2}(1-\rho^2)^2/n = 1/n$. A finer calculation replaces n by $n-3$, and z is very nearly Gaussian even for small n:

$$z = \operatorname{artanh} r \approx \mathcal{N}\Big(\operatorname{artanh}\rho,\ \frac{1}{n-3}\Big). \tag{3.41}$$

The inverse map is $\rho = \tanh z = (e^{2z}-1)/(e^{2z}+1)$. The map stretches the region near ± 1, which pulls the squashed tail back into a symmetric shape.

Example: An interval for a correlation. We observe $r = 0.6$ from $n = 20$ pairs. A 95% interval is built in z and mapped back [Wasserman p. 234]:

$$\hat z \overset{(1)}{=} \operatorname{artanh} 0.6 = \tfrac{1}{2}\ln 4 = 0.693, \qquad \frac{1.96}{\sqrt{n-3}} \overset{(2)}{=} \frac{1.96}{\sqrt{17}} = 0.475,$$

$$(a, b) \overset{(3)}{=} (0.218,\ 1.169), \qquad (\tanh a, \tanh b) \overset{(4)}{=} (0.214,\ 0.824).$$

(1) $\frac{1}{2}\ln\frac{1.6}{0.4} = \frac{1}{2}\ln 4$.
(2) The standard deviation of z from (3.41), times 1.96.
(3) $\hat z \mp 0.475$.
(4) Map each end back with tanh, which is increasing, so the order is kept.

The interval $(0.21, 0.82)$ is asymmetric about 0.6, longer on the side away from the wall at 1, exactly as the skewed sampling distribution demands.

Listing 3.5: Simulating the sample correlation for $\rho = 0.6$, $n = 20$ (`code/ch03/05_covariance_noise.py`)

```
rho, n2, M = 0.6, 20, 100_000
L = np.linalg.cholesky(np.array([[1, rho], [rho, 1]]))
Z = rng.standard_normal((M, n2, 2)) @ L.T
a = Z[..., 0] - Z[..., 0].mean(axis=1, keepdims=True)
b = Z[..., 1] - Z[..., 1].mean(axis=1, keepdims=True)
r = (a * b).sum(1) / np.sqrt((a * a).sum(1) * (b * b).sum(1))
z = np.arctanh(r)
```

The script makes correlated pairs by multiplying independent standard normals by the Cholesky factor L of the target 2×2 correlation matrix ($LL^{\mathsf T}$ equals that matrix, so $L\epsilon$ has exactly that covariance, Chapter 2). It removes the mean of each column within every experiment and applies (3.37) row by row. Over 100 000 experiments the mean of r is 0.5902 (from (3.39): 0.5904), its spread 0.1544 (leading order: 0.1431, an underestimate at this small n) and its skewness -0.8008. For z, the spread is 0.2416 against $1/\sqrt{17} = 0.2425$, and the skewness -8.926×10^{-4}: Fisher's transformation has removed it.

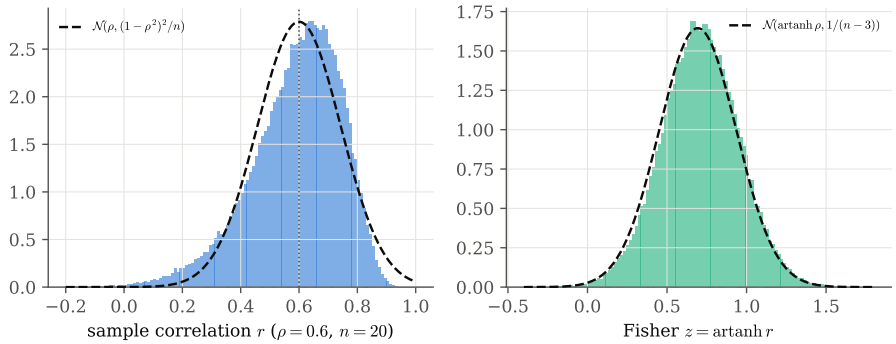


Figure 3.19: Left: the sample correlation from twenty pairs with true correlation 0.6 (dotted). It is skewed, with a long tail towards zero, and the leading-order Gaussian (dashed) is too narrow on that side. Right: the same experiments after Fisher's transformation $z = \operatorname{artanh} r$. The histogram now sits on a symmetric Gaussian of width $1/\sqrt{n-3}$.

3.5.4 How noisy is every element, and what inversion does

For Gaussian readings there is a clean formula for the scatter of each element of $\hat{\mathbf{C}}$. The rotation argument of Section 3.4.1, applied to each component, turns the n centred readings into $n-1$ independent vectors $z_k \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, with $(n-1)\hat{\mathbf{C}} = \sum_{k=1}^{n-1} z_k z_k^{\mathsf{T}}$. We need one fact about Gaussian fourth moments, familiar to physicists as Wick's theorem.

Isserlis (Wick) theorem. For a zero-mean Gaussian vector, $\mathbb{E}[z_a z_b z_c z_d] = \Sigma_{ab}\Sigma_{cd} + \Sigma_{ac}\Sigma_{bd} + \Sigma_{ad}\Sigma_{bc}$: the sum over all ways of pairing the four factors, each pair replaced by its covariance. (With $a = b = c = d$ it gives $\mathbb{E}\, z^4 = 3\sigma^4$.)

$$\operatorname{Var}\hat{C}_{ab} \overset{(1)}{=} \frac{1}{(n-1)^2}\sum_{k=1}^{n-1}\operatorname{Var}(z_{ak}z_{bk}) \overset{(2)}{=} \frac{1}{n-1}\Big(\mathbb{E}[z_a z_b z_a z_b] - \Sigma_{ab}^2\Big)$$
$$\overset{(3)}{=} \frac{1}{n-1}\Big(\Sigma_{aa}\Sigma_{bb} + 2\Sigma_{ab}^2 - \Sigma_{ab}^2\Big) = \frac{\Sigma_{aa}\Sigma_{bb} + \Sigma_{ab}^2}{n-1}. \tag{3.42}$$

(1) $\hat{C}_{ab}$ is $1/(n-1)$ times a sum of $n-1$ independent terms, so variances add.
(2) All $n-1$ terms have the same distribution, and $\operatorname{Var}(U) = \mathbb{E}\,U^2 - (\mathbb{E}\,U)^2$ with $\mathbb{E}[z_a z_b] = \Sigma_{ab}$.
(3) Isserlis with indices (a, b, a, b): the pairings give $\Sigma_{ab}\Sigma_{ab}$, $\Sigma_{aa}\Sigma_{bb}$ and $\Sigma_{ab}\Sigma_{ba}$.

For $a = b$ this is $2\Sigma_{aa}^2/(n-1)$, the Gaussian variance of S^2 in (3.34). For $\Sigma = \mathbb{1}$ and $a \neq b$ it is $1/(n-1)$, the scatter of spurious correlations found above. Each element of a covariance matrix estimated from $n = 50$ simulations is uncertain by about $1/\sqrt{49} = 14\%$ of the typical diagonal size.

Inverting a noisy matrix. The χ^2 needs $\mathbf{C}^{-1}$, and here noise does more than add scatter: it biases the result. The reason is that $1/x$ is a convex function. For a single number, Jensen's inequality gives $\mathbb{E}[1/\hat{\sigma}^2] > 1/\mathbb{E}[\hat{\sigma}^2]$: a fluctuation of $\hat{\sigma}^2$ towards zero raises $1/\hat{\sigma}^2$ by more than an equal fluctuation upwards lowers it. For a matrix, the noise spreads the eigenvalues of $\hat{\mathbf{C}}$ apart. The smallest ones come out too small, and inverting turns them into eigenvalues of $\hat{\mathbf{C}}^{-1}$ that are too large. For Gaussian readings and $n > p + 2$ the exact result is

$$\mathbb{E}\big[\hat{\mathbf{C}}^{-1}\big] = \frac{n-1}{n-p-2}\,\mathbf{C}^{-1}, \tag{3.43}$$

a property of the inverse-Wishart distribution. The unbiased estimate of the inverse covariance is therefore $\widehat{\mathbf{C}^{-1}} = \frac{n-p-2}{n-1}\hat{\mathbf{C}}^{-1}$, the **Hartlap factor** [6]. As p approaches n the bias explodes, and at $p \geq n$ the matrix $\hat{\mathbf{C}}$ is singular (its rank is at most $n-1$).

Example: How many simulations? With $n = 30$ simulations and $p = 10$ bins, $(n-1)/(n-p-2) = 29/18 = 1.611$: without the correction, every χ^2 is too large by 61%, and parameter error bars come out too small by a factor of about $\sqrt{1.61} = 1.27$. The simulation gives 1.616. To keep the bias below 10% we need $(n-1)/(n-p-2) \leq 1.1$. From $n - 1 \leq 1.1(n-p-2)$ we get $0.1\,n \geq 1.1p + 1.2$, i.e. $n \geq 11p + 12$: about 120 simulations for 10 bins. Even then, the *noise* in the inverse (not just its bias) widens the parameter constraints, a further effect that has to be propagated.

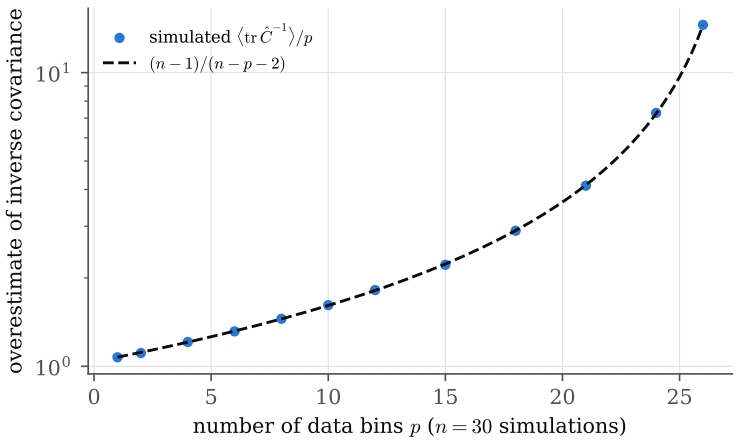


Figure 3.20: How much the inverse of a sample covariance matrix overestimates the true inverse, for thirty simulations and a growing number of bins (logarithmic vertical axis). The simulated overestimate follows $(n-1)/(n-p-2)$ (dashed): small for a few bins, a factor of two by $p \approx 14$, and diverging as p approaches $n-2$.

Watch out: Selected and misread correlations

A correlation coefficient estimated from the same data that were used to choose which pair to look at is not distributed as (3.41). Nor is a correlation evidence of a causal link, and a zero correlation does not mean independence (the correlation only sees the linear part of a relation, Chapter 2).

What this bought us. A covariance matrix estimated from n samples has element-wise noise of relative size $\sim 1/\sqrt{n-1}$, which shows up as spurious correlations of that size. Put error bars on a correlation in the variable $z = \operatorname{artanh} r$, where the width is $1/\sqrt{n-3}$. Before inverting an estimated covariance, multiply the inverse by the Hartlap factor $(n-p-2)/(n-1)$, and choose n well above p.

Verification and research. Everything above was verified against formulas valid for Gaussian readings. For band powers from masked skies or non-Gaussian fields, neither (3.42) nor (3.43) is exact, and the number of simulations needed for a stable χ^2 is itself found by simulation: rerun the covariance estimate with independent sets of simulations and watch how the parameter constraints move (Chapter 8).

3.6 Propagating errors through a formula

The pendulum student now measures the length, $L = 1.000 \pm 0.002\,\mathrm{m}$, and the period, $T = 2.006 \pm 0.004\,\mathrm{s}$, and computes $g = 4\pi^2 L/T^2$. What is the error on g? A particle physicist asks the same about a branching ratio $N_{\mathrm{signal}}/N_{\mathrm{norm}}$, a cosmologist about H_0 derived from a measured angle and a sound horizon. And sometimes the textbook answer is badly wrong: a ratio whose denominator is poorly measured has a distribution with tails that no Gaussian error bar describes.

Faced with a formula and a few uncertain inputs, the tempting non-answer is the one in Figure 3.21: we are not sure how big the error on each point is, so we give the error bar an error bar of its own, and that one an error bar too. It never ends in a number. The way out is to notice that the uncertainty of g is not a new, separate unknown. It is the uncertainty of L and T, carried once through the formula $g = 4\pi^2 L/T^2$. The rest of this section shows how.

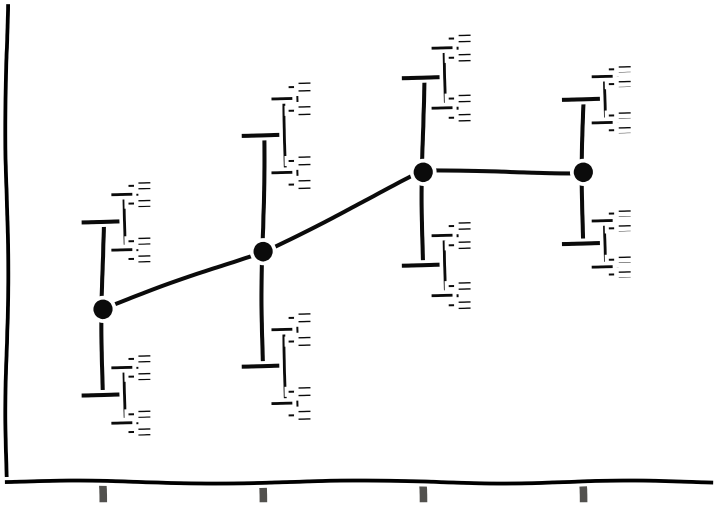


Figure 3.21: What not propagating an error looks like: every error bar gets error bars of its own, three levels deep, and the picture never settles to one number per point.

The standard answer replaces the formula by its tangent, which gives the output covariance as a "Jacobian sandwich" $A\mathbf{C}A^{\mathsf{T}}$ and shows exactly where correlations between the inputs enter. The answer is exact for linear formulas, and right for large samples (the delta method). Where it fails, a Monte Carlo propagation still works.

Every smooth function is linear up close. Zoom in on a smooth curve and it looks like its tangent line. If the cloud of possible input values is small compared with the scale on which the function bends, then over that cloud the function is effectively linear, and a linear map sends a Gaussian to a Gaussian, stretched by the slope. When the cloud is wide enough to feel the curvature, the output is skewed and its centre shifts. Near a pole, like $1/x$ at $x = 0$, all bets are off.

What we want to know.

1. Given the means and covariance of the inputs, what are the mean and covariance of the outputs, to the first order in the fluctuations?

2. When is that answer exact, and when is it an approximation?

3. How do correlations between inputs change the result?
4. How can we tell that the approximation has failed, and what then?

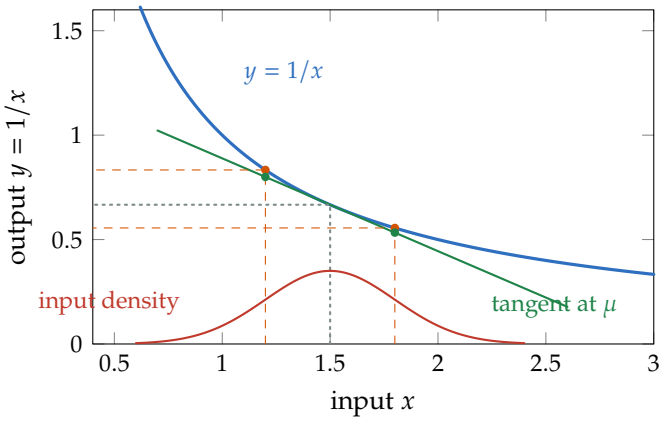


Figure 3.22: Linear propagation through $y = 1/x$ with input $\mu = 1.5$, $\sigma = 0.3$. The red curve is the input density. Orange dots: the true images of $\mu \pm \sigma$, at 0.556 and 0.833, lie asymmetrically about the image 0.667 of the mean (dotted). Green dots: the tangent line puts them symmetrically at 0.533 and 0.800. The curvature makes the output skewed towards large y, which is what the Monte Carlo shows later.

3.6.1 The first-order formula

Let the inputs $\boldsymbol{x} = (x_1, \dots, x_n)$ have means $\boldsymbol{\mu}$ and covariance matrix $V_{ij} = \operatorname{Cov}(x_i, x_j)$, and let $y(\boldsymbol{x})$ be a smooth function. Expand about the means,

$$y(\boldsymbol{x}) = y(\boldsymbol{\mu}) + \sum_{i=1}^{n} \left.\frac{\partial y}{\partial x_i}\right|_{\boldsymbol{\mu}} (x_i - \mu_i) + \frac{1}{2} \sum_{i,j} \left.\frac{\partial^2 y}{\partial x_i \partial x_j}\right|_{\boldsymbol{\mu}} (x_i - \mu_i)(x_j - \mu_j) + \cdots , \tag{3.44}$$

If the inputs scatter only a little about their means, the quadratic and higher terms are small, so we keep only the linear term. Write $g_i = \partial y / \partial x_i$ at $\boldsymbol{\mu}$. Then, to this order,

$$\mathbb{E}\, y \overset{(1)}{=} y(\boldsymbol{\mu}) + \sum_i g_i \, \mathbb{E}(x_i - \mu_i) \overset{(2)}{=} y(\boldsymbol{\mu}), \tag{3.45}$$

$$\sigma_y^2 = \mathbb{E}\big[(y - \mathbb{E}\, y)^2\big] \overset{(3)}{=} \mathbb{E}\Big[\Big(\sum_i g_i (x_i - \mu_i)\Big)^2\Big] \overset{(4)}{=} \sum_{i,j} g_i g_j \, \mathbb{E}\big[(x_i - \mu_i)(x_j - \mu_j)\big] \overset{(5)}{=} \sum_{i,j} g_i \, V_{ij} \, g_j . \tag{3.46}$$

(1) Take the expectation of the linearised (3.44). The g_i are constants.
(2) Each $\mathbb{E}(x_i - \mu_i) = 0$.
(3) Subtract (3.45) from the linearised y.
(4) Square of a sum is a double sum, then linearity.
(5) Definition of the covariance.

For several outputs $y_1, \dots, y_m$, the same steps with two different outputs y_k, y_l give the covariance matrix of the outputs,

$$U_{kl} = \operatorname{Cov}(y_k, y_l) \approx \sum_{i,j} \frac{\partial y_k}{\partial x_i} \frac{\partial y_l}{\partial x_j} V_{ij}, \qquad U = A\, V A^{\mathsf{T}}, \qquad A_{ki} = \left.\frac{\partial y_k}{\partial x_i}\right|_{\boldsymbol{\mu}}, \tag{3.47}$$

where A is the $m \times n$ **Jacobian** matrix [Cowan (1.53)–(1.56)].

Working definition 3.8 (linear error propagation)

Linear (first-order) **error propagation** replaces each function by its tangent plane at the input means [the first two terms of the Taylor series, valid when the inputs scatter over a region where the second derivatives hardly matter], so that the output covariance is $U = AVA^{\mathsf{T}}$ with A the Jacobian at the means. If the inputs are uncorrelated, V is diagonal and $\sigma_y^2 = \sum_i (\partial y/\partial x_i)^2 \sigma_i^2$ [Cowan (1.57)].

Four rules everybody uses. With two inputs and $V_{12} = \rho\sigma_1\sigma_2$, (3.46) gives [Cowan (1.59)–(1.60)]:

$$y = x_1 \pm x_2: \quad \sigma_y^2 = \sigma_1^2 + \sigma_2^2 \pm 2\rho\sigma_1\sigma_2, \tag{3.48}$$

$$y = x_1 x_2 \text{ or } x_1/x_2: \quad \frac{\sigma_y^2}{y^2} = \frac{\sigma_1^2}{x_1^2} + \frac{\sigma_2^2}{x_2^2} \pm \frac{2\rho\sigma_1\sigma_2}{x_1 x_2}, \tag{3.49}$$

$$y = x^k: \quad \frac{\sigma_y}{|y|} = |k|\,\frac{\sigma_x}{|x|}, \tag{3.50}$$

$$y = \ln x: \quad \sigma_y = \frac{\sigma_x}{x}. \tag{3.51}$$

(For the product take +, for the ratio −. The derivatives are $\partial(x_1x_2)/\partial x_1 = x_2$, $\partial(x_1/x_2)/\partial x_2 = -x_1/x_2^2$, and dividing by y^2 turns them into relative errors.) Absolute errors add in quadrature for sums, relative errors for products and ratios.

Example: The pendulum. $g = 4\pi^2 L T^{-2}$ is a product of powers, so relative errors add in quadrature with weights $|k|$:

$$\frac{\sigma_g^2}{g^2} \overset{(1)}{=} \left(\frac{\sigma_L}{L}\right)^2 + \left(2\,\frac{\sigma_T}{T}\right)^2 \overset{(2)}{=} (0.002)^2 + \left(\frac{2\times 0.004}{2.006}\right)^2 = 4.0\times10^{-6} + 1.59\times10^{-5},$$

$$g \overset{(3)}{=} \frac{4\pi^2 \times 1.000}{2.006^2} = 9.811\ \mathrm{m\,s^{-2}}, \qquad \sigma_g = 9.811 \times \sqrt{1.99\times10^{-5}} = 0.04377\ \mathrm{m\,s^{-2}}.$$

(1) (3.49) and (3.50) with $k = 1$ for L and $k = -2$ for T. L and T are measured independently, so $\rho = 0$.
(2) Insert the numbers.
(3) The central value is the function at the input means (3.45).

The period dominates. Its relative error (0.20%) is the same as that of L, but it enters squared, so it contributes four times the variance. To improve g, time more swings.

Example: A common calibration: correlated errors. Two energies E_1, E_2 are measured with the same calorimeter. Each has an independent statistical error σ_s and they share one calibration error σ_c: $E_i = E_i^{\text{true}} + \epsilon_i + \delta$, where the ϵ_i are independent with mean 0 and variance σ_s^2, and the common shift δ, independent of both, has mean 0 and variance σ_c^2. Then $V_{11} = V_{22} = \sigma_s^2 + \sigma_c^2$ and $V_{12} = \mathrm{Cov}(\epsilon_1 + \delta, \epsilon_2 + \delta) = \sigma_c^2$. For the difference and the sum, (3.48) gives

$$\mathrm{Var}(E_1 - E_2) \overset{(1)}{=} 2(\sigma_s^2 + \sigma_c^2) - 2\sigma_c^2 = 2\sigma_s^2, \qquad \mathrm{Var}(E_1 + E_2) \overset{(2)}{=} 2(\sigma_s^2 + \sigma_c^2) + 2\sigma_c^2 = 2\sigma_s^2 + 4\sigma_c^2.$$

(1) Minus sign: the common shift δ cancels in a difference.
(2) Plus sign: it adds coherently in a sum, so its standard deviation doubles.

Ignoring the correlation would overstate the error of a mass difference and understate the error of a total energy. This is why experiments publish full covariance matrices, not just error bars.

3.6.2 Linear maps are exact, and every covariance can be diagonalised

If $\boldsymbol{y} = A\boldsymbol{x}$ exactly (a linear map, as in a change of basis, a rebinning, or a sum of channels), there is no Taylor remainder, and

$$U_{kl} = \operatorname{Cov}(y_k, y_l) \overset{(1)}{=} \mathbb{E}\Big[\sum_i A_{ki}(x_i - \mu_i) \sum_j A_{lj}(x_j - \mu_j)\Big] \overset{(2)}{=} \sum_{i,j} A_{ki} V_{ij} A_{lj} = (AVA^{\mathsf{T}})_{kl} \tag{3.52}$$

(1) $y_k - \mathbb{E}\, y_k = \sum_i A_{ki}(x_i - \mu_i)$ by linearity, with no approximation.
(2) Multiply out and use linearity of expectation.

holds exactly [Cowan (1.62)]. A linear map of a Gaussian vector is Gaussian, so for Gaussian inputs the whole distribution of $\boldsymbol{y}$ is $\mathcal{N}(A\boldsymbol{\mu}, AVA^{\mathsf{T}})$.

Exactness has a useful consequence: a well-chosen linear map removes all correlations. V is real and symmetric, so it has n orthonormal eigenvectors $\boldsymbol{r}^{(k)}$ with $V\boldsymbol{r}^{(k)} = \lambda_k \boldsymbol{r}^{(k)}$ and $\boldsymbol{r}^{(k)} \cdot \boldsymbol{r}^{(l)} = \delta_{kl}$. Take the rows of A to be these eigenvectors, $A_{ki} = r_i^{(k)}$. Then

$$U_{kl} \overset{(1)}{=} \sum_{i,j} r_i^{(k)} V_{ij} r_j^{(l)} \overset{(2)}{=} \sum_i r_i^{(k)} \lambda_l \, r_i^{(l)} \overset{(3)}{=} \lambda_l \, \delta_{kl}. \tag{3.53}$$

(1) (3.52) with $A_{ki} = r_i^{(k)}$.
(2) $\sum_j V_{ij} r_j^{(l)} = \lambda_l r_i^{(l)}$, the eigenvalue equation.
(3) Orthonormality of the eigenvectors.

The new variables $y_k = \boldsymbol{r}^{(k)} \cdot \boldsymbol{x}$ are uncorrelated, with variances equal to the eigenvalues of V [Cowan (1.63)–(1.65)]. This is the statistics version of finding normal modes of coupled oscillators, and of the principal axes of an inertia tensor.

Example: Decorrelating two measurements. For $V = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ the eigenvectors are $(1,1)/\sqrt{2}$ and $(1,-1)/\sqrt{2}$ with eigenvalues $\sigma^2(1 \pm \rho)$ (check: $V(1,1)^{\mathsf{T}} = \sigma^2(1+\rho)(1,1)^{\mathsf{T}}$). The combinations $(x_1 + x_2)/\sqrt{2}$ and $(x_1 - x_2)/\sqrt{2}$ are uncorrelated, and with ρ close to 1 the difference is much better measured than either input. The calibration example above is this with $\rho = \sigma_c^2/(\sigma_s^2 + \sigma_c^2)$.

3.6.3 The delta method: why the first-order formula is right for large samples

Linear propagation is an approximation, so when can we trust it? When the inputs are themselves estimators built from n readings. Their scatter then shrinks like $1/\sqrt{n}$, the cloud of inputs shrinks onto the tangent, and the curvature terms become negligible. The delta method makes this precise.

The delta method. If $\sqrt{n}(Y_n - \mu) \rightsquigarrow \mathcal{N}(0, \sigma^2)$ and g is differentiable with $g'(\mu) \neq 0$, then

$$\sqrt{n}\big(g(Y_n) - g(\mu)\big) \rightsquigarrow \mathcal{N}\big(0,\ g'(\mu)^2\sigma^2\big), \qquad \text{i.e.} \qquad g(Y_n) \approx \mathcal{N}\Big(g(\mu),\ \frac{g'(\mu)^2\sigma^2}{n}\Big). \tag{3.54}$$

For a vector, $\sqrt{n}\big(g(\boldsymbol{Y}_n) - g(\boldsymbol{\mu})\big) \rightsquigarrow \mathcal{N}\big(0, \nabla g^{\mathsf{T}} \boldsymbol{\Sigma} \nabla g\big)$ [Wasserman Thms. 5.13, 5.15].

Why it is true: expand $g(Y_n) = g(\mu) + g'(\mu)(Y_n - \mu) + R_n$ with a remainder R_n of order $(Y_n - \mu)^2$. Multiply by $\sqrt{n}$. The linear term $g'(\mu)\sqrt{n}(Y_n - \mu)$ tends in distribution to $\mathcal{N}(0, g'(\mu)^2\sigma^2)$. The remainder is $\sqrt{n}\, R_n \sim \sqrt{n}\,(Y_n - \mu)^2 = [\sqrt{n}(Y_n - \mu)]^2/\sqrt{n}$, a quantity of fixed size divided by $\sqrt{n}$, so it tends to zero in probability, and Slutsky's theorem lets us drop it. The curvature does not disappear. It becomes negligible compared with the shrinking linear scatter.

Example: A function of one mean, and the product of two. In both cases the inputs are sample means, so by the central limit theorem $\sqrt{n}(\bar{X}_n - \mu) \rightsquigarrow \mathcal{N}(0, \sigma^2)$ and the delta method applies. (a) For $W_n = e^{\bar{X}_n}$, $g(s) = e^s$ has $g'(\mu) = e^\mu$, so $W_n \approx \mathcal{N}(e^\mu, e^{2\mu}\sigma^2/n)$ [Wasserman Ex. 5.14]. (b) For the product of two means, $Y = \bar{X}_1\bar{X}_2$ from n paired readings with covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ab})$, the gradient of $g(s_1, s_2) = s_1 s_2$ at $\boldsymbol{\mu}$ is (μ_2, μ_1) and

$$\nabla g^{\mathsf{T}} \boldsymbol{\Sigma} \nabla g \overset{(1)}{=} \begin{pmatrix} \mu_2 & \mu_1 \end{pmatrix} \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} \begin{pmatrix} \mu_2 \\ \mu_1 \end{pmatrix} \overset{(2)}{=} \mu_2^2\sigma_{11} + 2\mu_1\mu_2\sigma_{12} + \mu_1^2\sigma_{22},$$

(1) $\partial(s_1 s_2)/\partial s_1 = s_2$ and $\partial(s_1 s_2)/\partial s_2 = s_1$, at the means.
(2) Multiply out the quadratic form.

so $\sqrt{n}(\bar{X}_1\bar{X}_2 - \mu_1\mu_2) \rightsquigarrow \mathcal{N}(0, \mu_2^2\sigma_{11} + 2\mu_1\mu_2\sigma_{12} + \mu_1^2\sigma_{22})$ [Wasserman Ex. 5.16]. Dividing by $(\mu_1\mu_2)^2$ this is the product rule (3.49).

When the slope vanishes. If $g'(\mu) = 0$, the first-order formula says $\sigma_y = 0$, which is absurd. The second-order term takes over [CasellaBerger Thm. 5.5.26]: with $g(Y_n) - g(\mu) \approx \frac{1}{2}g''(\mu)(Y_n - \mu)^2$ and $n(Y_n - \mu)^2/\sigma^2 \rightsquigarrow \chi^2_1$,

$$n\big(g(Y_n) - g(\mu)\big) \rightsquigarrow \frac{\sigma^2 g''(\mu)}{2}\,\chi^2_1. \tag{3.55}$$

Estimating a power $|A|^2$ when the true amplitude A is zero is exactly this case. With $g(s) = s^2$ we have $g'(0) = 0$ and $g'' = 2$, so the estimate is $\sigma^2\chi^2_1/n$: never negative, and biased upward by σ^2/n (the mean of χ^2_1 is 1).

The second-order bias. Keeping the quadratic term of (3.44) in the mean gives $\mathbb{E}\,y \approx y(\boldsymbol{\mu}) + \frac{1}{2}\sum_{ij}\partial_i\partial_j y\, V_{ij}$. For $y = 1/x$ this is $\mathbb{E}(1/x) \approx \frac{1}{\mu}\big(1 + \sigma^2/\mu^2\big)$, always above $1/\mu$, as Figure 3.22 suggests. For the ratio of means $\bar{X}/\bar{Y}$ Casella and Berger derive in the same way [CasellaBerger Ex. 5.5.27]

$$\mathbb{E}\,\frac{\bar{X}}{\bar{Y}} \approx \frac{\mu_X}{\mu_Y}, \qquad \operatorname{Var}\frac{\bar{X}}{\bar{Y}} \approx \frac{1}{n}\,\frac{\mu_X^2}{\mu_Y^2}\Big(\frac{\sigma_X^2}{\mu_X^2} + \frac{\sigma_Y^2}{\mu_Y^2} - \frac{2\sigma_{XY}}{\mu_X\mu_Y}\Big),$$

the ratio rule (3.49) with $\sigma^2 \to \sigma^2/n$, and they warn that it can be far off when μ_Y is near zero.

3.6.4 When linear propagation fails: the ratio of Gaussians

Let $x_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $x_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ be independent and $y = x_1/x_2$. The linear rule gives a Gaussian of width $\sigma_y = |y_0|\sqrt{(\sigma_1/\mu_1)^2 + (\sigma_2/\mu_2)^2}$ around $y_0 = \mu_1/\mu_2$. But the denominator has a nonzero density at $x_2 = 0$, where y diverges, so y has no mean and no variance at all, however small σ_2/μ_2 is. The extreme case, both means zero, shows where the heavy tails come from.

Example: The ratio of two zero-mean Gaussians is a Breit–Wigner. Let $x_1, x_2 \sim \mathcal{N}(0, 1)$ independently. The pair (x_2, x_1) is an isotropic Gaussian in the plane, so its polar angle Θ is uniformly distributed, and $y = x_1/x_2 = \tan\Theta$. Since tan has period π, we can take Θ uniform on $(-\pi/2, \pi/2)$. Then

$$\mathrm{P}(y \le t) \overset{(1)}{=} \mathrm{P}(\Theta \le \arctan t) \overset{(2)}{=} \frac{\arctan t + \pi/2}{\pi}, \qquad f(t) \overset{(3)}{=} \frac{1}{\pi(1 + t^2)}.$$

(1) tan is increasing on $(-\pi/2, \pi/2)$.

(2) Uniform distribution on an interval of length π.
(3) Differentiate: $\frac{\mathrm{d}}{\mathrm{d}t}\arctan t = 1/(1+t^2)$.

This is the Cauchy (Breit–Wigner) distribution of Section 3.3.4 [CasellaBerger §5.5]. For $\mu_2 \neq 0$ the ratio keeps a Cauchy-like $1/y^2$ tail. Its weight is set by how much probability x_2 has near zero.

Listing 3.6: Monte Carlo propagation for a ratio of Gaussians (`code/ch03/06_error_propagation.py`)

```
# ---------- (ii) ratio of Gaussians ----------
mu1, s1, s2 = 10.0, 1.0, 1.0
fig, axs = plt.subplots(1, 3, figsize=(8.0, 2.7))
out = {}
for ax, mu2, c, tag in zip(axs, [10.0, 3.0, 1.0], SERIES, ["A", "B", "C"]):
    x1 = mu1 + s1 * rng.standard_normal(M)
    x2 = mu2 + s2 * rng.standard_normal(M)
    y = x1 / x2
    y0 = mu1 / mu2
    sy = abs(y0) * np.sqrt((s1 / mu1) ** 2 + (s2 / mu2) ** 2)
    lo, med, hi = np.percentile(y, [15.865, 50, 84.135])
    out[tag] = dict(y0=y0, sy=sy, med=med, half=0.5 * (hi - lo),
                    neg=np.mean(y < 0), far=np.mean(np.abs(y - y0) > 5 * sy))
```

The script takes $\mu_1 = 10$, $\sigma_1 = \sigma_2 = 1$ and three denominators, $\mu_2 = 10, 3, 1$ (relative error 0.1, 0.33, 1). For each it draws 10^6 pairs, forms the ratio, and compares the Monte Carlo distribution with the linear prediction. Because the variance may not exist, it measures the width by the half-distance between the 15.9 and 84.1 percentiles (which is σ for a Gaussian), and it counts negative ratios and "5σ" outliers.

	$\sigma_2/\mu_2 = 0.1$	0.33	1
linear prediction $y_0 \pm \sigma_y$	1 ± 0.1414	3.33 ± 1.16	10 ± 10.05
Monte Carlo median	0.9999	3.329	7.083
Monte Carlo central-68% half-width	0.1425	1.294	11.49
fraction with $y < 0$	0	0.001363	0.1589
fraction beyond $5\sigma_y$ (Gaussian: 5.7×10^{-7})	1.600×10^{-4}	0.02883	0.0961

Table 3.1: Linear propagation against Monte Carlo for $y = x_1/x_2$.

At 10% relative error the linear rule is fine in the core (width 0.1425 against 0.1414) but already has 300 times too many 5σ outliers. At 33% the core is 12% wider than predicted and 3% of ratios are beyond "5σ". At 100% the linear answer is meaningless: the median is 7.083, not 10, and 0.1589 of the ratios are negative although both means are positive.

For the pendulum, by contrast, the Monte Carlo spread of g is 0.04375 against the linear 0.04377, and the mean is 9.811 (Figure 3.24).

Taxonomy: Which propagation to use

situation	method
linear function ($\boldsymbol{y} = A\boldsymbol{x}$)	$U = AVA^{\mathsf{T}}$, exact
small relative errors, smooth y	first-order formula; check one case by Monte Carlo
y depends on a quantity whose error is not small (ratio, $1/x$, $\ln x$ near 0)	Monte Carlo propagation: sample inputs, apply y, histogram; quote median and percentiles
$g'(\mu) = 0$ at the input means	second-order delta method, χ^2_1 shape
inputs from an MCMC chain	apply y to every sample of the chain (Chapter 9)

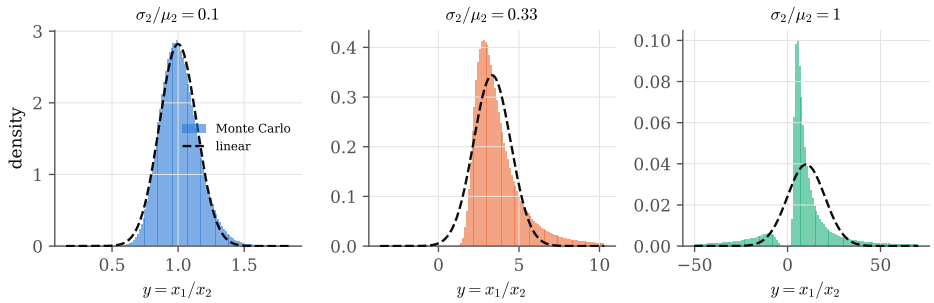


Figure 3.23: The ratio of two Gaussian quantities, simulated (histograms) and by linear propagation (dashed), as the denominator's relative error grows from left to right. At 10% the two agree. At 33% the true distribution is skewed, with a long right tail. At 100% it has a spike near zero, negative values and heavy tails, and bears no resemblance to the Gaussian.

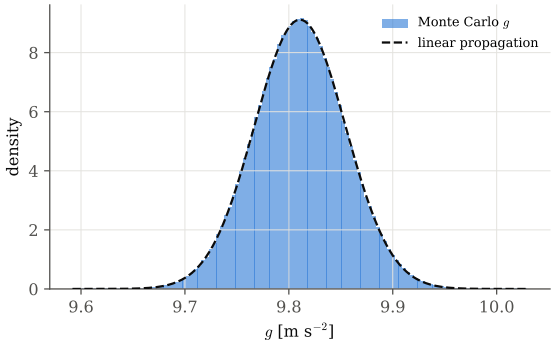


Figure 3.24: Monte Carlo propagation for the pendulum. A million pairs (L, T) drawn from their Gaussian errors give the histogram of g. The linear-propagation Gaussian (dashed) lies on top of it, because the relative errors of 0.2% are tiny compared with the scale on which L/T^2 bends.

Symmetric error bars on a ratio. An efficiency, a branching ratio or a signal-to-background ratio with a poorly known denominator should not be quoted as "central ± symmetric error". The median is shifted, the errors are asymmetric, and the tails are long. Quote percentiles from a Monte Carlo propagation or, better, build a likelihood for the ratio directly (second half of Chapter 3), as Cowan recommends [Cowan p. 22].

What this bought us. To first order, covariances propagate as $U = AVA^{\mathsf{T}}$ with the Jacobian A: absolute errors add in quadrature for sums, relative errors for products, and correlations add $\pm 2\rho\sigma_1\sigma_2$ terms. The rule is exact for linear maps and asymptotically right for smooth functions of estimators (delta method). It fails when the input scatter reaches a region of strong curvature or a pole, and then Monte Carlo propagation is the answer.

How this is used in real research: Verification and research

For the pendulum, Monte Carlo propagation merely verifies the linear formula. For a derived cosmological parameter (the age of the Universe from H_0, Ω_m and Ω_Λ; $S_8 = \sigma_8\sqrt{\Omega_m/0.3}$), the same idea is the working tool: take every sample of a posterior chain, compute the derived quantity, and histogram it. No Jacobian is needed, and correlations and nonlinearity are handled automatically (Chapter 9).

3.7 Combining measurements of different quality: the weighted mean

Eight laboratories have measured the same quantity μ: a particle mass, a lifetime, the Hubble constant. Their error bars range from 0.5 to 3 in some unit. A reviewer asks for one number. The plain average treats the best laboratory and the worst on an equal footing, and that feels wrong. A measurement with a small error bar should count for more. But how much more, and how small is the error of the combination?

One idea settles both questions with the tools we already have. The combination is again an estimator, a recipe applied to the data, so it has a sampling distribution, a bias and a variance (Sections 3.1 and 3.2). Among all unbiased linear recipes we look for the one with the smallest variance. The answer is the **inverse-variance weighted mean**, and the same answer drops out of the Gaussian likelihood, which is our first glimpse of the second half of the chapter. The questions, in order, are these.

1. Which weights w_i make $\sum_i w_i x_i$ unbiased with the smallest variance, and what is that variance?
2. How can we check that the measurements agree well enough for the combination to mean something?
3. What changes when the measurements are correlated, for example because they share a systematic error?
4. What if the readings also contain something we do not want, which enters each of them differently? How do we remove it?

When the readings carry more than the signal. The laboratories differ only in how noisy they are. Readings can also differ in what they contain. At the start of the chapter we wrote every reading as a piece that responds to what we want, plus everything else (3.1). So far, "everything else" has been noise of a known kind: it scatters around zero with a spread we can describe, and it is different in every reading. For that we already have good tools. We average, the noise partly cancels, and the error bar shrinks like $1/\sqrt{N}$.

Sometimes they carry something more than noise. Picture yourself at a noisy party with two microphones, trying to record a friend. She stands between them, so both pick up her voice equally loudly. A loudspeaker plays music, and it stands closer to the second microphone, so there the music is twice as loud. Averaging the two recordings does not help, because the music is not random noise that averages away: it is the same song in both, only louder in one. Yet there is a trick. Double the first recording and subtract the second. The music, twice in each, cancels. The voice, twice minus once, survives. Noise-cancelling headphones work on exactly this principle.

The sky does the same thing to an astronomer. The microwave background has the same brightness at two observing frequencies, while the dust of our own Galaxy is much brighter at the higher one. A calibration drift, a contaminating source, a foreground: whenever the unwanted part has a pattern of its own across the readings, different from the pattern of what we want, we can weight the readings so that the unwanted pattern cancels and the wanted one survives. Each reading gets its own weight, $F = \sum_i w_i x_i$, and the question is which weights to use.

The party in symbols. Call the voice (or the sky signal) s and the music (or the dust) f. It is twice as strong in the second reading, and each reading has its own noise. Combine them with

two numbers w_1, w_2:

$$x_1 = s + f + n_1, \qquad x_2 = s + 2f + n_2,$$
$$w_1 x_1 + w_2 x_2 \overset{(1)}{=} (w_1 + w_2)\, s + (w_1 + 2w_2)\, f + (w_1 n_1 + w_2 n_2). \tag{3.56}$$

(1) Multiply out and collect the terms in s, in f and in the noise.

Each bracket is a dial. The first sets the *normalisation*: unless $w_1 + w_2 = 1$ we are estimating a rescaled signal. The plain difference $x_2 - x_1 = f + n_2 - n_1$ shows how badly this can go: its weights sum to zero, so the signal cancels completely and what is left measures the foreground. The second bracket sets the *contamination*: $w_1 + 2w_2 = 0$ removes f whatever its value. Both conditions together fix the weights, $w_1 = 2$, $w_2 = -1$, and

$$2x_1 - x_2 \overset{(1)}{=} s + (2n_1 - n_2), \qquad \mathrm{Var}(2x_1 - x_2) \overset{(2)}{=} 4\sigma_1^2 + \sigma_2^2. \tag{3.57}$$

(1) (3.56) with $w_1 + w_2 = 1$ and $w_1 + 2w_2 = 0$.
(2) Independent noises: variances add, each weighted by w_i^2 (3.19).

The foreground is gone exactly, the signal keeps its unit normalisation, and the price is noise: with $\sigma_1 = \sigma_2 = \sigma$ the clean combination has variance $5\sigma^2$, ten times the $\sigma^2/2$ of the plain average $(x_1 + x_2)/2$, which in exchange keeps $1.5f$ of foreground. Weights that cancel unwanted pieces are easy to find; weights that also keep the noise small are a trade-off. Lesson: a linear combination can switch off anything whose response across the readings differs from that of s, while one constraint, $\sum_i w_i = 1$ for a signal that enters every reading equally, keeps s itself intact; the rest is the choice of the least noisy weights among those allowed.

A picture: weights point in a direction. Write the two readings as one list, $\boldsymbol{x} = (x_1, x_2)^\mathsf{T}$. The signal enters both readings equally, with the pattern $\boldsymbol{a} = (1, 1)^\mathsf{T}$; the dust enters with the pattern $\boldsymbol{c} = (1, 2)^\mathsf{T}$. The weights are a third list, $\boldsymbol{w} = (w_1, w_2)^\mathsf{T}$, and a weighted sum multiplies matching entries and adds them up. That is a dot product, so the readings and the two dials become

$$\boldsymbol{x} = \boldsymbol{a}\, s + \boldsymbol{c}\, f + \boldsymbol{n} \qquad \Longrightarrow \qquad \boldsymbol{w}^\mathsf{T}\boldsymbol{x} \overset{(1)}{=} (\boldsymbol{w}^\mathsf{T}\boldsymbol{a})\, s + (\boldsymbol{w}^\mathsf{T}\boldsymbol{c})\, f + \boldsymbol{w}^\mathsf{T}\boldsymbol{n}. \tag{3.58}$$

(1) Multiply each term by $\boldsymbol{w}^\mathsf{T}$; a number in front of a list comes out of the dot product.

For the weights $\boldsymbol{w} = (2, -1)^\mathsf{T}$ the signal gives $2 \cdot 1 - 1 \cdot 1 = 1$ and survives; the dust gives $2 \cdot 1 - 1 \cdot 2 = 0$ and vanishes. In the language of geometry, $\boldsymbol{w}$ points at right angles to the dust's pattern, so it is blind to the dust however bright it is.

Figure 3.25 shows what the weighted sum does to a point. Lay a ruler along $\boldsymbol{w}$, marked so that it reads $\boldsymbol{w}^\mathsf{T}\boldsymbol{x}$, and drop the data straight down onto it, at right angles. Where the point lands is the reading. Because the ruler stands at right angles to everything else, sliding the point in that direction does not move its shadow: the data and the signal alone land on the same mark. Only the signal moves the shadow along the ruler.

So $\boldsymbol{w}$ is a filter. It asks one question of the data: how much of the pattern we want is here, ignoring everything else? Notice that it does not point along the signal. It points where it must to be blind to everything else, so its direction is set as much by what we ignore as by what we want; only when the rest is plain noise does it line up with the signal itself. Choosing weights is choosing a direction: one that sees what we want and is blind to what we do not. With more readings there is room to be blind to more unwanted patterns at once, and finding the weights is a small set of linear equations. When the unwanted patterns are not known exactly, or the noise matters as much as they do, we look instead for the direction along which the combination scatters least, and that search is the rest of this section.

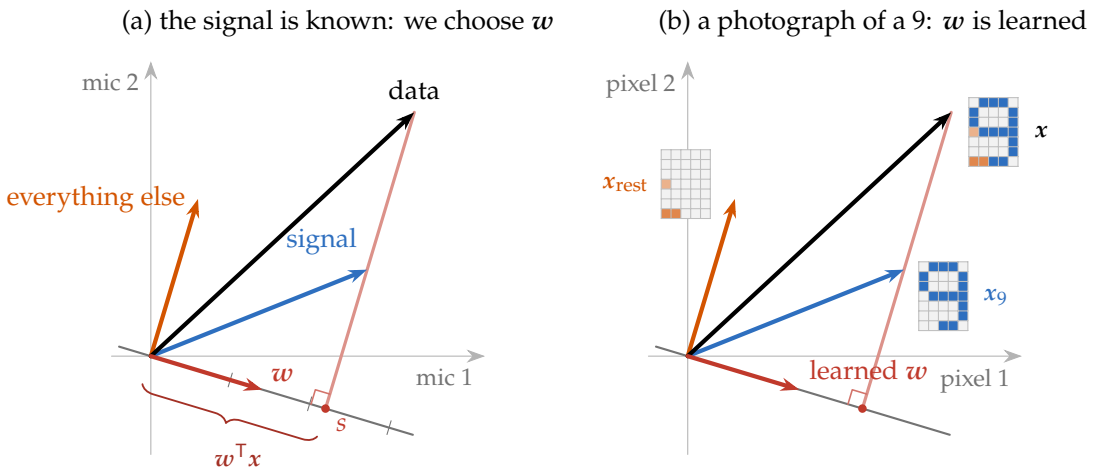


Figure 3.25: Choosing w when the signal is known, and learning it when it is not. (a) The data are the signal plus everything else. The weights w point along a ruler at right angles to everything else, and the reading $w^\top x$ is where the data fall when dropped straight onto it; the tip of the signal lies on the same drop line, so the data and the signal alone give the same reading. The two patterns are drawn further apart than at the party, so that the angles are easy to see. (b) The same split for a photograph of a 9. A picture is a long list of pixel values, so it is a vector too, drawn here in two of its many directions: the pixels of the written 9 are the signal, the paper, a smudge and uneven light are everything else, and the photograph is their sum, as in the first row of Figure 3.26. A network is not told this split; training finds weights whose reading keeps the 9 and ignores the rest, and so decides what counts as the signal.

The universal principle. The two microphones did one thing: each recording was the voice plus everything else, and the weights kept the voice and dropped the rest. Every weighted sum in this section does the same, and so, it turns out, does almost every analysis in the book. What makes it possible is a **feature**: a property that the part we want has and the rest does not. At the party it is how loud each source is in each microphone: the voice equally loud in both, the music twice as loud in one. Building a quantity that keeps the wanted feature and drops the others is **feature extraction**, and the weighted sum is its simplest form. The principle has one hard limit: without a feature there is no separation. Put the loudspeaker right next to your friend, and voice and music arrive equally loud in both microphones; no weights can pull them apart, because weights can use a difference the data contain but cannot invent one.

The principle does not stop at simple problems; even an artificial neural network runs on it. Suppose someone has written the digit 9 on a sheet of paper, we photograph it, and a network must tell from the photograph which digit was written. The photograph is a grid of pixel values: the pixels of the written 9, plus the paper, a smudge, uneven light. A grid of numbers is a long list of numbers, so the photograph is a vector, and this split is the same as the microphones' (Figure 3.25b). A first layer of weighted sums extracts the feature we want, the pixels that belong to the number. Those pixels are in turn a few shapes, a loop and a stroke running down from it, plus details that do not matter, such as the size of the loop; the next layer extracts the shapes. The last layers turn the shapes into the answer: one loop alone is a 0, two loops stacked are an 8, a loop on top with a tail below is a 9. Each layer is a set of weighted sums, each bent by a fixed curve, and each takes what the layer before it kept as its new data. The only real difference from our two microphones is where the weights come from: a network learns them from many labelled examples, while we work ours out from what we know about the patterns and the noise. So even the most complicated network is a long chain of the step we have just taken: isolate one feature, then the next. Figure 3.26 writes it with the same bookkeeping as the party, $x = a\,s + c\,f + n$: at each step the data are split into the part we

want plus the rest, and weights keep the part we want.

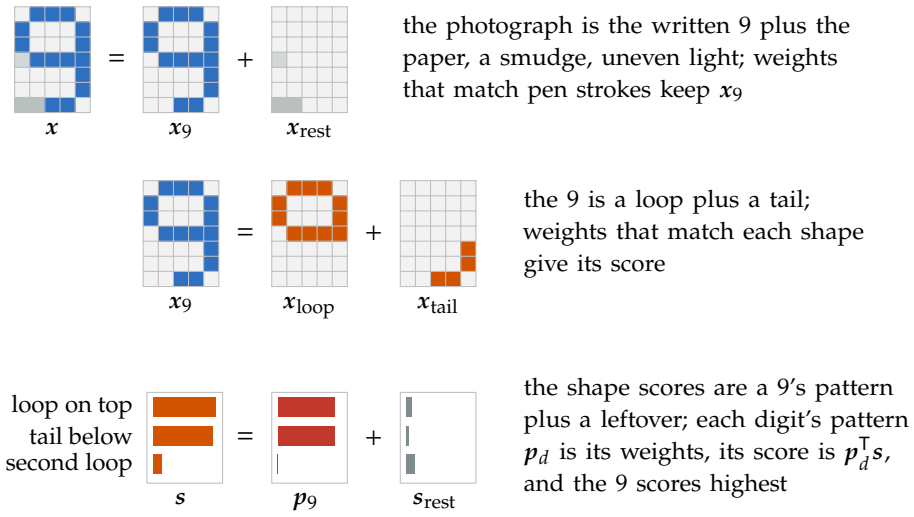


Figure 3.26: Reading a handwritten 9 with the idea of the two microphones, used again and again. Each row has the form *data = what we want + the rest*, exactly like $\boldsymbol{x} = \boldsymbol{a}\, s + \boldsymbol{c}\, f + \boldsymbol{n}$, and a set of weights keeps what we want: a weighted sum $\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}$ whose weights match the wanted pattern, bent by a fixed curve such as $\max(0, z)$. What one row keeps is the data of the next; in the last row the data are the shape scores, and each digit's weights are its own pattern of shapes. A neural network is this chain with many rows, and with weights learned from labelled examples.

The loop and the tail in the figure are our names, chosen so that the steps are easy to follow. A trained network need not split the 9 this way. Training fixes its weights, and the weights fix the split: at every layer, whatever the weights read becomes that layer's signal, and whatever they are blind to becomes its everything else, just as in Figure 3.25b. The pieces a real network ends up reading are often not shapes we would name. What carries over is the principle. Every layer splits its input into a part it keeps and a part it ignores, and it keeps that part with a weighted sum, the dot product of the two microphones, repeated many times and chained. Readings, patterns and weights are all lists of numbers, so the natural language for all of it is linear algebra.

We start with the simplest case: readings of different quality, no contaminant, and weights chosen so the combination scatters as little as possible. Section 3.7.4 then lets the unwanted pieces be shared between readings, and the same formula becomes the method used to separate the microwave background from the Galaxy.

3.7.1 The best linear combination

Let the measurements $x_1, \ldots, x_N$ be independent, each unbiased, $\mathbb{E}\, x_i = \mu$, with known standard deviations σ_i that differ from one measurement to the next. We try a *linear* recipe $\hat{\mu} = \sum_i w_i x_i$ with numbers w_i still to be chosen. Two properties follow at once:

$$\mathbb{E}\,\hat{\mu} \overset{(1)}{=} \sum_i w_i\, \mathbb{E}\, x_i \overset{(2)}{=} \mu \sum_i w_i, \tag{3.59}$$

$$\operatorname{Var}\hat{\mu} \overset{(3)}{=} \sum_i w_i^2\, \operatorname{Var} x_i \overset{(4)}{=} \sum_i w_i^2 \sigma_i^2. \tag{3.60}$$

(1) Linearity of expectation.
(2) Every measurement is unbiased, $\mathbb{E}\, x_i = \mu$.

(3) Variance of a sum (3.19): the covariances vanish because the x_i are independent, and a constant factor comes out squared, $\operatorname{Var}(wx) = w^2 \operatorname{Var} x$.
(4) $\operatorname{Var} x_i = \sigma_i^2$.

So $\hat{\mu}$ is unbiased for every value of μ exactly when the weights sum to one, $\sum_i w_i = 1$. Its mean squared error is then just its variance (3.13), and we want the weights that make $\sum_i w_i^2 \sigma_i^2$ as small as possible under that one constraint.

Minimising with a Lagrange multiplier. Recall the method from mechanics: to minimise $f(\boldsymbol{w})$ subject to $g(\boldsymbol{w}) = 0$, look for points where the gradient of f is parallel to the gradient of g, that is, where the gradient of $f - \lambda g$ vanishes for some number λ (the multiplier). Here $f = \sum_i w_i^2 \sigma_i^2$ and $g = \sum_i w_i - 1$:

$$
\begin{aligned}
0 &\overset{(1)}{=} \frac{\partial}{\partial w_k}\Big[\sum_i w_i^2\sigma_i^2 - \lambda\Big(\sum_i w_i - 1\Big)\Big] \overset{(2)}{=} 2w_k\sigma_k^2 - \lambda \quad\Longrightarrow\quad w_k = \frac{\lambda}{2\sigma_k^2}, \\
1 &\overset{(3)}{=} \sum_k w_k = \frac{\lambda}{2}\sum_k \frac{1}{\sigma_k^2} \quad\Longrightarrow\quad \frac{\lambda}{2} = \frac{1}{W}, \qquad W \equiv \sum_{k=1}^{N} \frac{1}{\sigma_k^2}, \\
w_k &\overset{(4)}{=} \frac{1/\sigma_k^2}{W}, \qquad \operatorname{Var}\hat{\mu} \overset{(5)}{=} \sum_k \frac{1}{\sigma_k^4 W^2}\,\sigma_k^2 = \frac{1}{W^2}\sum_k \frac{1}{\sigma_k^2} = \frac{1}{W}.
\end{aligned}
\tag{3.61}
$$

(1) Stationarity of $f - \lambda g$ in each w_k.
(2) Only the $i = k$ terms depend on w_k: $\partial(w_k^2\sigma_k^2)/\partial w_k = 2w_k\sigma_k^2$ and $\partial w_k/\partial w_k = 1$.
(3) Impose the constraint $\sum_k w_k = 1$ to fix λ.
(4) Insert $\lambda/2 = 1/W$ into $w_k = \lambda/(2\sigma_k^2)$.
(5) Insert these weights into (3.60).

A stationary point need not be a minimum. One inequality settles it: no other unbiased linear recipe does better. For any weights with $\sum_i w_i = 1$, the Cauchy–Schwarz inequality $(\sum_i a_i b_i)^2 \le \sum_i a_i^2 \sum_i b_i^2$ with $a_i = w_i\sigma_i$ and $b_i = 1/\sigma_i$ gives

$$
1 \overset{(1)}{=} \Big(\sum_i w_i\Big)^2 \overset{(2)}{=} \Big(\sum_i w_i\sigma_i \cdot \frac{1}{\sigma_i}\Big)^2 \overset{(3)}{\le} \sum_i w_i^2\sigma_i^2 \sum_i \frac{1}{\sigma_i^2} \overset{(4)}{=} \operatorname{Var}\hat{\mu}\cdot W. \tag{3.62}
$$

(1) The weights sum to one.
(2) Multiply and divide each term by σ_i.
(3) Cauchy–Schwarz. Equality holds only when $a_i \propto b_i$, that is $w_i\sigma_i \propto 1/\sigma_i$, or $w_i \propto 1/\sigma_i^2$.
(4) (3.60) and the definition of W.

So $\operatorname{Var}\hat{\mu} \ge 1/W$ for every unbiased linear combination, and the inverse-variance weights are the only ones that reach the bound.

Working definition 3.9 (inverse-variance weighted mean)

For independent unbiased measurements x_i of one quantity, with standard deviations σ_i,

$$
\hat{\mu} = \frac{\sum_i x_i/\sigma_i^2}{\sum_i 1/\sigma_i^2}, \qquad \sigma_{\hat{\mu}} = \Big(\sum_i \frac{1}{\sigma_i^2}\Big)^{-1/2}.
$$

Each measurement is weighted by its inverse variance $1/\sigma_i^2$ [a measurement twice as precise counts four times as much]. This is the unbiased linear combination with the smallest variance [Cowan (7.26)–(7.27)], [Sivia (2.30)–(2.31)].

What the formula says. Three consequences are worth reading off. First, since $W \geq 1/\sigma_k^2$ for every single k, the combined variance $1/W$ is never larger than the smallest individual variance: adding a measurement, however poor, never hurts [Cowan p. 107]. Second, if one measurement is much more precise than the rest, it dominates both W and the weighted sum, and the combination is essentially that measurement. Third, if all errors are equal, $\sigma_i = \sigma$, the weights are all $1/N$ and $\operatorname{Var}\hat{\mu} = \sigma^2/N$, the familiar variance of the sample mean (3.20).

Two pictures from physics make the formula feel familiar. Inverse variances add like the conductances of resistors in parallel, $1/R = \sum_i 1/R_i$: the combined "resistance to knowing μ" is smaller than the smallest single one. And $\hat{\mu}$ is the centre of mass of the points x_i with masses $1/\sigma_i^2$; a precise measurement is a heavy mass that drags the centre towards itself. Later we will call $1/\sigma^2$ the **information** carried by a measurement, and the statement "the information of independent measurements adds" becomes the additivity of Fisher matrices (Chapter 10).

The same answer from least squares and the likelihood. Suppose, in addition, that each x_i is Gaussian, $x_i \sim \mathcal{N}(\mu, \sigma_i^2)$. The probability density of the whole data set, read as a function of μ, is $\prod_i (2\pi\sigma_i^2)^{-1/2} \exp[-(x_i-\mu)^2/2\sigma_i^2]$, so its logarithm is a constant minus $\frac{1}{2}\chi^2(\mu)$ with

$$\chi^2(\mu) = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{\sigma_i^2} \tag{3.63}$$

[Cowan (7.25)]. Making the density as large as possible means making χ^2 as small as possible. We complete the square in μ:

$$\begin{aligned}
\chi^2(\mu) &\overset{(1)}{=} \sum_i \frac{x_i^2}{\sigma_i^2} - 2\mu \sum_i \frac{x_i}{\sigma_i^2} + \mu^2 \sum_i \frac{1}{\sigma_i^2} \\
&\overset{(2)}{=} \sum_i \frac{x_i^2}{\sigma_i^2} - 2\mu W\hat{\mu} + \mu^2 W \\
&\overset{(3)}{=} W(\mu - \hat{\mu})^2 + \left[\sum_i \frac{x_i^2}{\sigma_i^2} - W\hat{\mu}^2\right] \overset{(4)}{=} W(\mu - \hat{\mu})^2 + \chi^2_{\min}.
\end{aligned} \tag{3.64}$$

(1) Expand $(x_i - \mu)^2 = x_i^2 - 2\mu x_i + \mu^2$ and sum term by term.
(2) $\sum_i x_i/\sigma_i^2 = W\hat{\mu}$ by the definition of the weighted mean, and $\sum_i 1/\sigma_i^2 = W$.
(3) Add and subtract $W\hat{\mu}^2$: $\mu^2 W - 2\mu W\hat{\mu} + W\hat{\mu}^2 = W(\mu - \hat{\mu})^2$.
(4) The bracket does not depend on μ. It is the value of χ^2 at $\mu = \hat{\mu}$, which we call $\chi^2_{\min}$.

So χ^2 is a parabola in μ with its minimum exactly at the weighted mean [Cowan (7.26)]. Its curvature is $\mathrm{d}^2\chi^2/\mathrm{d}\mu^2 = 2W$, and the density of the data, as a function of μ, is proportional to $\exp[-W(\mu - \hat{\mu})^2/2]$: a Gaussian bump of width $W^{-1/2} = \sigma_{\hat{\mu}}$. So the width of the bump is exactly the error bar we derived from the sampling distribution.

Two more facts can be read off the parabola, and both return in the second half of the chapter. First, the variance is "two over the curvature of χ^2": $\sigma^2_{\hat{\mu}} = 1/W = 2/(2W)$ [Cowan (7.27)], [Sivia (2.31)]. Second, moving μ one standard error away from $\hat{\mu}$ raises χ^2 by exactly 1, because $W\sigma^2_{\hat{\mu}} = 1$. The same formulas also appear as the peak and width of a Bayesian posterior with a flat prior [Sivia pp. 30–31]. For Gaussian data and a flat prior the two readings coincide.

Key idea: Inverse-variance weighting

To combine independent unbiased measurements, weight each by $1/\sigma_i^2$. Inverse variances add: $1/\sigma_{\hat\mu}^2 = \sum_i 1/\sigma_i^2$. For correlated measurements with covariance V, the weights become $\boldsymbol{w} = V^{-1}\mathbf{1}/(\mathbf{1}^{\mathsf T}V^{-1}\mathbf{1})$ and $1/\sigma_{\hat\mu}^2 = \mathbf{1}^{\mathsf T}V^{-1}\mathbf{1}$ (Section 3.7.4).

Example: two measurements of a mass. Two experiments report $x_1 = 10.0 \pm 0.3$ and $x_2 = 10.8 \pm 0.4$ (in GeV). The weights and the combination are

$$\frac{1}{\sigma_1^2} = \frac{1}{0.09} = 11.11, \qquad \frac{1}{\sigma_2^2} = \frac{1}{0.16} = 6.25, \qquad W \overset{(1)}{=} 17.36,$$
$$\hat\mu \overset{(2)}{=} \frac{11.11\times 10.0 + 6.25\times 10.8}{17.36} = \frac{111.1 + 67.5}{17.36} = 10.29, \qquad \sigma_{\hat\mu} \overset{(3)}{=} \frac{1}{\sqrt{17.36}} = 0.24.$$

(1) W is the sum of the inverse variances.
(2) Weighted mean. The first measurement carries weight $11.11/17.36 = 0.64$.
(3) $\sigma_{\hat\mu} = W^{-1/2}$.

The result, 10.29 ± 0.24, is closer to the more precise measurement and is more precise than either. The plain average 10.4 would have the error $\frac{1}{2}\sqrt{0.09 + 0.16} = 0.25$, slightly worse, and would be pulled towards the poorer measurement.

Example: equal errors and one outstanding measurement. Take $N - 1$ measurements with error σ and one with error σ/k. Then $W = (N-1)/\sigma^2 + k^2/\sigma^2$, and the precise measurement carries the weight

$$w_{\text{best}} \overset{(1)}{=} \frac{k^2/\sigma^2}{(N - 1 + k^2)/\sigma^2} = \frac{k^2}{N - 1 + k^2}, \qquad \sigma_{\hat\mu} \overset{(2)}{=} \frac{\sigma}{\sqrt{N - 1 + k^2}}.$$

(1) (3.61), with the common factor $1/\sigma^2$ cancelled.
(2) $\sigma_{\hat\mu} = W^{-1/2}$.

One measurement ten times more precise ($k = 10$) is worth $k^2 = 100$ ordinary ones: with $N - 1 = 9$ ordinary measurements it carries $100/109 = 92\%$ of the weight. Precision is worth its square.

3.7.2 Do the measurements agree? The χ^2 of the residuals

The error $\sigma_{\hat\mu} = W^{-1/2}$ uses only the quoted σ_i, never the spread of the x_i themselves. Two measurements 10.0 ± 0.3 and 10.8 ± 0.4 give the same $\sigma_{\hat\mu}$ as 10.0 ± 0.3 and 15.0 ± 0.4, although the second pair plainly contradicts itself. The value of χ^2 at its minimum is the missing check: it measures how far the data scatter about $\hat\mu$ in units of their own error bars.

What value should we expect? Write $z_i = (x_i - \mu)/\sigma_i$ for the standardised errors: they are independent with mean 0 and variance 1. Then

$$\chi^2_{\min} \overset{(1)}{=} \sum_i \frac{\big[(x_i - \mu) - (\hat\mu - \mu)\big]^2}{\sigma_i^2}$$
$$\overset{(2)}{=} \sum_i z_i^2 - 2(\hat\mu - \mu)\sum_i \frac{x_i - \mu}{\sigma_i^2} + (\hat\mu - \mu)^2 W$$

$$\overset{(3)}{=} \sum_i z_i^2 - 2(\hat\mu-\mu)^2 W + (\hat\mu-\mu)^2 W = \sum_i z_i^2 - W(\hat\mu-\mu)^2,$$

$$\mathbb{E}\,\chi^2_{\min} \overset{(4)}{=} N - W\cdot\frac{1}{W} = N-1. \tag{3.65}$$

(1) Add and subtract the true μ inside $x_i-\hat\mu$.
(2) Expand the square. The first term is z_i^2, and $\sum_i 1/\sigma_i^2 = W$.
(3) $\sum_i (x_i-\mu)/\sigma_i^2 = \sum_i x_i/\sigma_i^2 - \mu W = W(\hat\mu-\mu)$.
(4) $\mathbb{E}\,z_i^2 = 1$ for each of the N terms, and $\mathbb{E}(\hat\mu-\mu)^2 = \operatorname{Var}\hat\mu = 1/W$ because $\hat\mu$ is unbiased.

This is the weighted twin of the $N-1$ in the sample variance (Section 3.4.1): fitting one parameter to the data uses up one of the N independent directions of scatter. For Gaussian measurements the geometry of Section 3.4.1 goes through unchanged in the rescaled coordinates z_i. The noise vector $(z_1,\dots,z_N)$ is isotropic, the fit removes its component along the single direction $(1/\sigma_1,\dots,1/\sigma_N)$, and the remaining $N-1$ perpendicular components are independent standard normals. Hence

$$\chi^2_{\min} \sim \chi^2_{N-1}, \qquad \mathbb{E}\,\chi^2_{\min} = N-1, \qquad \operatorname{Var}\chi^2_{\min} = 2(N-1), \tag{3.66}$$

with the χ^2 distribution of Chapter 2. A value of $\chi^2_{\min}/(N-1)$ near 1 says the scatter matches the error bars. A value far above 1 says it does not.

Example: the two masses again. For two measurements, $\chi^2_{\min}$ depends only on their difference. With $w_1 = W^{-1}/\sigma_1^2$ and $w_2 = 1-w_1$,

$$\chi^2_{\min} \overset{(1)}{=} \frac{(x_1-\hat\mu)^2}{\sigma_1^2} + \frac{(x_2-\hat\mu)^2}{\sigma_2^2} \overset{(2)}{=} (x_1-x_2)^2\Big[\frac{w_2^2}{\sigma_1^2} + \frac{w_1^2}{\sigma_2^2}\Big] \overset{(3)}{=} \frac{(x_1-x_2)^2}{\sigma_1^2+\sigma_2^2},$$
$$\overset{(4)}{=} \frac{(10.0-10.8)^2}{0.09+0.16} = \frac{0.64}{0.25} = 2.56.$$

(1) Definition (3.63) at $\mu=\hat\mu$.
(2) $x_1-\hat\mu = x_1 - w_1x_1 - w_2x_2 = w_2(x_1-x_2)$, and likewise $x_2-\hat\mu = -w_1(x_1-x_2)$.
(3) $w_1 = \sigma_2^2/(\sigma_1^2+\sigma_2^2)$ and $w_2 = \sigma_1^2/(\sigma_1^2+\sigma_2^2)$ (multiply numerator and denominator of w_1 by $\sigma_1^2\sigma_2^2$). The bracket becomes $\big[\sigma_1^4/\sigma_1^2 + \sigma_2^4/\sigma_2^2\big]/(\sigma_1^2+\sigma_2^2)^2 = 1/(\sigma_1^2+\sigma_2^2)$.
(4) Insert the numbers.

The result is natural: x_1-x_2 has mean zero if both measure the same μ and variance $\sigma_1^2+\sigma_2^2$ (3.48), so $\chi^2_{\min}$ is the squared number of standard deviations by which the two disagree, here 1.6. For one degree of freedom, a value of 2.56 or more occurs in 11% of repetitions when both measurements are right, so the pair is acceptable. (This is a first p-value; Chapter 4 develops them properly.) The contradictory pair 10.0 ± 0.3 and 15.0 ± 0.4 gives $\chi^2_{\min} = 25/0.25 = 100$, a ten-standard-deviation disagreement.

What to do when they disagree. A large $\chi^2_{\min}$ means that at least one error bar is underestimated, or that the measurements do not in fact measure the same thing. There is no purely statistical cure. A common convention, used for example in particle-physics averages, is to inflate the combined error by the factor $\sqrt{\chi^2_{\min}/(N-1)}$ when that factor exceeds one. This is equivalent to assuming that every quoted σ_i is too small by the same unknown factor, and it is only as good as that assumption.

Watch out: A tiny error bar from inconsistent data

The formula $\sigma_{\hat{\mu}} = W^{-1/2}$ never looks at the data, so it happily returns a small error for measurements that contradict each other. Always report $\chi^2_{\min}$ and $N-1$ with a weighted mean. When two precise determinations disagree by several standard deviations (as early- and late-Universe measurements of H_0 do in the Hubble tension of Section T2.12.3), their weighted mean is a number that neither experiment supports, and its error bar is meaningless.

3.7.3 Eight laboratories in Python

A simulation checks every claim. The eight laboratories have $\sigma_i = 0.5, 0.7, 1.0, 1.2, 1.5, 2.0, 2.5$ and 3.0, and the true value is $\mu = 0$. We repeat the whole campaign of eight measurements 200 000 times and compute, in each repetition, the plain average, the weighted mean and $\chi^2_{\min}$.

Listing 3.7: Eight laboratories, many times over (`code/ch03/07_weighted_mean.py`)

```
# ---------- (i) eight labs ----------
mu = 0.0
sig = np.array([0.5, 0.7, 1.0, 1.2, 1.5, 2.0, 2.5, 3.0])
N = len(sig)
x = mu + sig * rng.standard_normal((M, N))
w = 1 / sig**2
xw = (x * w).sum(1) / w.sum()
xu = x.mean(1)
chi2 = (((x - xw[:, None]) / sig) ** 2).sum(1)
sd_w_th = 1 / np.sqrt(w.sum())
sd_u_th = np.sqrt((sig**2).sum()) / N
```

The array `x` has one row per repetition and one column per laboratory; multiplying standard normals by the row vector `sig` gives each column its own error bar (numpy broadcasting). The weights `w` are the inverse variances. The weighted mean of each row is the weighted sum divided by `w.sum()`, which is W. In the line for `chi2`, `xw[:, None]` turns the vector of weighted means into a column, so that each row's own $\hat{\mu}$ is subtracted from that row's eight measurements. The last two lines are the predictions, $W^{-1/2}$ for the weighted mean and $\sqrt{\sum_i \sigma_i^2}/N$ for the plain average (the variance of a sum of independent terms divided by N^2).

The weighted mean scatters with standard deviation 0.3391, against the prediction $W^{-1/2} = 0.339$. The plain average scatters with 0.6228, against the prediction $\sqrt{24.68}/8 = 0.621$. The plain average is not merely worse than the weighted mean: it is worse than the best laboratory on its own, whose error is 0.5. Averaging in poor measurements with equal weight destroys information. The simulated $\chi^2_{\min}$ has mean 6.998 and variance 14.05, against $N-1 = 7$ and $2(N-1) = 14$ from (3.66), and Figure 3.27 shows that its whole histogram is χ^2_7.

3.7.4 Correlated measurements

Measurements are often correlated. Two analyses may use the same data in part, the same luminosity, the same energy calibration (Section 3.6.1) or the same theory input. Let the measurements $\boldsymbol{x} = (x_1, \dots, x_N)^{\mathsf{T}}$ be unbiased, $\mathbb{E}\, x_i = \mu$, with a known covariance matrix V, and write $\mathbf{1} = (1, \dots, 1)^{\mathsf{T}}$ for the column of ones. The linear combination $\hat{\mu} = \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}$ is still unbiased exactly when $\mathbf{1}^{\mathsf{T}}\boldsymbol{w} = \sum_i w_i = 1$ [Cowan (7.31)–(7.32)], since the derivation of (3.59) never used independence. Its variance is the error-propagation formula (3.47) with the $1 \times N$ Jacobian

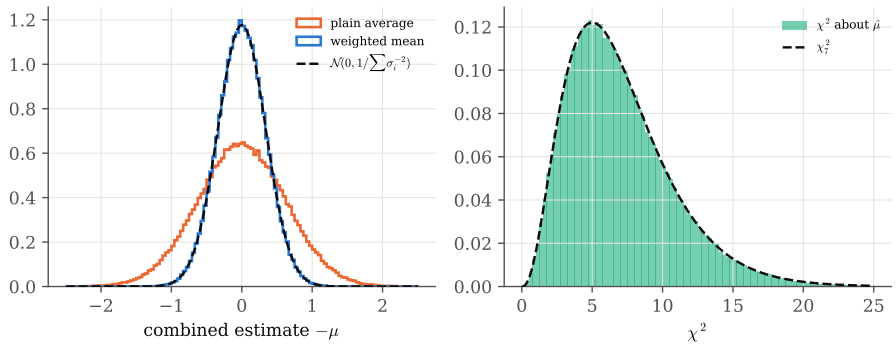


Figure 3.27: Left: over many repetitions of the eight-laboratory campaign, the weighted mean (blue) is a narrow Gaussian of width $W^{-1/2}$ (dashed), while the plain average (orange) is almost twice as wide. Right: the histogram of $\chi^2_{\min}$ about the weighted mean lies on the χ^2 curve with $7 = N - 1$ degrees of freedom, not 8.

$A = \boldsymbol{w}^{\mathsf{T}}$, and this is exact because the map is linear (Section 3.6.2):

$$\operatorname{Var}\hat{\mu} = \sum_{i,j} w_i V_{ij} w_j = \boldsymbol{w}^{\mathsf{T}} V \boldsymbol{w} \tag{3.67}$$

[Cowan (7.33)]. We minimise it with a multiplier as before (Cowan eliminates w_N instead; a multiplier is quicker):

$$\begin{aligned} \mathbf{0} &\overset{(1)}{=} \nabla_{\boldsymbol{w}}\left[\boldsymbol{w}^{\mathsf{T}} V \boldsymbol{w} - \lambda(\mathbf{1}^{\mathsf{T}}\boldsymbol{w} - 1)\right] \overset{(2)}{=} 2V\boldsymbol{w} - \lambda\mathbf{1} \quad\Longrightarrow\quad \boldsymbol{w} = \frac{\lambda}{2} V^{-1}\mathbf{1}, \\ 1 &\overset{(3)}{=} \mathbf{1}^{\mathsf{T}}\boldsymbol{w} = \frac{\lambda}{2}\,\mathbf{1}^{\mathsf{T}} V^{-1}\mathbf{1} \quad\Longrightarrow\quad \boldsymbol{w} = \frac{V^{-1}\mathbf{1}}{\mathbf{1}^{\mathsf{T}} V^{-1}\mathbf{1}}, \\ \operatorname{Var}\hat{\mu} &\overset{(4)}{=} \frac{\mathbf{1}^{\mathsf{T}} V^{-1} V V^{-1}\mathbf{1}}{(\mathbf{1}^{\mathsf{T}} V^{-1}\mathbf{1})^2} = \frac{1}{\mathbf{1}^{\mathsf{T}} V^{-1}\mathbf{1}}. \end{aligned} \tag{3.68}$$

(1) Stationarity in all w_k at once.
(2) $\partial(\boldsymbol{w}^{\mathsf{T}} V \boldsymbol{w})/\partial w_k = \sum_j (V_{kj} + V_{jk}) w_j = 2(V\boldsymbol{w})_k$, because V is symmetric.
(3) Impose the constraint to fix λ.
(4) Insert $\boldsymbol{w}$ into (3.67). V^{-1} is symmetric, so $\boldsymbol{w}^{\mathsf{T}} = \mathbf{1}^{\mathsf{T}} V^{-1}/(\mathbf{1}^{\mathsf{T}} V^{-1}\mathbf{1})$, and $V^{-1} V V^{-1} = V^{-1}$.

In components, $w_i = \sum_j (V^{-1})_{ij} \big/ \sum_{k,l} (V^{-1})_{kl}$ [Cowan (7.30)]. For uncorrelated measurements $(V^{-1})_{ij} = \delta_{ij}/\sigma_i^2$ and we are back at (3.61). The same weights minimise the generalised $\chi^2(\mu) = (\boldsymbol{x} - \mu\mathbf{1})^{\mathsf{T}} V^{-1} (\boldsymbol{x} - \mu\mathbf{1})$ [Cowan (7.28)], whose derivative $-2\mathbf{1}^{\mathsf{T}} V^{-1}(\boldsymbol{x} - \mu\mathbf{1})$ vanishes at $\mu = \mathbf{1}^{\mathsf{T}} V^{-1}\boldsymbol{x} / \mathbf{1}^{\mathsf{T}} V^{-1}\mathbf{1}$. That this least-squares combination is also the minimum-variance unbiased linear one is an instance of the Gauss–Markov theorem [Cowan p. 108].

Two correlated measurements. Two measurements already show every new effect of correlation. With correlation coefficient ρ,

$$V = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}, \qquad V^{-1} = \frac{1}{1-\rho^2} \begin{pmatrix} \dfrac{1}{\sigma_1^2} & \dfrac{-\rho}{\sigma_1\sigma_2} \\ \dfrac{-\rho}{\sigma_1\sigma_2} & \dfrac{1}{\sigma_2^2} \end{pmatrix} \tag{3.69}$$

[Cowan (7.34)–(7.35)]. (Check: the inverse of $\left(\begin{smallmatrix} a & b \\ b & d \end{smallmatrix}\right)$ is $\left(\begin{smallmatrix} d & -b \\ -b & a \end{smallmatrix}\right)/(ad - b^2)$, and here $ad - b^2 = \sigma_1^2\sigma_2^2(1-\rho^2)$.) Then

$$\mathbf{1}^{\mathsf{T}} V^{-1}\mathbf{1} \overset{(1)}{=} \frac{1}{1-\rho^2}\Big[\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} - \frac{2\rho}{\sigma_1\sigma_2}\Big] \overset{(2)}{=} \frac{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}{(1-\rho^2)\sigma_1^2\sigma_2^2}, \tag{3.70}$$

$$w \equiv w_1 \overset{(3)}{=} \frac{\big[1/\sigma_1^2 - \rho/(\sigma_1\sigma_2)\big]/(1-\rho^2)}{\mathbf{1}^{\mathsf{T}} V^{-1}\mathbf{1}} \overset{(4)}{=} \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}, \qquad w_2 = 1 - w, \tag{3.71}$$

$$\operatorname{Var}\hat{\mu} \overset{(5)}{=} \frac{(1-\rho^2)\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} \tag{3.72}$$

[Cowan (7.36)–(7.39)].

(1) The sum of all four entries of V^{-1}.
(2) Common denominator $\sigma_1^2\sigma_2^2$.
(3) $(V^{-1}\mathbf{1})_1$ is the sum of the first row of V^{-1}.
(4) Multiply numerator and denominator by $(1-\rho^2)\sigma_1^2\sigma_2^2$.
(5) $\operatorname{Var}\hat{\mu} = 1/(\mathbf{1}^{\mathsf{T}} V^{-1}\mathbf{1})$ from (3.68) and (3.70).

How much does the second measurement help? Subtract the information of the first alone:

$$\frac{1}{\operatorname{Var}\hat{\mu}} - \frac{1}{\sigma_1^2} \overset{(1)}{=} \frac{1}{1-\rho^2}\Big[\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} - \frac{2\rho}{\sigma_1\sigma_2} - \frac{1-\rho^2}{\sigma_1^2}\Big] \overset{(2)}{=} \frac{1}{1-\rho^2}\Big[\frac{\rho^2}{\sigma_1^2} - \frac{2\rho}{\sigma_1\sigma_2} + \frac{1}{\sigma_2^2}\Big] \overset{(3)}{=} \frac{1}{1-\rho^2}\Big(\frac{\rho}{\sigma_1} - \frac{1}{\sigma_2}\Big)^2. \tag{3.73}$$

(1) (3.70), with $1/\sigma_1^2$ written over the same denominator $1-\rho^2$.
(2) $1/\sigma_1^2 - (1-\rho^2)/\sigma_1^2 = \rho^2/\sigma_1^2$.
(3) The bracket is a perfect square.

This is never negative [Cowan (7.40)]: a second measurement never hurts, even a correlated one. The gain is zero exactly when $\rho = \sigma_1/\sigma_2$, which includes the trivial case of the same measurement entered twice ($\rho = 1$, $\sigma_1 = \sigma_2$). And for $\rho > \sigma_1/\sigma_2$ the weight w of (3.71) exceeds one, so the worse measurement gets a *negative* weight and $\hat{\mu} = wx_1 + (1-w)x_2$ lies outside the interval between x_1 and x_2 [Cowan p. 109].

This is not a paradox. With a strong positive correlation, both measurements are likely to err in the same direction. If the noisier one reads higher than the precise one, it is telling us that the common error pushed both up, the precise one too, so the true value probably lies *below* both. The optimal combination extrapolates. As $\rho \to 1$ the variance (3.72) goes to zero (for $\sigma_1 \neq \sigma_2$): the difference $x_1 - x_2$ then measures the common error exactly, and it can be subtracted.

The same formula cleans the microwave sky. A CMB experiment observes the sky in several frequency channels. In the temperature units used for the CMB, the CMB itself has the same amplitude in every channel, while the Galactic dust, synchrotron and free–free emission change strongly with frequency, and every channel has its own noise. So each channel map is exactly a reading of the form (3.1), $x_\nu = s + (\text{foregrounds} + \text{noise})_\nu$, at every pixel. Take as V the covariance of the channel maps measured from the maps themselves, foregrounds included, and apply (3.68): the weights $\boldsymbol{w} = V^{-1}\mathbf{1}/(\mathbf{1}^{\mathsf{T}} V^{-1}\mathbf{1})$ keep the CMB at unit normalisation, since $\mathbf{1}^{\mathsf{T}}\boldsymbol{w} = 1$, and minimise everything else together, which is possible because the foregrounds are strongly correlated between channels, the large-ρ regime just described, with negative weights doing the subtracting. This is the **internal linear combination** (ILC) [17]. The foregrounds dominate on some angular scales and in some parts of the sky and not in others, so a single set

of weights for the whole map is a compromise. The needlet ILC, NILC [18], splits the maps into pieces localised both in scale and on the sky (needlets) and computes a separate V, hence separate weights, in each piece. It is one of the four methods behind the Planck maps that Chapters 12 and 13 analyse [19]. One caveat comes straight from the derivation: the weights are computed from the same data they are applied to, so the chance correlation between CMB and foregrounds in a finite patch is partly subtracted too, which biases the cleaned map slightly low (the ILC bias); it is small when each piece contains many pixels.

The same thing in Python. The second half of `07_weighted_mean.py` takes $\sigma_1 = 1$, $\sigma_2 = 2$, evaluates (3.71) and (3.72) on a grid of ρ, and, at seven values of ρ, draws 200 000 correlated pairs with `rng.multivariate_normal`, forms $wx_1 + (1-w)x_2$ and measures its variance. Figure 3.28 shows the result. At $\rho = 0.9$ the simulation gives 0.5453 against the formula's 0.5429, well below the better measurement's own variance $\sigma_1^2 = 1$.

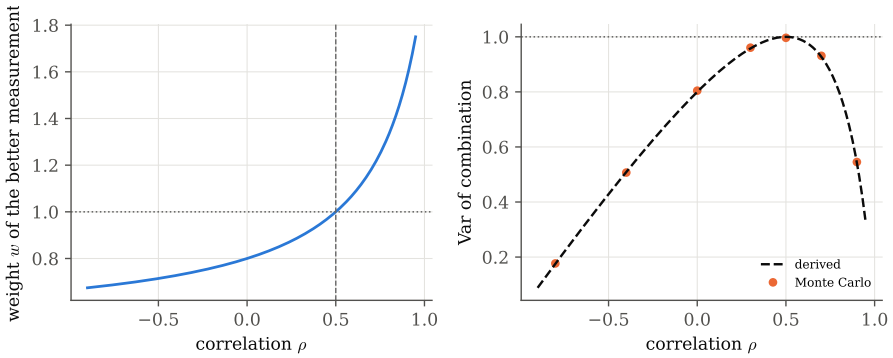


Figure 3.28: Combining two measurements with $\sigma_1 = 1$ and $\sigma_2 = 2$ as their correlation ρ varies. Left: the weight of the better measurement passes 1 at $\rho = \sigma_1/\sigma_2 = 0.5$ (dashed vertical line), and beyond it the poorer measurement is subtracted. Right: the variance of the optimal combination (dashed curve, formula; dots, simulation) touches the better measurement's variance 1 only at $\rho = 0.5$ and falls towards zero as $\rho \to 1$.

Example: two rulers and one thermometer. Cowan's example makes the negative weight concrete [Cowan §7.6.1]. We measure a length λ with two rulers of different materials, calibrated at temperature T_0, with thermal expansion coefficients c_1 and c_2. Each reading L_i is corrected with the measured temperature T,

$$y_i = L_i + c_i(T - T_0), \tag{3.74}$$

[Cowan (7.41)]. The readings have independent errors σ_L, and the one thermometer reading T has error σ_T and enters both corrections. Linear propagation, exact here, gives

$$\sigma_i^2 \overset{(1)}{=} \sigma_L^2 + c_i^2\sigma_T^2, \qquad V_{12} \overset{(2)}{=} c_1c_2\sigma_T^2, \qquad \rho \overset{(3)}{=} \frac{c_1c_2\sigma_T^2}{\sqrt{(\sigma_L^2 + c_1^2\sigma_T^2)(\sigma_L^2 + c_2^2\sigma_T^2)}}$$

[Cowan (7.42)–(7.44)].

(1) Variance of a sum of independent terms, with $\mathrm{Var}[c_i(T - T_0)] = c_i^2\sigma_T^2$.
(2) The only error the two corrected lengths share is the temperature error, as in Section 3.6.1: $\mathrm{Cov}(c_1T, c_2T) = c_1c_2\sigma_T^2$.
(3) $\rho = V_{12}/(\sigma_1\sigma_2)$.

Take $c_1 = 0.02$ and $c_2 = 0.05\,\mathrm{mm\,K^{-1}}$, $\sigma_L = 0.01\,\mathrm{mm}$ and $\sigma_T = 2\,\mathrm{K}$, and suppose the room is really at 22°C, two degrees above $T_0 = 20°\mathrm{C}$, for a true length of 1000.00 mm; the thermometer reads 25°C. The corrected lengths come out as $y_1 = 1000.06$ and $y_2 = 1000.15\,\mathrm{mm}$. The numbers are

$$\sigma_1^2 = 10^{-4} + 4\times10^{-4}\times 4 = 0.0017, \qquad \sigma_2^2 = 10^{-4} + 25\times10^{-4}\times 4 = 0.0101, \qquad V_{12} = 0.004,$$

$$\rho = \frac{0.004}{\sqrt{0.0017\times 0.0101}} = 0.965 \;>\; \frac{\sigma_1}{\sigma_2} = 0.410,$$

$$w \overset{(1)}{=} \frac{\sigma_2^2 - V_{12}}{\sigma_1^2 + \sigma_2^2 - 2V_{12}} = \frac{0.0061}{0.0038} = 1.605,$$

$$\hat\lambda \overset{(2)}{=} 1000 + 1.605\times 0.06 - 0.605\times 0.15 = 1000.006\,\mathrm{mm}, \qquad \sigma_{\hat\lambda} \overset{(3)}{=} \sqrt{\frac{\sigma_1^2\sigma_2^2 - V_{12}^2}{0.0038}} = 0.018\,\mathrm{mm}.$$

(1) (3.71) with $\rho\sigma_1\sigma_2 = V_{12}$.
(2) $\hat\lambda = wy_1 + (1-w)y_2$, written relative to 1000 mm.
(3) (3.72), again with $\rho\sigma_1\sigma_2 = V_{12}$, so that $(1-\rho^2)\sigma_1^2\sigma_2^2 = \sigma_1^2\sigma_2^2 - V_{12}^2$.

The combination 1000.006 ± 0.018 mm lies *outside* the interval [1000.06, 1000.15] and is right, whereas the naive weighted mean that ignores the correlation gives 1000.073 ± 0.038 mm, too high by twice its own error.

Why is a value outside both readings right? Figure 3.29 shows it. Each corrected length is a straight line in the unknown true temperature. Both rulers measure the same length, so they must agree at the true temperature. The estimate is therefore pulled towards the point where the two lines cross [Cowan Fig. 7.7]. The same fit returns an estimate of the temperature itself, $\hat T = 22.2°\mathrm{C}$, from [Cowan (7.47)]: the disagreement between the rulers has been used to correct the thermometer. In the limit $\sigma_L \to 0$ ($\rho \to 1$) the two lines give $\hat\lambda = (c_2y_1 - c_1y_2)/(c_2 - c_1)$ with zero variance.[1]

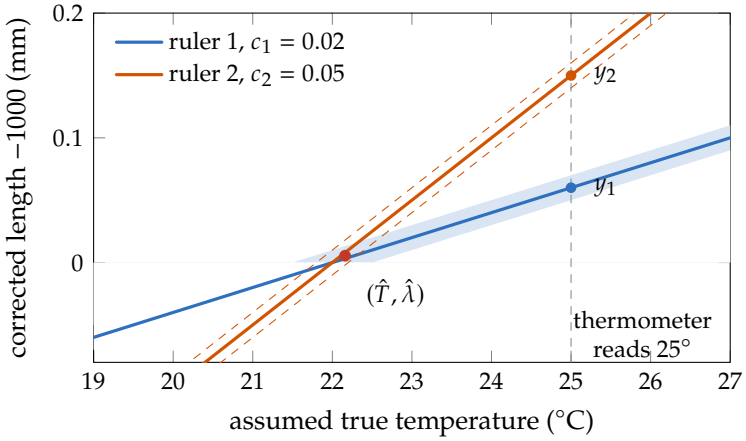


Figure 3.29: Two rulers read at a thermometer reading of 25°C give the corrected lengths y_1 and y_2 (dots on the dashed vertical line). If the temperature were different, each corrected length would slide along its own line, whose slope is the ruler's expansion coefficient; the shaded strip and the thin dashed lines show the $\pm\sigma_L$ reading error. Both rulers measure one length, so the truth sits where the lines cross. The correlated weighted mean (red dot) lands there, below both y_1 and y_2, and at a temperature near 22°C.

[1] This follows from (3.74) with exact readings: then $y_i = \lambda + c_i(T - T_{\mathrm{true}})$, and eliminating $T - T_{\mathrm{true}}$ between $i = 1, 2$ gives $c_2y_1 - c_1y_2 = (c_2 - c_1)\lambda$. Cowan's (7.46) has c_1 and c_2 interchanged in the numerators; the check $c_1 = 0$ (a ruler that does not expand, so $\lambda = y_1$) confirms the version given here.

How research uses it. Every world average of a particle mass or lifetime is an inverse-variance weighted mean, quoted with its $\chi^2_{\min}$ and, when needed, the inflation factor of Section 3.7.2. When several experiments share systematic errors, the correlated formula (3.68) is used under the name "best linear unbiased estimate" (the original discussion is by Lyons and collaborators, [Cowan p. 109]). The hard part is not the algebra but the off-diagonal elements of V, which must be estimated.

Here our Monte Carlo only *verifies* formulas we derived. In research the covariance of the inputs is itself often taken from simulations, and then its noise (Section 3.5.4) passes into the weights.

In cosmology the same algebra appears in a more general form. Combining independent probes with Gaussian likelihoods multiplies the likelihoods, and multiplying Gaussians adds their inverse covariances. For one parameter that is exactly $1/\sigma^2 = \sum_i 1/\sigma_i^2$; for many parameters it is the statement that Fisher matrices add (Chapter 10).

Where this leaves us. Every error bar so far came from one idea: treat a summary of the data as a random variable over imagined repetitions, and ask for its centre and width. The weighted mean showed a second way of seeing the same numbers. The density of the data, read as a function of the unknown μ, peaked at the estimator and had a width equal to its standard error. That reading of the density is the likelihood, the subject of the second half of the chapter. Maximising it gives estimators, and its curvature gives their error bars.

Recap: Part 3a in one page

- An estimator $\hat{\theta} = g(x_1, \ldots, x_N)$ is a random variable. Its sampling distribution over imagined repetitions gives the error bar (the standard error). Simulate it when in doubt.
- MSE = bias^2 + variance. Consistency: $\hat{\theta} \to \theta$ in probability, guaranteed when bias and variance both go to zero.
- Var $\bar{x} = \sigma^2/N$: errors shrink like $1/\sqrt{N}$ (law of large numbers via Chebyshev). The central limit theorem makes the sampling distribution of a mean Gaussian.
- S^2 divides by $N-1$ to be unbiased. Its own error is Var $S^2 = 2\sigma^4/(N-1)$ for Gaussian data (cosmic variance is the same fact).
- Sample covariance and correlation matrices are noisy: sample correlations of independent variables scatter by $\sim 1/\sqrt{N}$, and inverting a simulated covariance needs the Hartlap factor.
- Error propagation: $U = AVA^{\mathsf{T}}$ with the Jacobian A, exact for linear maps. It fails near poles and for noisy denominators; use Monte Carlo propagation there.
- Combining measurements: weight by $1/\sigma_i^2$, so $1/\sigma_{\hat{\mu}}^2 = \sum_i 1/\sigma_i^2$. Check $\chi^2_{\min} \sim \chi^2_{N-1}$. With correlations, $\boldsymbol{w} = V^{-1}\mathbf{1}/(\mathbf{1}^{\mathsf{T}}V^{-1}\mathbf{1})$, and weights can be negative.

3.8 The likelihood: data are the model plus noise

We measure a voltage, count photons in a detector, record the arrival times of decays, or estimate a power spectrum from a map of the sky. In every case we also have a theory with a knob on it: a resistance, a cross-section, a lifetime, a cosmological parameter. Set the knob to a value θ

and the theory predicts what the apparatus should have shown. Call that prediction $d(\theta)$. The apparatus never shows exactly $d(\theta)$. It shows the prediction plus whatever the noise happened to be on that day. Until now the question has run one way: *given* the truth, how does a summary of the data scatter when the experiment is repeated? Now the data are in, and we ask the opposite question: which values of the knob could have produced the data we actually have?

The tool for this backward question is the **likelihood**. It is not a new probability distribution. It is the old one, the distribution of the data, read in the other direction. To understand it we need answers to four questions.

1. How does a model of the noise turn into a number attached to each value of θ?
2. What happens with many measurements?
3. Why is that number *not* a probability distribution for θ, even though it is built from one?
4. Where does it sit in Bayes' theorem, and what information in the data does it keep?

The pipeline of Chapter 0 ends with an arrow from how a statistic S scatters, $P(S \mid \theta)$, to what the observed value of S says about the physics. The likelihood is that arrow. It is the sampling distribution of the data (or of a summary S of them), evaluated at the data we observed and read as a function of the physics θ.

3.8.1 One measurement: from the noise to the likelihood

Start with the simplest experiment there is. The model predicts a number $d(\theta)$, and the measurement adds noise:

$$D = d(\theta) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2), \tag{3.75}$$

with σ known from calibrating the instrument. The noise ϵ [the random amount by which this particular reading differs from the ideal, noiseless one] has density $\phi_\sigma(\epsilon) = (2\pi\sigma^2)^{-1/2}\exp(-\epsilon^2/2\sigma^2)$. What is the density of the reading D itself? Fix θ and ask for the probability that D lands in a small window $[D_0, D_0 + \mathrm{d}D]$:

$$\begin{aligned} \mathrm{P}\big(D_0 \le D \le D_0 + \mathrm{d}D \mid \theta\big) &\overset{(1)}{=} \mathrm{P}\big(D_0 - d(\theta) \le \epsilon \le D_0 - d(\theta) + \mathrm{d}D\big) \\ &\overset{(2)}{=} \phi_\sigma\big(D_0 - d(\theta)\big)\,\mathrm{d}D \\ &\overset{(3)}{=} \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{\big(D_0 - d(\theta)\big)^2}{2\sigma^2}\right]\mathrm{d}D. \end{aligned} \tag{3.76}$$

(1) At fixed θ the prediction $d(\theta)$ is just a number, so $D = d(\theta) + \epsilon$ lies in the window exactly when ϵ lies in the same window shifted down by $d(\theta)$.
(2) For a small window, probability is density times width. A shift does not stretch the window, so the width is still $\mathrm{d}D$.
(3) Insert the Gaussian density of the noise.

So $D \sim \mathcal{N}(d(\theta), \sigma^2)$: the reading is Gaussian, centred on the prediction, with the width of the noise. Nothing new so far. This is the *forward* or *prediction* direction: fix the physics, ask how the data would scatter.

Now the data are in. We hold $D = D_0$ fixed at the value on the lab book and let θ vary instead. The same expression becomes a function of θ,

$$\mathcal{L}(\theta) \equiv p(D_0 \mid \theta) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{\big(D_0 - d(\theta)\big)^2}{2\sigma^2}\right] = \phi_\sigma\big(\underbrace{D_0 - d(\theta)}_{\text{required noise}}\big). \tag{3.77}$$

The last form says what the likelihood measures. Take a trial value θ. The model predicts $d(\theta)$. To turn that prediction into the reading we actually got, the noise must have been exactly $\epsilon = D_0 - d(\theta)$. The likelihood is the probability density of *that* noise value. So for each trial θ it answers one question: how probable is the noise that this θ needs in order to explain our reading?

Key idea: what the likelihood measures

Pick a trial value θ. The model then says what the reading should have been, $d(\theta)$, and the gap $D_0 - d(\theta)$ is the noise that nature would have had to add to turn that prediction into the reading we actually got. The likelihood of θ is how probable this particular piece of noise is. In symbols it is nothing more than the distribution of the data: if $D = d(\theta) + \epsilon$ with $\epsilon \sim \mathcal{N}(0, 1)$, then $D \sim \mathcal{N}(d(\theta), 1)$, and this density, with D held at the observed value and θ left free, is the likelihood function. For noise of width σ, as in (3.77), the 1 becomes σ^2.

A value of θ whose prediction sits close to the data needs only a small, ordinary fluctuation and gets a high likelihood. A value whose prediction sits five noise widths away needs a five-sigma fluctuation and gets a likelihood smaller by a factor $e^{-25/2} \approx 4 \times 10^{-6}$. Figure 3.30 draws exactly this for several data points and two trial models.

In the language of the backward question, each trial value of θ is a candidate explanation, the reading D_0 is what we saw, and the likelihood ranks the candidates by how easily each one would have produced that reading. Parameter estimation in Chapter 5 is built on exactly this.

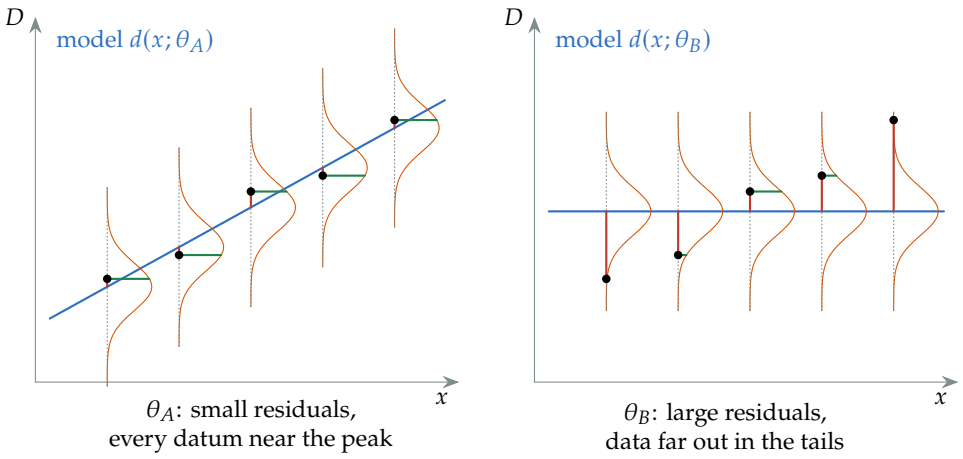


Figure 3.30: The likelihood of two trial parameter values. Black dots are the same five measurements in both panels. The blue line is the model prediction for the trial value, and the orange curve on its side at each x is the noise density $\mathcal{N}(d, \sigma^2)$ centred on the prediction. The red segment is the fluctuation the noise would have had to produce, and the length of the green tick is the height of the noise density at the datum. The likelihood is the product of the five green lengths. On the left they are all long. On the right two of them are almost zero, so θ_B needs an implausible noise realisation.

Working definition 3.10 (likelihood)

For data $\boldsymbol{d}$ and a model that gives the data a probability (or density) $p(\boldsymbol{d} \mid \boldsymbol{\theta})$ for each parameter value, the **likelihood** is that same function with the data held fixed at their observed values and the parameters varied,

$$\mathcal{L}(\boldsymbol{\theta}) \equiv p(\boldsymbol{d}_{\text{obs}} \mid \boldsymbol{\theta}).$$

Operationally: write down the noise model that turns a prediction $d(\boldsymbol{\theta})$ into data [for

example "add independent Gaussian noise of known width"], compute the probability of the observed data under it, and treat the answer as a function of $\boldsymbol{\theta}$ alone. The **log-likelihood** is $\ell(\boldsymbol{\theta}) = \ln \mathcal{L}(\boldsymbol{\theta})$.

Example: one reading of a resistance. We pass a known current $I = 2\,\mathrm{A}$ through a resistor and read the voltage with a voltmeter whose noise is $\sigma = 0.2\,\mathrm{V}$. The model is Ohm's law, $d(R) = IR$, and the reading is $D_0 = 10.4\,\mathrm{V}$. Then

$$\mathcal{L}(R) \overset{(1)}{=} \frac{1}{\sqrt{2\pi}\,(0.2\,\mathrm{V})} \exp\left[-\frac{(10.4\,\mathrm{V} - 2\,\mathrm{A}\cdot R)^2}{2(0.2\,\mathrm{V})^2}\right] \overset{(2)}{=} \frac{1}{\sqrt{2\pi}\,(0.2\,\mathrm{V})} \exp\left[-\frac{(R - 5.2\,\Omega)^2}{2(0.1\,\Omega)^2}\right].$$

(1) (3.77) with $d(R) = IR$.
(2) Factor I^2 out of the square: $(D_0 - IR)^2 = I^2(R - D_0/I)^2$, and $\sigma^2/I^2 = (0.1\,\Omega)^2$.

As a function of R the likelihood is a Gaussian bump centred at $5.2\,\Omega$ with width $0.1\,\Omega$. The peak is at the resistance whose prediction needs zero noise: $IR = D_0$ gives $R = 10.4\,\mathrm{V}/2\,\mathrm{A} = 5.2\,\Omega$. The width is the noise width divided by the slope $\mathrm{d}d/\mathrm{d}R = I$, because a change δR moves the prediction by $I\,\delta R$. The prefactor carries units of V^{-1}, because it is a density in the *voltage*, not in R. That is the first sign that the likelihood is not a probability density for R (Section 3.8.3).

3.8.2 Many measurements: the product and the logarithm

Real experiments have many data points $D_1, \ldots, D_N$, taken at settings $x_1, \ldots, x_N$, with predictions $d_i(\theta) = d(x_i; \theta)$. If the noise in different readings is independent, the joint density factorises, because independence means precisely that the joint density is the product of the single ones [Sivia p. 65, (3.62)]:

$$\mathcal{L}(\theta) \overset{(1)}{=} p(D_1, \ldots, D_N \mid \theta) \overset{(2)}{=} \prod_{i=1}^{N} p(D_i \mid \theta) \overset{(3)}{=} \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{\big(D_i - d_i(\theta)\big)^2}{2\sigma_i^2}\right], \tag{3.78}$$

$$\ell(\theta) \overset{(4)}{=} -\frac{1}{2}\sum_{i=1}^{N} \frac{\big(D_i - d_i(\theta)\big)^2}{\sigma_i^2} \; - \; \sum_{i=1}^{N} \ln\sqrt{2\pi\sigma_i^2}. \tag{3.79}$$

(1) Definition of the likelihood for the whole data vector.
(2) Independent noise: the probability of the joint noise realisation is the product of the probabilities of the individual ones.
(3) Insert (3.76) for each reading, with its own noise width σ_i.
(4) The logarithm of a product is the sum of the logarithms, and $\ln e^u = u$.

The first sum in (3.79) is $\chi^2(\theta)$: each residual $D_i - d_i(\theta)$ is divided by its error bar σ_i, squared, and the squares are added. The second sum does not depend on θ at all. So, for Gaussian noise, $\ell(\theta) = -\tfrac{1}{2}\chi^2(\theta) + \text{const}$. This is why least squares works (Section 3.10).

For identically distributed data, $x_1, \ldots, x_n$ drawn independently from one density $f(x;\theta)$, the same steps give Wasserman's definition [Wasserman p. 122, Def. 9.7],

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(x_i;\theta), \qquad \ell_n(\theta) = \sum_{i=1}^{n} \ln f(x_i;\theta). \tag{3.80}$$

Three practical remarks follow from the product structure.

(i) *Work with logarithms.* A product of a thousand densities of size 0.1 is 10^{-1000} and underflows every computer. Its logarithm, -2303, is harmless. Sums are also what we differentiate most easily.

(ii) *Constants do not matter.* Multiplying $\mathcal{L}$ by any positive number c that does not depend on θ changes nothing we will do with it: the location of the maximum, the shape, and ratios $\mathcal{L}(\theta_1)/\mathcal{L}(\theta_2)$ are all unchanged [Wasserman p. 123, Remark 9.9]. We drop such constants freely and write $\mathcal{L} \propto \dots$.

(iii) *Only ratios carry meaning.* Since the overall constant is arbitrary, the statement "$\mathcal{L}(\theta_1) = 0.03$" means nothing. The statement "$\mathcal{L}(\theta_1)/\mathcal{L}(\theta_2) = 20$" means that the data are twenty times more probable under θ_1 than under θ_2. The cosmology literature often reports $-2\ln[\mathcal{L}(\theta)/\mathcal{L}_{\max}]$ [Padilla (26)], which is a $\Delta\chi^2$ for Gaussian noise.

Example: a coin, or any yes/no detector. A detector fires ($x_i = 1$) or not ($x_i = 0$) with unknown probability p. Each trial is Bernoulli, $f(x;p) = p^x(1-p)^{1-x}$. In $n = 20$ trials it fires $S = \sum_i x_i = 12$ times [Wasserman p. 123, Ex. 9.10]. Then

$$\mathcal{L}(p) \overset{(1)}{=} \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \overset{(2)}{=} p^{\sum_i x_i}(1-p)^{n-\sum_i x_i} = p^{12}(1-p)^{8}.$$

(1) (3.80) with the Bernoulli probability.
(2) Collect the powers: $\prod_i p^{x_i} = p^{\sum_i x_i}$ and likewise for $1-p$.

The left panel of Figure 3.31 shows it. It is a smooth bump with its peak at $p = S/n = 0.6$. The individual zeros and ones have disappeared: only their sum S enters.

Example: a uniform distribution with an unknown end point. Let $x_i \sim \text{Uniform}(0, \theta)$, with density $1/\theta$ on $[0, \theta]$ and zero outside [Wasserman p. 124, Ex. 9.12]. Each factor of the product is $1/\theta$ if $x_i \le \theta$ and zero if $x_i > \theta$. The product is nonzero only if *every* $x_i \le \theta$, that is, if $\theta \ge x_{(n)} \equiv \max_i x_i$:

$$\mathcal{L}(\theta) = \begin{cases} \theta^{-n}, & \theta \ge x_{(n)}, \\ 0, & \theta < x_{(n)}. \end{cases}$$

A value θ smaller than the largest observation is not merely implausible, it is impossible: no noise realisation can put a point above the end of the box. For ten simulated points with largest value $x_{(10)} = 0.9072$, the middle panel of Figure 3.31 shows a cliff at $x_{(10)}$ followed by a decay as θ^{-10}. It has no smooth peak. Because of that cliff, this model breaks rules that smooth likelihoods obey: the maximum is not found by setting a derivative to zero (Section 3.9.1), and the Cramér–Rao bound does not apply (Section 3.13.3).

3.8.3 The likelihood is not a probability density for θ

The likelihood is not the probability density of θ. The temptation is real: it is built from a density, and in the resistor example it even looked like a Gaussian in R. Two facts show that it is not a density for θ. Its area over θ is arbitrary, and it does not change the way a density must when we rename the parameter.

The area under the likelihood is arbitrary. Integrate the resistor likelihood over R:

$$\int_{-\infty}^{\infty} \mathcal{L}(R)\,\mathrm{d}R \overset{(1)}{=} \frac{1}{\sqrt{2\pi}\,\sigma}\int_{-\infty}^{\infty} \exp\!\left[-\frac{(R-D_0/I)^2}{2(\sigma/I)^2}\right]\mathrm{d}R \overset{(2)}{=} \frac{1}{\sqrt{2\pi}\,\sigma}\sqrt{2\pi}\,\frac{\sigma}{I} = \frac{1}{I} = 0.5\,\mathrm{A}^{-1}.$$

(1) The form found in the resistor example, with $\sigma = 0.2\,\mathrm{V}$.

(2) The Gaussian integral $\int e^{-(u-u_0)^2/2s^2}\mathrm{d}u = \sqrt{2\pi}\,s$ with $s = \sigma/I$.

If $\mathcal{L}(R)$ were a probability density for R, this area would be 1. It is 0.5, and it is not even a pure number: it has units of inverse amperes. Change the current to 1 A and the area becomes 1. Measure the current in milliamperes and it changes again. The reason is that the likelihood is a density in the *data*: it integrates to one over D, by construction, and its units are those of $1/D$. Nothing fixes its integral over θ.

The same happens in every example. For the detector,

$$\int_0^1 \mathcal{L}(p)\,\mathrm{d}p \overset{(1)}{=} \int_0^1 p^S(1-p)^{n-S}\,\mathrm{d}p \overset{(2)}{=} B(S+1, n-S+1) \overset{(3)}{=} \frac{S!\,(n-S)!}{(n+1)!} = \frac{12!\,8!}{21!} = 3.780\times10^{-7},$$

(1) The Bernoulli likelihood of the example above.

(2) The Beta integral $\int_0^1 p^{a-1}(1-p)^{b-1}\mathrm{d}p = B(a,b)$, the normalising constant of the Beta distribution (Chapter 2), with $a = S+1$, $b = n-S+1$.

(3) $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ and $\Gamma(m+1) = m!$.

and the numerical integral in `08_likelihood_shapes.py` gives the same number, 3.780×10^{-7}. The peak value is $\mathcal{L}(0.6) = 1.427\times10^{-6}$, so the area is only 0.265 times the peak height (in units of p). Nothing about it is one.

Renaming the parameter changes the area but not the values. The second fact goes deeper. When we change the name of the parameter, a density picks up a Jacobian factor and the likelihood does not. To see it, take $k = 5$ decay times with total $T = \sum_i t_i = 10.51$ (in units of the true lifetime), and describe the decay either by the mean life τ or by the rate $\lambda = 1/\tau$. With $f(t;\tau) = \tau^{-1}e^{-t/\tau}$,

$$\mathcal{L}(\tau) \overset{(1)}{=} \prod_{i=1}^{k}\frac{1}{\tau}e^{-t_i/\tau} = \tau^{-k}e^{-T/\tau}, \qquad \int_0^\infty \mathcal{L}(\tau)\,\mathrm{d}\tau \overset{(2)}{=} \frac{\Gamma(k-1)}{T^{k-1}} = 4.920\times10^{-4},$$

$$\mathcal{L}(\lambda) \overset{(3)}{=} \prod_{i=1}^{k}\lambda e^{-\lambda t_i} = \lambda^k e^{-\lambda T}, \qquad \int_0^\infty \mathcal{L}(\lambda)\,\mathrm{d}\lambda \overset{(4)}{=} \frac{\Gamma(k+1)}{T^{k+1}} = 8.910\times10^{-5}.$$

(1) (3.80) with the exponential density.

(2) Substitute $u = T/\tau$, $\mathrm{d}\tau = -T\,\mathrm{d}u/u^2$: $\int_0^\infty \tau^{-k}e^{-T/\tau}\mathrm{d}\tau = T^{1-k}\int_0^\infty u^{k-2}e^{-u}\mathrm{d}u = T^{1-k}\Gamma(k-1)$.

(3) The same density written with λ: $f(t;\lambda) = \lambda e^{-\lambda t}$.

(4) The Gamma integral $\int_0^\infty \lambda^k e^{-\lambda T}\mathrm{d}\lambda = \Gamma(k+1)/T^{k+1}$.

The two areas differ by a factor $T^2/[k(k-1)] = 5.522$. Yet at corresponding points the two functions are *equal*: putting $\lambda = 1/\tau$ into $\lambda^k e^{-\lambda T}$ gives $\tau^{-k}e^{-T/\tau}$, so $\mathcal{L}(\lambda = 1/\tau) = \mathcal{L}(\tau)$. The likelihood behaves like a scalar field. It assigns a number to each physical hypothesis ("the lifetime is 1.3 ns"), whatever coordinate we use to name that hypothesis.

A genuine probability density behaves differently. If $p_\tau(\tau)$ were a density for τ, the probability of a small interval $\mathrm{d}\tau$ would be $p_\tau\,\mathrm{d}\tau$. The same interval written in λ has width $\mathrm{d}\lambda = \mathrm{d}\tau/\tau^2$, and the probability must be the same, so the density for λ would be

$p_\lambda(\lambda) = p_\tau(1/\lambda)\,|\mathrm{d}\tau/\mathrm{d}\lambda| = p_\tau(1/\lambda)/\lambda^2$. The factor $1/\lambda^2$ is the Jacobian. The likelihood never picks one up, so the areas under it in different coordinates disagree, and neither area means anything. The right panel of Figure 3.31 shows the two curves, each scaled to peak height one: the same heights at corresponding points, but different shapes along the axis.

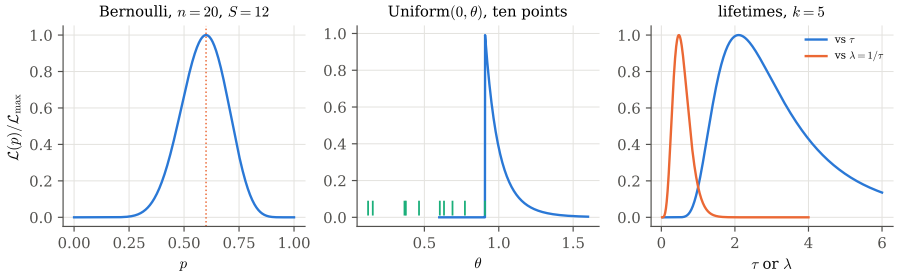


Figure 3.31: Three likelihoods, each scaled to a maximum of one. Left: a yes/no detector that fired 12 times in 20 trials, a smooth bump peaked at $p = 0.6$. Middle: ten points from a uniform box of unknown length (ticks at the bottom). The likelihood is zero until θ reaches the largest point, then decays as θ^{-10}. Right: five decay times, with the likelihood drawn once against the mean life τ and once against the rate $\lambda = 1/\tau$. The two curves take the same value at $\lambda = 1/\tau$, but they enclose different areas.

Key idea: The likelihood is a function of θ, not a distribution over θ

$\mathcal{L}(\theta) = p(\boldsymbol{d}_{\text{obs}} \mid \theta)$ is normalised over the *data*, $\int p(\boldsymbol{d} \mid \theta)\,\mathrm{d}\boldsymbol{d} = 1$ for every θ. Over θ it has no normalisation, it changes no Jacobian under a change of parameter, and only ratios $\mathcal{L}(\theta_1)/\mathcal{L}(\theta_2)$ are meaningful [Wasserman p. 122]. To get a probability distribution over θ one needs a prior, and Bayes' theorem (Chapter 5).

3.8.4 Prediction and inference: where the likelihood sits

The two readings of $p(\boldsymbol{d} \mid \theta)$ are the two directions of probability that run through the whole book (Chapter 1). Read forward, it starts from a fixed θ and says which data sets that θ can produce, and how often. Read backward, the data set is already in hand, and we ask which values of θ could have produced it and how much each one deserves to be believed now that we have it.

Prediction is the first reading, *inference* the second. Figure 3.32 draws both as trees. In the prediction tree one parameter value fans out into many possible data sets, and their probabilities add to one. In the inference tree many parameter values point at the one data set we have, and the numbers attached to the branches are the likelihoods. They need not add to anything.

So, facing data, we ask two questions in words before writing anything. First: which values of θ could have produced these data? Second: how probable would these data have been under each of them? The answers to the second question, read across all θ, are the likelihood. The first question fixes what the posterior will be a distribution over. Turning the likelihood into that distribution needs one more ingredient, a prior (Chapter 5).

Bayes' theorem is the bridge from one tree to the other. Written for parameters it reads [HobsonBMC p. 11, (1.25)]

$$\underbrace{p(\theta)}_{\text{prior}}\ \underbrace{p(\boldsymbol{d} \mid \theta)}_{\text{likelihood}} = p(\theta, \boldsymbol{d}) = \underbrace{p(\boldsymbol{d})}_{\text{evidence } Z}\ \underbrace{p(\theta \mid \boldsymbol{d})}_{\text{posterior}}. \tag{3.81}$$

Both sides are the joint probability of parameter and data, $p(\theta, \boldsymbol{d})$, split into two factors in the two possible orders. The likelihood is the part the experiment supplies. It is also the least

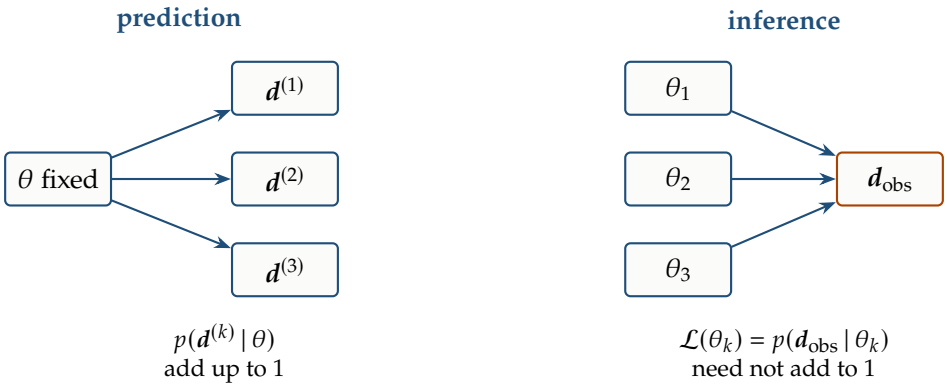


Figure 3.32: The same conditional probability read two ways. Left: one hypothesis produces many possible data sets, and the probabilities of all of them sum to one. Right: many hypotheses could have produced the single data set we observed (orange). The likelihood attaches to each hypothesis the probability that it would have produced exactly these data. These numbers compare hypotheses with each other but are not a distribution over them.

controversial part. We can calibrate the instrument with known inputs and count how often it produces each output, which pins $p(\boldsymbol{d} \mid \theta)$ down to any desired precision [HobsonBMC p. 11]. The prior, the evidence and the posterior belong to Chapter 5. Two facts about them are needed now. First, with a flat prior $p(\theta) = \text{const}$ the posterior is proportional to the likelihood, so the peak of the posterior is the peak of the likelihood [Sivia p. 64, (3.61)]. Second, "flat" depends on the parametrisation. A prior flat in τ is not flat in $\lambda = 1/\tau$, because of the Jacobian $1/\lambda^2$ found above. From here on we stay on the frequentist side and use the likelihood by itself, without any prior.

What the likelihood keeps: sufficiency. In each of the three examples the likelihood depended on the data through one number only: the count S of detector firings, the total decay time T, and the largest value $x_{(n)}$ of the uniform box. Two detector data sets with the same S give the same likelihood function, so nothing computed from the likelihood can tell them apart. A statistic with this property is called **sufficient** [knowing its value lets us write down the whole likelihood function, up to a constant factor] [Wasserman p. 137, Def. 9.32]. The factorisation theorem [Wasserman p. 140, Thm. 9.40] gives a quick test: $T(\boldsymbol{d})$ is sufficient exactly when $p(\boldsymbol{d} \mid \theta) = g\big(T(\boldsymbol{d}), \theta\big)\, h(\boldsymbol{d})$. For Gaussian data with unknown μ and σ the pair $(\bar{x}, S^2)$ is sufficient [Wasserman p. 138, Ex. 9.34]: the likelihood $\propto \sigma^{-n} \exp\{-[(n-1)S^2 + n(\bar{x}-\mu)^2]/2\sigma^2\}$ uses nothing else. Cosmologists use the same compression every day: when a map is Gaussian and statistically isotropic, its power spectrum carries all the information about the parameters. Sufficiency also improves estimators. The Rao–Blackwell theorem says that averaging any estimator over all data sets that share the observed value of a sufficient statistic gives a new estimator whose mean squared error is never larger [Wasserman p. 140, Thm. 9.42]: an estimator that uses anything beyond the sufficient statistic is only adding noise.

The likelihood principle. The **likelihood principle** goes one step further [CasellaBerger §6.3.1]. It says that if two data sets $\boldsymbol{d}$ and $\boldsymbol{d}'$ give proportional likelihoods, $\mathcal{L}(\theta \mid \boldsymbol{d}) = c\, \mathcal{L}(\theta \mid \boldsymbol{d}')$ for all θ, then they carry the same evidence about θ and should lead to the same conclusion. Two coin experiments show what this means. Experimenter A decides to toss a coin $n = 20$ times and sees $S = 12$ heads. The count is binomial, and $\mathcal{L} \propto p^{12}(1-p)^8$. Experimenter B decides to toss until the eighth tail, and the eighth tail comes on toss 20, so B also saw 12 heads and 8 tails. That count is negative binomial, and again $\mathcal{L} \propto p^{12}(1-p)^8$, only with a different constant in front. The likelihoods are proportional, so by the principle A and B have learned the same

thing about p. Estimates and intervals built from the likelihood alone respect the principle. Procedures that average over data sets that could have been observed but were not, such as p-values (Chapter 4), depend on the stopping rule, so they violate it. Bayesian inference satisfies it automatically, because the posterior uses the data only through $\mathcal{L}(\theta)$ [CasellaBerger §6.3.2].

3.9 Maximum likelihood: pick the value that needs the least unusual noise

For fixed data, the likelihood $\mathcal{L}(\theta)$ of Section 3.8 ranks candidate parameter values by the probability or density they assign to the observation. Selecting the value that maximises this quantity defines the **maximum likelihood estimator**. This rule underlies many measurements in experimental physics. Almost every fitted mass, lifetime, cross-section and cosmological parameter you have seen was obtained this way, or by its Gaussian special case, least squares.

Turning the idea into recipes raises three questions.

1. How do we find the maximum when we can do the calculus by hand, and what do the answers look like for the standard models?
2. How does it compare with the older recipe of matching moments?
3. How do we find the maximum when we cannot do the calculus, and how does that look in Python?

The recipe is worth using because, for large samples, it homes in on the truth, is nearly unbiased, and is as precise as any estimator can be. These properties are proved in Section 3.13. Its error bar comes from the width of the likelihood (Section 3.12).

Working definition 3.11 (maximum likelihood estimator)

The **maximum likelihood estimator** (MLE) $\hat{\theta}$ is the value of θ that maximises $\mathcal{L}(\theta)$, or equivalently $\ell(\theta) = \ln \mathcal{L}(\theta)$ [Wasserman p. 122, Def. 9.8]. When the maximum is inside the allowed range and ℓ is smooth, it solves the **score equation**

$$S(\theta) \equiv \frac{\partial \ell}{\partial \theta} = 0, \qquad \frac{\partial^2 \ell}{\partial \theta^2} < 0,$$

where the **score** $S(\theta)$ [the slope of the log-likelihood, a random quantity because it depends on the data] measures how much the data pull θ up or down. For several parameters every component of the gradient vanishes and the matrix of second derivatives is negative definite.

Before differentiating we may throw away every factor of $\mathcal{L}$ that does not contain θ. Two facts allow it. First, ln is increasing, so $\mathcal{L}$ and ℓ peak at the same θ. Second, a constant factor in $\mathcal{L}$ becomes an added constant in ℓ, and the derivative of a constant is zero.

3.9.1 Solving the score equation by hand

Example: the mean lifetime of an unstable particle. We record n decay times $t_1, \dots, t_n$ from $f(t;\tau) = \tau^{-1} e^{-t/\tau}$ [Cowan §6.2]. Then

$$\ell(\tau) \overset{(1)}{=} \sum_{i=1}^{n} \left(-\ln \tau - \frac{t_i}{\tau} \right) = -n \ln \tau - \frac{T}{\tau}, \qquad T \equiv \sum_i t_i,$$

$$S(\tau) \overset{(2)}{=} -\frac{n}{\tau} + \frac{T}{\tau^2} \overset{(3)}{=} 0 \quad\Longrightarrow\quad \hat\tau = \frac{T}{n} = \bar t, \tag{3.82}$$
$$\ell''(\hat\tau) \overset{(4)}{=} \frac{n}{\hat\tau^2} - \frac{2T}{\hat\tau^3} \overset{(5)}{=} -\frac{n}{\hat\tau^2} < 0.$$

(1) Logarithm of the product (3.80): $\ln(\tau^{-1}e^{-t/\tau}) = -\ln\tau - t/\tau$.
(2) Differentiate term by term: $\mathrm{d}(-n\ln\tau)/\mathrm{d}\tau = -n/\tau$ and $\mathrm{d}(-T/\tau)/\mathrm{d}\tau = T/\tau^2$.
(3) The score equation. Multiply by τ^2 to get $-n\tau + T = 0$.
(4) Differentiate once more: $\mathrm{d}(-n/\tau)/\mathrm{d}\tau = n/\tau^2$, $\mathrm{d}(T/\tau^2)/\mathrm{d}\tau = -2T/\tau^3$.
(5) At $\hat\tau$ we have $T = n\hat\tau$, so $2T/\hat\tau^3 = 2n/\hat\tau^2$.

The maximum likelihood lifetime is the average decay time. It is unbiased, $\mathbb{E}\,\bar t = \tau$, and its variance is τ^2/n, because the exponential has variance τ^2 and averaging n independent times divides the variance by n. This is the same estimator whose sampling distribution we simulated for the muon lifetime in Section 3.1.3. The number $-\ell''(\hat\tau) = n/\hat\tau^2$ is one over the square of the error bar, as Section 3.12 shows.

Example: the centre and width of a Gaussian. Let $x_i \sim \mathcal{N}(\mu, \sigma^2)$ with both parameters unknown [Wasserman p. 123, Ex. 9.11], [Cowan §6.3], [CasellaBerger Ex. 7.2.11]. Dropping the constant $-\frac{n}{2}\ln 2\pi$,

$$\ell(\mu,\sigma) \overset{(1)}{=} -n\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2,$$
$$\frac{\partial\ell}{\partial\mu} \overset{(2)}{=} \frac{1}{\sigma^2}\sum_i (x_i-\mu) = 0 \implies \hat\mu = \bar x,$$
$$\frac{\partial\ell}{\partial\sigma} \overset{(3)}{=} -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_i (x_i-\mu)^2 = 0 \implies \hat\sigma^2 = \frac{1}{n}\sum_i (x_i-\bar x)^2.$$

(1) ln of $\prod_i (2\pi\sigma^2)^{-1/2} e^{-(x_i-\mu)^2/2\sigma^2}$, with $\ln(\sigma^2)^{-n/2} = -n\ln\sigma$.
(2) The chain rule on $(x_i-\mu)^2$ gives $-2(x_i-\mu)$. The sum vanishes when $n\mu = \sum_i x_i$.
(3) Differentiate $-n\ln\sigma$ and σ^{-2}; multiply the equation by σ^3/n; and insert the solution $\mu = \bar x$ of the first equation, because both equations must hold at once.

The centre is the sample mean, as one would guess. The width divides by n, not $n-1$, and so it is *biased*. In Section 3.4.1 we showed that $\mathbb{E}\sum_i (x_i-\bar x)^2 = (n-1)\sigma^2$. Dividing by n gives $\mathbb{E}\,\hat\sigma^2 = \frac{n-1}{n}\sigma^2$: the MLE is low by a factor $(n-1)/n$. The bias is the price of estimating μ from the same data, and it vanishes as $n \to \infty$. Maximum likelihood promises good behaviour for large samples, not unbiasedness at every n.

Example: the firing probability of a detector. For the Bernoulli likelihood $p^S(1-p)^{n-S}$ of Section 3.8.2,

$$\ell(p) = S\ln p + (n-S)\ln(1-p), \qquad S(p) = \frac{S}{p} - \frac{n-S}{1-p} = 0 \implies \hat p = \frac{S}{n},$$

because multiplying the score equation by $p(1-p)$ gives $S(1-p) - (n-S)p = S - np = 0$. For $S = 12$ in $n = 20$ trials, $\hat p = 0.6$, the peak seen in Figure 3.31.

Example: when the maximum sits on an edge. For the uniform box the likelihood θ^{-n} decreases for all $\theta \geq x_{(n)}$ and is zero below, so the maximum is at the cliff, $\hat\theta = x_{(n)}$ [Wasserman p. 124]. The score equation $-n/\theta = 0$ has no solution at all; the maximum is found by

looking, not by differentiating. The estimator is biased low, because the largest of n points always falls short of the end of the box: $\mathbb{E}\, x_{(n)} = \frac{n}{n+1}\theta$. A related case arises whenever physics restricts the parameter. A signal strength μ cannot be negative. If we measure $x_i \sim \mathcal{N}(\mu, \sigma^2)$ and find $\bar{x} < 0$, the likelihood restricted to $\mu \geq 0$ is largest at the edge, so $\hat{\mu} = 0$ [CasellaBerger Ex. 7.2.8]. The derivative there is not zero, and several of the nice properties of Section 3.13 fail at such boundaries.

Watch out: The MLE is not "the most probable value of θ"

$\hat{\theta}$ is the value under which the *data* are most probable. It is the most probable value of θ only in a Bayesian analysis with a flat prior, and then only in the parametrisation in which the prior is flat [Sivia p. 64]. Because the likelihood carries no Jacobian, the MLE itself does not care about the parametrisation: $\hat{\lambda} = 1/\hat{\tau}$ (Section 3.13). The peak of a posterior *density* does move under a change of variables.

3.9.2 The method of moments, for comparison

Before maximum likelihood there was a simpler idea. A model with k parameters predicts the first k moments $\alpha_j(\theta) = \mathbb{E}_\theta\, X^j$ [the average of the j-th power of the data, as a function of the parameters]. The data give the sample moments $\hat{\alpha}_j = \frac{1}{n}\sum_i x_i^j$. The **method of moments** estimator sets the two equal and solves, $\alpha_j(\tilde{\theta}) = \hat{\alpha}_j$ for $j = 1, \ldots, k$ [Wasserman p. 121, Def. 9.3], [Cowan Ch. 8]. For the Bernoulli model, $\mathbb{E}\, X = p$ gives $\tilde{p} = \bar{x} = S/n$; for the Gaussian, $\mathbb{E}\, X = \mu$ and $\mathbb{E}\, X^2 = \sigma^2 + \mu^2$ give $\tilde{\mu} = \bar{x}$ and $\tilde{\sigma}^2 = \frac{1}{n}\sum x_i^2 - \bar{x}^2 = \frac{1}{n}\sum (x_i - \bar{x})^2$ [Wasserman p. 121, Exs. 9.4–9.5]. Both coincide with the MLE. Under mild conditions the moment estimator is consistent and asymptotically Gaussian [Wasserman p. 122, Thm. 9.6], but in general it throws information away, and the MLE is at least as precise for large samples (Section 3.13).

Example: an angular distribution. In $e^+e^- \to \mu^+\mu^-$ the distribution of $x = \cos\theta$ of the outgoing muon has the shape $1 + \alpha x + \beta x^2$, where β comes from the photon exchange and α, the forward–backward asymmetry, from its interference with the Z. The detector accepts only $x_{\min} \leq x \leq x_{\max}$, so the density is [Cowan §6.8, (8.9)]

$$f(x;\alpha,\beta) = \frac{1 + \alpha x + \beta x^2}{d_1 + \alpha d_2 + \beta d_3}, \qquad d_k \equiv \frac{x_{\max}^k - x_{\min}^k}{k}, \tag{3.83}$$

because $\int_{x_{\min}}^{x_{\max}} (1 + \alpha x + \beta x^2)\,\mathrm{d}x = d_1 + \alpha d_2 + \beta d_3$. The first two moments are ratios of the same kind,

$$\mathbb{E}\, x \overset{(1)}{=} \frac{d_2 + \alpha d_3 + \beta d_4}{d_1 + \alpha d_2 + \beta d_3}, \qquad \mathbb{E}\, x^2 \overset{(2)}{=} \frac{d_3 + \alpha d_4 + \beta d_5}{d_1 + \alpha d_2 + \beta d_3},$$

(1) $\int x(1 + \alpha x + \beta x^2)\,\mathrm{d}x = d_2 + \alpha d_3 + \beta d_4$.
(2) $\int x^2(1 + \alpha x + \beta x^2)\,\mathrm{d}x = d_3 + \alpha d_4 + \beta d_5$.

and setting them equal to $\hat{\alpha}_1 = \overline{x}$ and $\hat{\alpha}_2 = \overline{x^2}$ and multiplying out gives two *linear* equations for (α, β),

$$\alpha(\hat{\alpha}_1 d_2 - d_3) + \beta(\hat{\alpha}_1 d_3 - d_4) = d_2 - \hat{\alpha}_1 d_1, \qquad \alpha(\hat{\alpha}_2 d_2 - d_4) + \beta(\hat{\alpha}_2 d_3 - d_5) = d_3 - \hat{\alpha}_2 d_1.$$

For 2000 simulated events with $\alpha = \beta = 0.5$ on $-0.95 \leq x \leq 0.95$ (Cowan's set-up), these give $(\tilde{\alpha}, \tilde{\beta}) = (0.49, 0.4458)$. The maximum likelihood equations for the same model are not linear, so they need a numerical method.

3.9.3 When the calculus cannot be done: Newton–Raphson

For (3.83) the log-likelihood is $\ell(\alpha,\beta)=\sum_i \ln(1+\alpha x_i+\beta x_i^2)-n\ln(d_1+\alpha d_2+\beta d_3)$, and its score equations have no closed-form solution. The standard numerical maximiser replaces ℓ near a trial point by a parabola and jumps to the parabola's top. The same parabola gives the error bar in Section 3.12. Near a trial point θ_j,

$$\ell(\theta)\overset{(1)}{=}\ell(\theta_j)+\ell'(\theta_j)(\theta-\theta_j)+\tfrac12\ell''(\theta_j)(\theta-\theta_j)^2+\dots,$$
$$0\overset{(2)}{=}\ell'(\theta_j)+\ell''(\theta_j)(\theta_{j+1}-\theta_j)\quad\Longrightarrow\quad \theta_{j+1}=\theta_j-\frac{\ell'(\theta_j)}{\ell''(\theta_j)}. \tag{3.84}$$

(1) Taylor expansion to second order: we replace ℓ by the parabola that has the same value, slope and curvature at θ_j.
(2) Jump to the top of that parabola, where its derivative vanishes.

With several parameters, ℓ' becomes the gradient $\boldsymbol{g}$ and ℓ'' the matrix of second derivatives H (the **Hessian**), and the step is $\boldsymbol{\theta}_{j+1}=\boldsymbol{\theta}_j-H^{-1}\boldsymbol{g}$. Near the maximum the error shrinks quadratically, roughly squaring at every step, so a handful of iterations suffice. Figure 3.33 shows the first two steps on a lifetime likelihood.

Example: Newton by hand. Take $n=5$ decays with total time $T=6$, so that $\ell(\tau)=-5\ln\tau-6/\tau$, $\ell'=-5/\tau+6/\tau^2$, $\ell''=5/\tau^2-12/\tau^3$, and the exact answer is $\hat\tau=T/n=1.2$. Start at $\tau_0=0.8$:

$$\tau_1\overset{(1)}{=}0.8-\frac{-6.25+9.375}{7.8125-23.4375}=0.8+\frac{3.125}{15.625}=1.0,$$
$$\tau_2\overset{(2)}{=}1.0-\frac{-5+6}{5-12}=1+\frac17=1.1429,$$
$$\tau_3\overset{(3)}{=}1.1429+\frac{0.2188}{4.2109}=1.1948,\qquad \tau_4=1.19996,\qquad \tau_5=1.2000.$$

(1) (3.84) with $\ell'(0.8)=-5/0.8+6/0.64$ and $\ell''(0.8)=5/0.64-12/0.512$.
(2) The same at $\tau_1=1$.
(3) The same at τ_2, where $\ell'=0.2188$ and $\ell''=-4.2109$.

The distance to the answer goes $0.4,\ 0.2,\ 0.057,\ 0.0052,\ 0.00005$: the number of correct digits roughly doubles per step once we are close.

Newton's method trusts the parabola, and the parabola can lie. Here $\ell''=5/\tau^2-12/\tau^3$ is negative only when $5\tau<12$, that is for $\tau<2T/n=2.4$. Beyond that the curve bends upwards and the parabola has a minimum, not a maximum. Started at $\tau_0=3$, the step therefore points *away* from the maximum. Practical minimisers (the `MINUIT` of particle physics, `scipy.optimize`) therefore check that each step increases ℓ, shorten it if not, and start from a sensible guess. The method of moments is an excellent source of such guesses.

The angular fit in Python. The script `10_angular_ml.py` generates 2000 events from (3.83) by accept–reject (Chapter 6), computes the moment estimate, and hands it to Newton's method. The heart of it is the analytic gradient and Hessian,

$$\frac{\partial\ell}{\partial\alpha}=\sum_i\frac{x_i}{g_i}-\frac{n\,d_2}{D},\qquad \frac{\partial^2\ell}{\partial\alpha^2}=-\sum_i\frac{x_i^2}{g_i^2}+\frac{n\,d_2^2}{D^2},\qquad g_i=1+\alpha x_i+\beta x_i^2,\quad D=d_1+\alpha d_2+\beta d_3,$$

and the same with $x_i\to x_i^2$, $d_2\to d_3$ for β (the mixed derivative has x_i^3 and $d_2 d_3$). These follow from $\mathrm{d}\ln g/\mathrm{d}\alpha=x/g$ and $\mathrm{d}\ln D/\mathrm{d}\alpha=d_2/D$ and one more derivative of each.

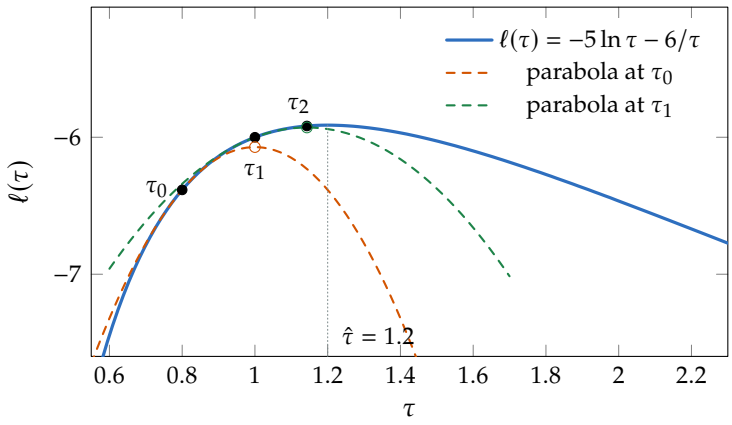


Figure 3.33: Newton–Raphson on a log-likelihood. At the starting point τ_0 the method replaces ℓ by the parabola (orange) with the same height, slope and curvature, and jumps to its top (open circle), which lies straight above τ_1. Repeating from τ_1 (green parabola) lands at τ_2, already close to the true maximum at $\hat\tau = 1.2$ (dotted line). Each parabola hugs the curve better than the last because the curve looks more and more like a parabola near its peak.

Listing 3.8: Newton–Raphson for the angular distribution, and the covariance from the Hessian (`code/ch03/10_angular_ml.py`)

```
def newton(x, th0, tol=1e-10, maxit=50):
    th, path = np.array(th0, float), [np.array(th0, float)]
    for it in range(maxit):
        g, H = grad_hess(th, x)
        step = -np.linalg.solve(H, g)                # theta_{j+1} = theta_j - H^{-1} grad
        th = th + step
        path.append(th.copy())
        if np.max(np.abs(step)) < tol:
            break
    return th, np.array(path)


# ---------- one experiment ----------
x = sample(n)
th_mom = mom(x)
th_ml, path = newton(x, th_mom)
_, H = grad_hess(th_ml, x)
V = np.linalg.inv(-H)                                # covariance = inverse of minus the
    Hessian
sa, sb = np.sqrt(np.diag(V))
rho = V[0, 1] / (sa * sb)
```

The loop is (3.84) in matrix form: `np.linalg.solve(H, g)` computes $H^{-1}g$ without forming the inverse. The last four lines compute the covariance of the estimates as the inverse of minus the Hessian at the maximum, a rule derived in Section 3.12. Starting from the moment estimate, Newton reaches $(\hat\alpha, \hat\beta) = (0.4885, 0.4396)$ in 4 iterations; after the first step the answer already agrees with the final one far better than its error bar. The left panel of Figure 3.34 shows the fitted density on the histogram of the events.

Repeating the whole experiment 500 times, the maximum likelihood estimates average $(0.502, 0.4986)$ (truth $(0.5, 0.5)$) and scatter by $\sigma_{\hat\alpha} = 0.05376$ and $\sigma_{\hat\beta} = 0.1081$. The moment estimates scatter by 0.05422 and 0.1109. Both methods were applied to the same 500 data sets, so the ratios of the spreads are measured well: 1.009 ± 0.007096 for α and 1.026 ± 0.009865 for β. For this mild distribution the two methods are close; the gap widens for sharply peaked densities, where low moments are poor summaries of the shape.

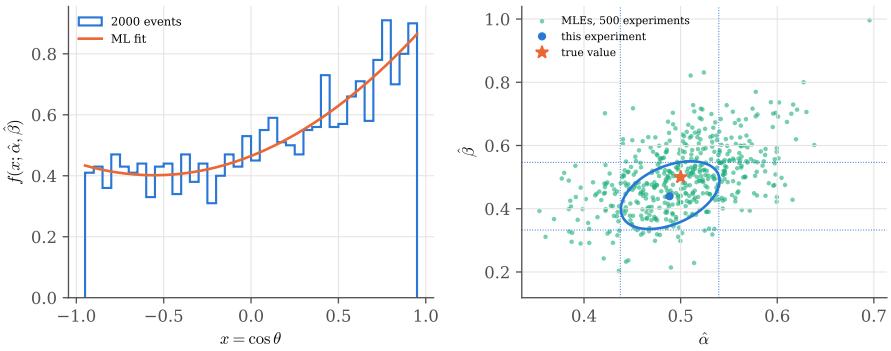


Figure 3.34: A two-parameter maximum likelihood fit. Left: 2000 simulated values of $\cos\theta$ and the fitted density $1+\hat\alpha x+\hat\beta x^2$ (normalised). Right: the fitted $(\hat\alpha,\hat\beta)$ of 500 independent repetitions of the experiment (green), the fit of the first experiment (blue dot), the truth (star), and the contour on which $\ln\mathcal{L}$ of the first experiment has dropped by $\frac{1}{2}$ from its maximum. The dotted lines tangent to the contour mark $\hat\alpha\pm\sigma_{\hat\alpha}$ and $\hat\beta\pm\sigma_{\hat\beta}$ from the inverse Hessian. The cloud of repetitions has the size, tilt and shape of the contour.

When the number of events is itself random. In a counting experiment the total number of events n is not fixed in advance: it is Poisson distributed with a mean ν that the theory may predict (a cross-section times a luminosity). Then the likelihood acquires one more factor [Cowan §6.9],

$$\mathcal{L}(\nu,\theta)=\frac{\nu^n e^{-\nu}}{n!}\prod_{i=1}^{n}f(x_i;\theta)=\frac{e^{-\nu}}{n!}\prod_{i=1}^{n}\nu f(x_i;\theta), \tag{3.85}$$

the **extended likelihood**. If ν is a free parameter unrelated to θ, the ν-dependent part of ℓ is $n\ln\nu-\nu$, whose derivative $n/\nu-1$ vanishes at $\hat\nu=n$. The θ-dependent part is the same as before, so $\hat\theta$ is unchanged. If $\nu=\nu(\theta)$, the total rate carries information about θ in addition to the shape, and the extended likelihood uses both. The product $\nu f(x;\theta)$ is the expected number of events per unit x, which is why the extended likelihood is the natural starting point for histograms of counts (Section 3.11).

Hidden variables. Some likelihoods are easy to maximise only if we knew something we were not told, for example which of two Gaussian populations each point came from (a mixture model). The **EM algorithm** alternates between guessing the hidden labels from the current parameters (the "E" step, an expectation) and maximising the likelihood as if the labels were known (the "M" step). No cycle can decrease $\mathcal{L}$ [Wasserman pp. 143–145]. It is the standard tool for fitting mixtures.

3.10 Gaussian noise: least squares and the χ^2

Every physics student has fitted a straight line through points with error bars by "minimising χ^2". The rule is usually taught without a reason. Why squares and not absolute values? Why divide by σ_i^2? What is the error on the slope, and how do we know whether the line is any good? All of this follows in a few lines from the likelihood of Section 3.8, once we assume that the noise is Gaussian. The same lines then cover noise that is correlated from point to point, which is the normal situation for binned power spectra and for gravitational-wave strain. Four questions need answers.

1. Why is least squares the maximum likelihood estimator for Gaussian noise, and when is it not?

2. For a model linear in its parameters, what are the estimates and their covariance, exactly?
3. What value of $\chi^2_{\min}$ should a correct model give, and how do we turn it into a test of the fit?
4. What changes when the noise is correlated, and what goes wrong if we ignore the correlation?

3.10.1 Least squares is Gaussian maximum likelihood

Take data $D_i = d_i(\boldsymbol{\theta}) + \epsilon_i$, $i = 1, \dots, N$: the model prediction $d_i(\boldsymbol{\theta})$ plus independent noise $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ with known σ_i. Its log-likelihood is (3.79) of Section 3.8.2. Written with the **normalised residuals** $R_i = (D_i - d_i(\boldsymbol{\theta}))/\sigma_i$ it reads [Sivia p. 65, (3.65)–(3.67)]

$$\ell(\boldsymbol{\theta}) \overset{(1)}{=} -\frac{1}{2}\sum_{i=1}^{N} R_i^2 - \sum_{i=1}^{N} \ln \sqrt{2\pi}\,\sigma_i \overset{(2)}{=} -\frac{1}{2}\chi^2(\boldsymbol{\theta}) + \text{const}, \qquad \chi^2(\boldsymbol{\theta}) \equiv \sum_{i=1}^{N} \frac{\big(D_i - d_i(\boldsymbol{\theta})\big)^2}{\sigma_i^2}. \tag{3.86}$$

(1) (3.79), with the residuals in units of their error bars.
(2) The error bars are known numbers, so the second sum does not depend on $\boldsymbol{\theta}$.

Maximising ℓ is minimising χ^2, because of the minus sign. So the **least-squares estimator** is the maximum likelihood estimator [Cowan §7.1], [HobsonBMC p. 50, (2.24)–(2.25)]. This answers the two "why" questions. The squares come from the exponent of the Gaussian. The $1/\sigma_i^2$ measures each residual against its own error bar: a residual of 0.3 is a one-sigma event for a point with $\sigma_i = 0.3$ and a ten-sigma event for a point with $\sigma_i = 0.03$. With equal error bars all points count equally. For a constant model, $\chi^2 = \sum_i (D_i - \mu)^2/\sigma_i^2$, and minimising it over μ gives the inverse-variance weighted mean of Section 3.7.

The derivation also says when least squares is *not* maximum likelihood. It used two assumptions, and each can fail. If the noise is not Gaussian (heavy tails, outliers, counts), the exponent is not a square. If the error bars depend on the parameters, for example $\sigma_i^2 = d_i(\boldsymbol{\theta})$ for counts, the term $\sum \ln \sigma_i$ depends on $\boldsymbol{\theta}$ and cannot be dropped. So least squares is what the likelihood reduces to under these two assumptions. When it misbehaves, the cure is to go back to the likelihood, not to patch the recipe [Sivia p. 66].

3.10.2 Linear models: the normal equations

Many models are linear in their parameters even when they are not linear in x: a straight line $a + bx$, a polynomial, a sum of templates with unknown amplitudes. Write such a model as $d_i(\boldsymbol{\theta}) = \sum_{k=1}^{m} A_{ik}\theta_k$, or $\boldsymbol{d} = A\boldsymbol{\theta}$, where the $N \times m$ **design matrix** A [column k is the model curve for $\theta_k = 1$ and all other parameters zero] is known. For a line, row i of A is $(1,\ x_i)$. Collect the error bars into the diagonal covariance $\mathbf{C} = \mathrm{diag}(\sigma_1^2, \dots, \sigma_N^2)$; everything below also holds for a general $\mathbf{C}$, which we use in Section 3.10.5. Then

$$\begin{aligned} \chi^2(\boldsymbol{\theta}) &\overset{(1)}{=} (\boldsymbol{D} - A\boldsymbol{\theta})^{\mathsf{T}}\mathbf{C}^{-1}(\boldsymbol{D} - A\boldsymbol{\theta}), \\ \frac{\partial \chi^2}{\partial \boldsymbol{\theta}} &\overset{(2)}{=} -2A^{\mathsf{T}}\mathbf{C}^{-1}(\boldsymbol{D} - A\boldsymbol{\theta}) \overset{(3)}{=} \mathbf{0} \quad\Longrightarrow\quad (A^{\mathsf{T}}\mathbf{C}^{-1}A)\,\hat{\boldsymbol{\theta}} = A^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{D}, \\ \hat{\boldsymbol{\theta}} &\overset{(4)}{=} (A^{\mathsf{T}}\mathbf{C}^{-1}A)^{-1}A^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{D} \;\equiv\; B\boldsymbol{D}. \end{aligned} \tag{3.87}$$

(1) With a diagonal $\mathbf{C}$ this is $\sum_i (D_i - d_i)^2/\sigma_i^2$ written as a matrix product.
(2) Differentiate the quadratic form: for a symmetric matrix M, $\partial(\boldsymbol{v}^{\mathsf{T}}M\boldsymbol{v})/\partial\boldsymbol{\theta} = 2(\partial\boldsymbol{v}/\partial\boldsymbol{\theta})^{\mathsf{T}}M\boldsymbol{v}$, and here $\partial\boldsymbol{v}/\partial\boldsymbol{\theta} = -A$.

(3) The score equation. These m linear equations are the **normal equations**.
(4) The matrix $A^{\mathsf{T}}\mathbf{C}^{-1}A$ is $m \times m$ and invertible as long as the columns of A are linearly independent (no two parameters do the same thing to the model).

The estimator is a fixed matrix B times the data. So its mean and covariance follow from linear error propagation, which is exact for linear maps (Section 3.6.2):

$$\begin{aligned} \mathbb{E}\,\hat{\boldsymbol{\theta}} &\overset{(1)}{=} B\,\mathbb{E}\,\boldsymbol{D} \overset{(2)}{=} (A^{\mathsf{T}}\mathbf{C}^{-1}A)^{-1}A^{\mathsf{T}}\mathbf{C}^{-1}A\boldsymbol{\theta} = \boldsymbol{\theta}, \\ \operatorname{Cov}\hat{\boldsymbol{\theta}} &\overset{(3)}{=} B\,\mathbf{C}\,B^{\mathsf{T}} \overset{(4)}{=} (A^{\mathsf{T}}\mathbf{C}^{-1}A)^{-1}A^{\mathsf{T}}\mathbf{C}^{-1}\,\mathbf{C}\,\mathbf{C}^{-1}A(A^{\mathsf{T}}\mathbf{C}^{-1}A)^{-1} \overset{(5)}{=} (A^{\mathsf{T}}\mathbf{C}^{-1}A)^{-1} \equiv U. \end{aligned} \tag{3.88}$$

(1) Expectation is linear.
(2) The noise has mean zero, so $\mathbb{E}\,\boldsymbol{D} = A\boldsymbol{\theta}$.
(3) Covariance of a linear map, $\operatorname{Cov}(B\boldsymbol{D}) = B\operatorname{Cov}(\boldsymbol{D})B^{\mathsf{T}}$.
(4) Insert B and B^{T}; $\mathbf{C}$ and $A^{\mathsf{T}}\mathbf{C}^{-1}A$ are symmetric.
(5) $\mathbf{C}^{-1}\mathbf{C} = \mathbb{1}$, and then $(A^{\mathsf{T}}\mathbf{C}^{-1}A)^{-1}(A^{\mathsf{T}}\mathbf{C}^{-1}A) = \mathbb{1}$.

So for a linear model with Gaussian noise the least-squares estimates are unbiased and have covariance $U = (A^{\mathsf{T}}\mathbf{C}^{-1}A)^{-1}$, with no approximation. Now look at the curvature of χ^2. Differentiating (3.87) once more gives $\partial^2\chi^2/\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^{\mathsf{T}} = 2A^{\mathsf{T}}\mathbf{C}^{-1}A = 2U^{-1}$. Because χ^2 is exactly quadratic in $\boldsymbol{\theta}$, its Taylor series stops there:

$$\chi^2(\boldsymbol{\theta}) = \chi^2_{\min} + (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^{\mathsf{T}}U^{-1}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}), \qquad \ell(\boldsymbol{\theta}) = \ell_{\max} - \tfrac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^{\mathsf{T}}U^{-1}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}). \tag{3.89}$$

The likelihood is an exact Gaussian in $\boldsymbol{\theta}$, centred on $\hat{\boldsymbol{\theta}}$, whose width matrix is the covariance of the estimator. In one dimension $\chi^2(\theta) = \chi^2_{\min} + (\theta-\hat{\theta})^2/\sigma^2_{\hat{\theta}}$, so the parameter values at which χ^2 has risen by one from its minimum are exactly $\hat{\theta} \pm \sigma_{\hat{\theta}}$ [Cowan §7.2]. This is the model case for the rule "the curvature of the log-likelihood is the error", which Section 3.12 extends to every smooth likelihood.

Key idea: Linear least squares

For data $\boldsymbol{D} = A\boldsymbol{\theta} + \boldsymbol{\epsilon}$ with Gaussian noise of covariance $\mathbf{C}$,

$$\hat{\boldsymbol{\theta}} = (A^{\mathsf{T}}\mathbf{C}^{-1}A)^{-1}A^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{D}, \qquad \operatorname{Cov}\hat{\boldsymbol{\theta}} = (A^{\mathsf{T}}\mathbf{C}^{-1}A)^{-1},$$

the estimator is unbiased and Gaussian, and χ^2 rises by exactly one at $\hat{\theta}_k \pm \sigma_{\hat{\theta}_k}$ when the other parameters are re-minimised. Nothing here is asymptotic.

Example: a straight line through three points. Fit $y = a + bx$ to $(x_i, y_i) = (0, 1.1), (1, 1.9), (2, 3.2)$, all with $\sigma = 0.1$. For a line the normal equations are 2×2, and it pays to name the sums $S = \sum 1/\sigma_i^2$, $S_x = \sum x_i/\sigma_i^2$, $S_{xx} = \sum x_i^2/\sigma_i^2$, $S_y = \sum y_i/\sigma_i^2$, $S_{xy} = \sum x_i y_i/\sigma_i^2$ and $\Delta = SS_{xx} - S_x^2$. In units of $1/\sigma^2 = 100$ they are $S = 3$, $S_x = 3$, $S_{xx} = 5$, $S_y = 6.2$, $S_{xy} = 8.3$. Then

$$\begin{aligned} \begin{pmatrix}\hat{a}\\ \hat{b}\end{pmatrix} &\overset{(1)}{=} \begin{pmatrix} S & S_x \\ S_x & S_{xx}\end{pmatrix}^{-1}\begin{pmatrix} S_y \\ S_{xy}\end{pmatrix} \overset{(2)}{=} \frac{1}{\Delta}\begin{pmatrix} S_{xx}S_y - S_xS_{xy} \\ SS_{xy} - S_xS_y\end{pmatrix} = \frac{1}{6}\begin{pmatrix} 31 - 24.9 \\ 24.9 - 18.6\end{pmatrix} = \begin{pmatrix} 1.017 \\ 1.050\end{pmatrix}, \\ U &\overset{(3)}{=} \frac{1}{\Delta}\begin{pmatrix} S_{xx} & -S_x \\ -S_x & S\end{pmatrix} = \sigma^2\begin{pmatrix} 5/6 & -1/2 \\ -1/2 & 1/2\end{pmatrix}, \qquad \sigma_{\hat{a}} = 0.091, \quad \sigma_{\hat{b}} = 0.071, \quad \rho = -0.775. \end{aligned}$$

(1) The normal equations (3.87): with rows $(1, x_i)$, $A^{\mathsf{T}}\mathbf{C}^{-1}A$ has entries S, S_x, S_x, S_{xx} and $A^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{y} = (S_y, S_{xy})^{\mathsf{T}}$.
(2) Inverse of a 2×2 matrix: swap the diagonal, negate the off-diagonal, divide by the determinant Δ. Here $\Delta = 3 \cdot 5 - 3^2 = 6$ in units of σ^{-4}; the powers of σ cancel in $\hat{a}, \hat{b}$.

(3) $U = (A^\mathsf{T}\mathbf{C}^{-1}A)^{-1}$ from (3.88); restoring the units, $1/\Delta = \sigma^4/6$ and the matrix entries carry $1/\sigma^2$.

The intercept and slope are strongly anticorrelated, $\rho = -(1/2)/\sqrt{(5/6)(1/2)}$: if the fitted slope is a little too steep, the line must start a little lower to pass through the middle of the data. Choosing the origin of x at the weighted mean, $\bar{x} = S_x/S = 1$, would make $S_x = 0$ and the two estimates uncorrelated. The residuals are $y_i - \hat{a} - \hat{b}x_i = (0.083, -0.167, 0.083)$, so $\chi^2_{\min} = (0.0069 + 0.0278 + 0.0069)/0.01 = 4.17$. Is that good or bad?

3.10.3 How good is the fit? The distribution of $\chi^2_{\min}$

If the model is right, every normalised residual is a standard Gaussian before fitting, and $\chi^2(\boldsymbol{\theta}_{\text{true}})$ is a sum of N squared standard Gaussians, so it follows the χ^2_N distribution (Chapter 2). Fitting lowers it: the fitted curve bends towards the data and eats part of the noise. By how much? The cleanest way to see it is geometric.

Whiten the data first. Factor the covariance as $\mathbf{C} = LL^\mathsf{T}$ (for a diagonal $\mathbf{C}$, simply $L = \operatorname{diag}(\sigma_i)$; in general the Cholesky factor). Then $\boldsymbol{z} = L^{-1}\boldsymbol{D}$, $\tilde{A} = L^{-1}A$ and $\boldsymbol{\eta} = L^{-1}\boldsymbol{\epsilon}$ satisfy $\boldsymbol{z} = \tilde{A}\boldsymbol{\theta} + \boldsymbol{\eta}$ with $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \mathbb{1}_N)$, and $\chi^2 = |\boldsymbol{z} - \tilde{A}\boldsymbol{\theta}|^2$ is an ordinary squared distance in $\mathbb{R}^N$. The model predictions $\tilde{A}\boldsymbol{\theta}$, as $\boldsymbol{\theta}$ ranges over $\mathbb{R}^m$, fill an m-dimensional flat subspace (the column space of $\tilde{A}$). Minimising the distance means dropping a perpendicular: the fit is the orthogonal projection of $\boldsymbol{z}$ onto that subspace (Figure 3.35). With the projector $P = \tilde{A}(\tilde{A}^\mathsf{T}\tilde{A})^{-1}\tilde{A}^\mathsf{T}$ [the matrix that maps any vector to its closest point in the model subspace. It satisfies $P^2 = P$ and $P^\mathsf{T} = P$],

$$\begin{aligned} \chi^2_{\min} &\overset{(1)}{=} \big|(1-P)\boldsymbol{z}\big|^2 \overset{(2)}{=} \big|(1-P)\boldsymbol{\eta}\big|^2 \overset{(3)}{=} \boldsymbol{\eta}^\mathsf{T}(\mathbb{1}-P)\boldsymbol{\eta}, \\ \mathbb{E}\,\chi^2_{\min} &\overset{(4)}{=} \operatorname{tr}\big[(\mathbb{1}-P)\;\mathbb{E}(\boldsymbol{\eta}\boldsymbol{\eta}^\mathsf{T})\big] \overset{(5)}{=} \operatorname{tr}(\mathbb{1}_N) - \operatorname{tr} P \overset{(6)}{=} N - m \equiv \nu. \end{aligned} \tag{3.90}$$

(1) The fitted point is $P\boldsymbol{z}$, so the residual vector is $\boldsymbol{z} - P\boldsymbol{z}$.
(2) $\boldsymbol{z} = \tilde{A}\boldsymbol{\theta} + \boldsymbol{\eta}$, and $(\mathbb{1} - P)\tilde{A} = 0$ because P leaves every vector of the model subspace unchanged. The true model drops out: only the noise is left.
(3) $|\boldsymbol{v}|^2 = \boldsymbol{v}^\mathsf{T}\boldsymbol{v}$ with $(\mathbb{1} - P)^\mathsf{T}(\mathbb{1} - P) = \mathbb{1} - P$, using $P^\mathsf{T} = P$ and $P^2 = P$.
(4) A number equals its own trace, $\boldsymbol{\eta}^\mathsf{T}M\boldsymbol{\eta} = \operatorname{tr}(M\boldsymbol{\eta}\boldsymbol{\eta}^\mathsf{T})$, and the trace is linear, so it commutes with $\mathbb{E}$.
(5) Whitened noise has $\mathbb{E}(\boldsymbol{\eta}\boldsymbol{\eta}^\mathsf{T}) = \mathbb{1}_N$.
(6) $\operatorname{tr} P = \operatorname{tr}[(\tilde{A}^\mathsf{T}\tilde{A})^{-1}\tilde{A}^\mathsf{T}\tilde{A}] = \operatorname{tr}\mathbb{1}_m = m$, by the cyclic property of the trace.

The fit removes on average exactly one unit of χ^2 per fitted parameter. The whole distribution follows too. P is a projector of rank m, so it has m eigenvalues 1 and $N - m$ eigenvalues 0, and $\mathbb{1} - P$ has the reverse. Rotate to the eigenbasis of P. A rotation of independent standard Gaussians gives independent standard Gaussians again, because $\mathcal{N}(\mathbf{0}, \mathbb{1})$ looks the same in every direction. In the new basis $\chi^2_{\min} = \boldsymbol{\eta}^\mathsf{T}(\mathbb{1} - P)\boldsymbol{\eta}$ keeps only the $N - m$ components with eigenvalue 1, so it is a sum of $N - m$ squared independent standard Gaussians:

$$\chi^2_{\min} \sim \chi^2_\nu, \qquad \nu = N - m, \qquad \mathbb{E}\,\chi^2_{\min} = \nu, \qquad \operatorname{Var}\chi^2_{\min} = 2\nu. \tag{3.91}$$

This is the argument that gave the divisor $N - 1$ in the sample variance (Section 3.4.1), where one parameter, the mean, was fitted. Now there are m. The number ν is the number of **degrees of freedom** of the fit.

Turning $\chi^2_{\min}$ into a test. Given (3.91), a correct model with correct error bars gives $\chi^2_{\min} \approx \nu \pm \sqrt{2\nu}$. The standard summary is the **goodness-of-fit p-value**, the probability that a correct

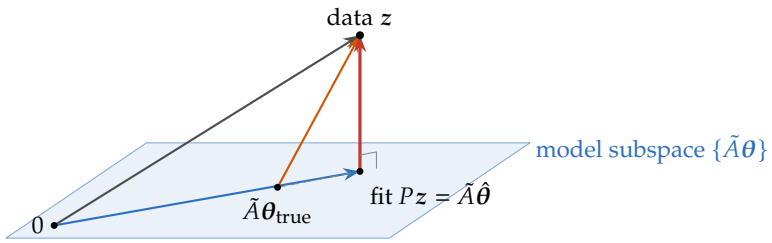


Figure 3.35: Least squares as a projection, after whitening so that the noise is the same in every direction. The data vector lies off the model subspace by the noise vector (orange). The fit drops a perpendicular onto the subspace. The residual (red) is the part of the noise perpendicular to the subspace; the part lying inside it has been absorbed into the fitted parameters, which is why the fit sits away from the truth along the plane. With m parameters, m of the N noise components are absorbed and $N - m$ are left in χ^2_{min}.

model would give a χ^2_{min} at least as large as the one observed, $p = \text{P}(\chi^2_\nu \geq \chi^2_{\text{min}})$ [Padilla §3.5], [Cowan §7.5]. A tiny p says that the data are hard to reconcile with the model and the error bars together: the model is wrong, or the errors are underestimated, or the noise is not Gaussian. A p very close to one ($\chi^2_{\text{min}} \ll \nu$) is also a warning: the errors are overestimated, or there are correlations that were ignored (Section 3.10.5), or the data were "cleaned" by hand. The ratio χ^2_{min}/ν, the **reduced** χ^2, should be near one, with a spread of $\sqrt{2/\nu}$. The full logic of tests, and what a p-value does and does not mean, is Chapter 4.

Example: the three-point line again. Two parameters from three points leave $\nu = 1$. The observed $\chi^2_{\text{min}} = 4.17$ has $p = \text{P}(\chi^2_1 \geq 4.17) = \text{P}(|Z| \geq 2.04) = 0.041$, where $Z \sim \mathcal{N}(0, 1)$ and we used that a χ^2_1 variable is the square of one standard Gaussian. Such a large χ^2 happens in about four per cent of experiments if the line and the error bars are both right. With one degree of freedom this is weak evidence, but it would be worth checking whether $\sigma = 0.1$ is honest.

Confidence regions from $\Delta\chi^2$. The quadratic form (3.89), evaluated at the true parameters, is $\chi^2(\boldsymbol{\theta}_{\text{true}}) - \chi^2_{\text{min}} = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\text{true}})^{\text{T}} U^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\text{true}})$, and since $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\text{true}} \sim \mathcal{N}(\mathbf{0}, U)$, it is a sum of m squared standard Gaussians (whiten with U as we whitened with $\mathbf{C}$): $\Delta\chi^2 \sim \chi^2_m$. The region $\{\boldsymbol{\theta} : \chi^2(\boldsymbol{\theta}) \leq \chi^2_{\text{min}} + \Delta\chi^2_\alpha\}$ therefore covers the truth in a fraction $1 - \alpha$ of repeated experiments if $\Delta\chi^2_\alpha$ is the corresponding quantile of χ^2_m [Padilla §3.6, Table III]:

coverage	$m = 1$	$m = 2$	$m = 3$
68.3%	1.00	2.30	3.53
95.4%	4.00	6.17	8.02
99.73%	9.00	11.8	14.2

The threshold depends on the number m of parameters in the joint region, not on the number N of data points.[2] For one parameter at a time, with the others re-minimised, $\Delta\chi^2 = 1$ gives the $\hat{\theta} \pm \sigma_{\hat{\theta}}$ interval found from (3.89). A joint two-parameter 68% region needs $\Delta\chi^2 = 2.30$, a larger ellipse. Since $\ell = -\frac{1}{2}\chi^2 + \text{const}$, the same ellipse is the contour on which $\ln \mathcal{L}$ has dropped by 2.30/2 from its maximum (Section 3.12).

[2] Padilla et al. write the recipe as $Q(n - M, \min \chi^2 + \Delta\chi^2) = p$ [Padilla (37)], which mixes the goodness-of-fit distribution (with $n - M$ degrees of freedom) with the distribution of $\Delta\chi^2$ (with M). Their Table III is correct and follows from $\text{P}(\chi^2_M \leq \Delta\chi^2) = p$.

3.10.4 Fitting a line in Python

The script `11_line_fit.py` fits $y = a + bx$ to $N = 10$ points at $x = 1, \dots, 10$ with error bars growing from 0.48 to 1.2, true values $a = 1$, $b = 0.5$. The whole of linear least squares is five lines.

Listing 3.9: Generalised least squares in five lines (`code/ch03/11_line_fit.py`)

```
def gls(y, C):
    """generalised least squares: theta = (A^T C^-1 A)^-1 A^T C^-1 y and its covariance"""
    Ci = np.linalg.inv(C)
    U = np.linalg.inv(A.T @ Ci @ A)
    th = U @ A.T @ Ci @ y.T
    r = y - (A @ th).T
    chi2 = np.einsum("...i,ij,...j->...", r, Ci, r)
    return th.T, U, chi2
```

Line by line: invert the covariance, form $U = (A^{\mathsf{T}}\mathbf{C}^{-1}A)^{-1}$, apply (3.87), form the residuals and $\chi^2 = \boldsymbol{r}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{r}$. The function accepts a whole stack of data sets at once (one per row of `y`), which is how we fit 10^4 simulated experiments in one call.

For the first data set the fit gives $\hat{a} = 1.238 \pm 0.4113$, $\hat{b} = 0.4187 \pm 0.08767$ with correlation -0.8223, and $\chi^2_{\min} = 11.01$ for $\nu = 8$, a goodness-of-fit $p = 0.2012$: an unremarkable fit. Over 10^4 repetitions the estimates scatter by 0.4067 and 0.08717 with correlation -0.8182, exactly as U predicts, and $\chi^2_{\min}$ has mean 7.932 and variance 15.69, against $\nu = 8$ and $2\nu = 16$ from (3.91). Fitting the wrong model, a constant, to the same data gives an average $\chi^2_{\min}$ of 41.36 for 9 degrees of freedom: the test sees the missing slope at once (Figure 3.36).

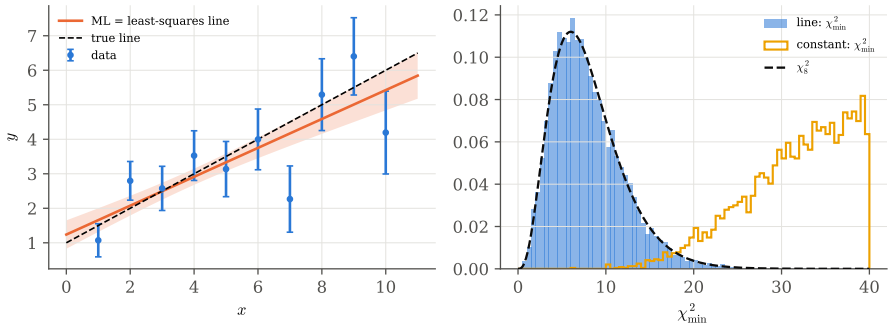


Figure 3.36: Least squares as Gaussian maximum likelihood. Left: ten points with error bars that grow to the right, the fitted line with its $\pm 1\sigma$ band from $U = (A^{\mathsf{T}}\mathbf{C}^{-1}A)^{-1}$ (narrowest near the weighted centre of the data), and the true line (dashed). Right: $\chi^2_{\min}$ of 10^4 simulated experiments for the correct model (filled), which follows the χ^2_8 curve, and for a constant fitted to the same data (outline), which sits far to the right.

3.10.5 Correlated noise: $\chi^2 = \boldsymbol{r}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{r}$

Noise is often correlated from one data point to the next. Neighbouring bins of a power spectrum share modes. A slowly drifting baseline pushes consecutive readings up together. A common calibration error moves every point at once. The noise is then a Gaussian vector $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$ with a non-diagonal covariance, and the likelihood is the multivariate normal density (Chapter 2) evaluated at the required noise $\boldsymbol{r} = \boldsymbol{D} - \boldsymbol{d}(\boldsymbol{\theta})$:

$$\mathcal{L}(\boldsymbol{\theta}) \overset{(1)}{=} p_{\boldsymbol{\eta}}\big(L^{-1}\boldsymbol{r}\big)\,\big|\det L^{-1}\big| \overset{(2)}{=} \frac{1}{(2\pi)^{N/2}}\, e^{-\frac{1}{2}(L^{-1}\boldsymbol{r})^{\mathsf{T}}(L^{-1}\boldsymbol{r})}\, \frac{1}{|\det L|}$$

$$\overset{(3)}{=} \frac{1}{\sqrt{(2\pi)^N \det \mathbf{C}}} \exp\!\left[-\tfrac{1}{2}\, \boldsymbol{r}^{\mathsf{T}} \mathbf{C}^{-1} \boldsymbol{r}\right], \qquad -2\ell(\boldsymbol{\theta}) = \boldsymbol{r}^{\mathsf{T}} \mathbf{C}^{-1} \boldsymbol{r} + \ln\det \mathbf{C} + N \ln 2\pi. \tag{3.92}$$

(1) Write the noise as $\boldsymbol{\epsilon} = L\boldsymbol{\eta}$ with $\mathbf{C} = LL^{\mathsf{T}}$ and $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \mathbb{1})$. Changing variables from $\boldsymbol{\eta}$ to $\boldsymbol{\epsilon}$ multiplies the density by the Jacobian $|\det(\partial\boldsymbol{\eta}/\partial\boldsymbol{\epsilon})| = |\det L^{-1}|$.

(2) Independent standard Gaussians: the density is the product of N factors $(2\pi)^{-1/2} e^{-\eta_i^2/2}$.

(3) $(L^{-1}\boldsymbol{r})^{\mathsf{T}}(L^{-1}\boldsymbol{r}) = \boldsymbol{r}^{\mathsf{T}}(LL^{\mathsf{T}})^{-1}\boldsymbol{r} = \boldsymbol{r}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{r}$, and $(\det L)^2 = \det \mathbf{C}$.

So the correlated χ^2 is $\boldsymbol{r}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{r}$, and every result above, from the normal equations to $\chi^2_{\min} \sim \chi^2_{N-m}$, holds with this $\mathbf{C}$, because those derivations never used that $\mathbf{C}$ was diagonal. This is the likelihood written for cosmological data with correlated errors. One term needs care. The $\ln\det \mathbf{C}$ is a constant only if the covariance does not depend on the parameters. For the power spectrum of a Gaussian random field the covariance *is* the signal (cosmic variance scales with C_ℓ itself), and then $\ln\det \mathbf{C}(\boldsymbol{\theta})$ must be kept. Dropping it biases the fit (Chapter 7).

What goes wrong if we ignore the correlation? Suppose the noise has covariance $\mathbf{C}_{ij} = \sigma_i \sigma_j e^{-|x_i - x_j|/2}$ (neighbours correlated at $e^{-1/2} = 0.61$) but we fit with only the diagonal. The estimator is still $\hat{\boldsymbol{\theta}} = B_{\text{diag}}\boldsymbol{D}$, a linear function of the data with $B_{\text{diag}}A = \mathbb{1}$, so it is still *unbiased*. What is wrong is the error bar. The quoted covariance $(A^{\mathsf{T}}\mathbf{C}_{\text{diag}}^{-1}A)^{-1}$ assumes the noise we did not have; the true covariance is $B_{\text{diag}}\mathbf{C}B_{\text{diag}}^{\mathsf{T}}$ by the same propagation rule as (3.88). The second half of `11_line_fit.py` draws correlated noise as $L\boldsymbol{\eta}$ with the Cholesky factor and fits 10^4 data sets both ways. For the slope:

fit	mean $\hat{b}$	quoted $\sigma_{\hat{b}}$	true scatter of $\hat{b}$
diagonal χ^2	0.4986	0.08767	0.1282 (predicted 0.1284)
full $\boldsymbol{r}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{r}$	0.4994	0.121	0.1212

Both fits centre on the true slope 0.5. The diagonal fit claims an error of 0.08767 but actually scatters by 0.1282, so every quoted interval is too short by a third. The full fit quotes its error correctly, and that error is also smaller than the true error of the diagonal fit: using the correlations is not only honest but more precise. The goodness of fit misleads too. With the diagonal χ^2 the average $\chi^2_{\min}$ is 4.77 instead of 8, a fit that looks suspiciously good; with the full $\mathbf{C}$ it is 8.069 (Figure 3.37).

Watch out: Correlated noise makes a bad error bar look like a good fit

Positively correlated noise moves neighbouring points up and down together. A smooth model can follow such coherent wiggles, so the residuals come out smaller than independent noise of the same size would give, and $\chi^2_{\min} \ll \nu$. At the same moment the parameter errors computed from the diagonal are too small. A reduced χ^2 well below one is a hint to look for correlations, not a reason to celebrate.

The same idea for a gravitational-wave detector. Detector noise is correlated in time but, if stationary, it is uncorrelated between different frequencies: the Fourier transform diagonalises its covariance, and the diagonal entries are the noise power spectral density $S_n(f)$ [the noise variance per unit frequency. Frequencies where the detector is noisy have large S_n]. The sum $\boldsymbol{r}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{r}$ then becomes an integral over frequency, with $\tilde{r}(f) = \int r(t) e^{-2\pi i f t}\,\mathrm{d}t$,

$$-2\ell(\boldsymbol{\theta}) = (D - h_{\boldsymbol{\theta}} \,|\, D - h_{\boldsymbol{\theta}}) + \text{const}, \qquad (a \,|\, b) \equiv 4\,\mathrm{Re}\int_0^\infty \frac{\tilde{a}(f)\,\tilde{b}^*(f)}{S_n(f)}\,\mathrm{d}f,$$

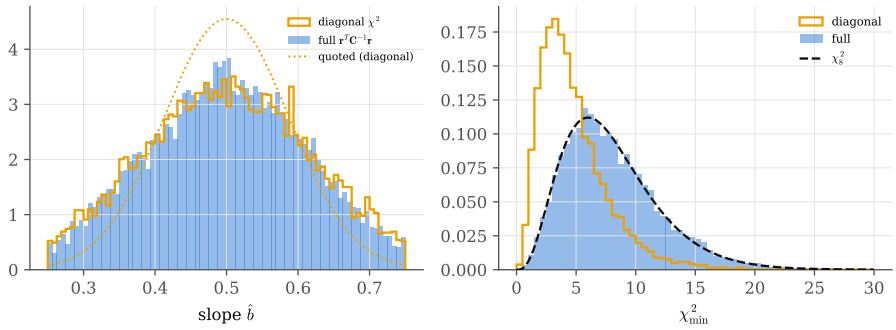


Figure 3.37: Fitting data with correlated noise, 10^4 simulated experiments. Left: the slope from the diagonal χ^2 (outline) and from the full $\boldsymbol{r}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{r}$ (filled). Both are centred on the truth, but the diagonal fit is wider than the Gaussian it quotes (dotted). Right: the diagonal fit's $\chi^2_{\min}$ piles up well below the χ^2_8 curve, while the full fit's follows it.

where $h_{\boldsymbol{\theta}}$ is the waveform predicted for source parameters $\boldsymbol{\theta}$. The **noise-weighted inner product** $(a|b)$ is $\boldsymbol{a}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{b}$ in the basis where $\mathbf{C}$ is diagonal, with one-sided S_n and positive frequencies only, hence the factor 4. Every gravitational-wave parameter estimate, and the Fisher forecasts of Chapter 10, start from this one line.

3.11 Counting experiments: the Poisson likelihood of a histogram

A search at a collider histograms the invariant mass of candidate events into bins; an X-ray telescope histograms photon energies; a decay experiment histograms decay times. Far from the peak, or in a short run, many bins hold a handful of events, and some hold none. The habit of putting an error bar $\sqrt{n_i}$ on every bin and minimising χ^2 breaks down here. A bin with $n_i = 0$ gets an error bar of zero, and bins that fluctuated low get error bars that are too small. The reason is that the noise in a counting experiment is not Gaussian and is not added to the prediction. The count in a bin is a Poisson random number whose mean is the prediction. The likelihood has no trouble with this, because all it needs is the probability of what was observed. Four questions need answers.

1. What is the likelihood of a histogram of counts, and what does maximising it do to the total number of events?
2. Can the same likelihood also tell us whether the fit is good?
3. What exactly goes wrong with the two χ^2 recipes that use $\sqrt{n_i}$ or $\sqrt{\mu_i}$ as the error bar?
4. How does the idea carry over to other data: measurements with Gaussian errors, gravitational-wave strain, CMB maps?

3.11.1 The likelihood of a histogram

Divide the range of x into N bins. If events arrive independently at a constant rate, the count n_i in bin i is Poisson distributed (Chapter 2) with mean

$$\mu_i(\nu, \boldsymbol{\theta}) = \nu \int_{\text{bin } i} f(x;\boldsymbol{\theta})\,\mathrm{d}x \equiv \nu\, p_i(\boldsymbol{\theta}), \qquad \sum_{i=1}^{N} p_i(\boldsymbol{\theta}) = 1,$$

where ν is the expected total and p_i the probability that an event falls in bin i [Cowan §6.10]. Different bins are independent, so

$$\mathcal{L}(\nu, \boldsymbol{\theta}) \overset{(1)}{=} \prod_{i=1}^{N} \frac{\mu_i^{n_i} e^{-\mu_i}}{n_i!},$$
$$\ell(\nu, \boldsymbol{\theta}) \overset{(2)}{=} \sum_{i=1}^{N} \big(n_i \ln \mu_i - \mu_i\big) - \sum_{i=1}^{N} \ln n_i! \overset{(3)}{=} -\tfrac{1}{2}\, C(\nu, \boldsymbol{\theta}) + \text{const}, \qquad C \equiv 2 \sum_{i=1}^{N} \big(\mu_i - n_i \ln \mu_i\big). \tag{3.93}$$

(1) Poisson probability of n_i counts for mean μ_i, multiplied over independent bins [Cowan (6.44)].
(2) $\ln(\mu^n e^{-\mu}/n!) = n \ln \mu - \mu - \ln n!$.
(3) The counts are data, so $\sum \ln n_i!$ does not depend on the parameters.

The quantity $C = -2\ell + \text{const}$ is the **Cash statistic**, introduced for X-ray astronomy [20]; minimising it is maximising the Poisson likelihood. In the language of Section 3.8, each bin asks how probable it is that a Poisson source with mean $\mu_i(\boldsymbol{\theta})$ produced the count we saw, and an empty bin contributes simply $e^{-\mu_i}$, the probability of seeing nothing.

The total number of events is conserved. Split the Poisson product into a statement about the total and a statement about the shape. With $n = \sum_i n_i$ and $\mu_i = \nu p_i$,

$$\prod_{i=1}^{N} \frac{\mu_i^{n_i} e^{-\mu_i}}{n_i!} \overset{(1)}{=} \frac{\nu^n e^{-\nu}}{n!} \, \frac{n!}{\prod_i n_i!} \prod_{i=1}^{N} p_i^{n_i}, \tag{3.94}$$
$$\frac{\partial \ell}{\partial \nu} \overset{(2)}{=} \frac{n}{\nu} - 1 \overset{(3)}{=} 0 \quad \Longrightarrow \quad \hat{\nu} = n.$$

(1) $\prod_i \mu_i^{n_i} = \nu^{\sum n_i} \prod_i p_i^{n_i}$ and $\prod_i e^{-\mu_i} = e^{-\nu \sum p_i} = e^{-\nu}$; then multiply and divide by $n!$. The first factor is the Poisson probability of the total, the second the multinomial probability of the split into bins [Cowan (6.41), (6.43)].
(2) Only the first factor depends on ν, and $\partial_\nu(n \ln \nu - \nu) = n/\nu - 1$.
(3) The score equation for ν; the multinomial factor is untouched, so $\hat{\boldsymbol{\theta}}$ is the same as in a fit with the total fixed.

The fitted normalisation equals the observed number of events, exactly, in every experiment, whatever the shape parameters do. This is the binned form of the extended likelihood (3.85). It is also the property the two χ^2 recipes below lack.

Working definition 3.12 (Poisson likelihood of a histogram)

For counts n_i with predicted means $\mu_i(\boldsymbol{\theta})$,

$$-2\ell(\boldsymbol{\theta}) = 2 \sum_i \big(\mu_i - n_i \ln \mu_i\big) + \text{const}.$$

Operationally: compute the expected count in every bin from the model (integrate the density over the bin and multiply by the expected total), evaluate this sum, minimise. Empty bins stay in the sum and contribute $2\mu_i$. No bins need to be merged, and no error bars are assigned.

Example: one bin. For a single bin with $n = 9$ counts, $\ell(\mu) = 9 \ln \mu - \mu$ gives $\hat{\mu} = 9$ and $-\ell''(\hat{\mu}) = n/\hat{\mu}^2 = 1/9$, a curvature error of $3 = \sqrt{n}$, the familiar rule. The likelihood itself is

skewed: solving $\ell(\mu) = \ell_{\max} - \frac{1}{2}$ gives $\mu \in [6.32, 12.34]$, that is $9^{+3.34}_{-2.68}$. The interval extends further above the estimate than below it because the spread $\sqrt{\mu}$ of a Poisson count grows with its mean μ. A large mean such as 12.34 has a large spread, 3.5, so a count of 9 is less than one spread below it. A small mean such as 6.32 has a small spread, 2.5, so the same count is slightly more than one spread above it. Large means can therefore stray further from the data and still explain them about as well (Section 3.12).

3.11.2 Fitting and testing with one number: the Baker–Cousins χ^2_λ

The value of C at its minimum does not by itself say whether the fit is good. Its constant is arbitrary, and it grows with the total number of events even for a perfect model. To make it a measure of fit, compare the model with the best any model could do [21]. That best is the **saturated model** [a "model" with one free mean per bin, which therefore sets $\mu_i = n_i$ and fits the data perfectly]. The likelihood ratio is

$$\chi^2_\lambda(\boldsymbol{\theta}) \overset{(1)}{=} -2\ln\frac{\mathcal{L}(\boldsymbol{\mu}(\boldsymbol{\theta}))}{\mathcal{L}(\boldsymbol{\mu}=\boldsymbol{n})} \overset{(2)}{=} -2\sum_i\Big[\big(n_i\ln\mu_i-\mu_i\big)-\big(n_i\ln n_i-n_i\big)\Big]$$
$$\overset{(3)}{=} 2\sum_{i=1}^{N}\Big[\mu_i-n_i+n_i\ln\frac{n_i}{\mu_i}\Big] \tag{3.95}$$

[Cowan (6.52)].

(1) Definition of the likelihood-ratio statistic against the saturated model.
(2) (3.93) at μ_i and at $\mu_i = n_i$; the $\ln n_i!$ cancel in the ratio.
(3) Collect terms; $n_i\ln n_i - n_i\ln\mu_i = n_i\ln(n_i/\mu_i)$. An empty bin contributes $2\mu_i$ (since $n\ln n \to 0$ as $n \to 0$).

Since $\chi^2_\lambda = C +$(terms in n_i only), minimising it is the same as maximising the Poisson likelihood. Every term is non-negative. To see it, write the bracket as $n_i(u - 1 - \ln u)$ with $u = \mu_i/n_i$, and note that $u - 1 - \ln u \geq 0$, with equality only at $u = 1$ (the line $u - 1$ lies above the curve $\ln u$ and touches it at $u = 1$). So χ^2_λ vanishes only for a perfect fit, $\mu_i = n_i$ in every bin.

Why call it a χ^2? Expand one term for a count close to its mean, $n = \mu + \delta$ with $|\delta| \ll \mu$:

$$\mu - n + n\ln\frac{n}{\mu} \overset{(1)}{=} -\delta + (\mu+\delta)\ln\Big(1+\frac{\delta}{\mu}\Big) \overset{(2)}{=} -\delta + (\mu+\delta)\Big(\frac{\delta}{\mu}-\frac{\delta^2}{2\mu^2}+O\big(\tfrac{\delta^3}{\mu^3}\big)\Big)$$
$$\overset{(3)}{=} -\delta+\delta+\frac{\delta^2}{\mu}-\frac{\delta^2}{2\mu}+O\Big(\frac{\delta^3}{\mu^2}\Big) = \frac{\delta^2}{2\mu}+O\Big(\frac{\delta^3}{\mu^2}\Big).$$

(1) Substitute $n = \mu + \delta$, and $n/\mu = 1 + \delta/\mu$.
(2) Taylor series $\ln(1 + u) = u - u^2/2 + O(u^3)$.
(3) Multiply out and keep terms up to δ^2/μ.

So $\chi^2_\lambda \approx \sum_i (n_i - \mu_i)^2/\mu_i$ when every bin has many counts: a Pearson χ^2, a sum of squared standardised Poisson fluctuations. In that limit $\chi^2_{\lambda,\min} \sim \chi^2_{N-m}$ for m fitted parameters, the Poisson analogue of (3.91). The general statement is **Wilks' theorem** [22]: for nested models, $-2\ln$ of the ratio of maximised likelihoods is asymptotically χ^2 with as many degrees of freedom as the difference in the number of free parameters, here $N - m$ (the saturated model has N). Chapter 4 treats it properly. Baker and Cousins' point is that χ^2_λ reaches this limit faster than the χ^2 recipes, and one statistic serves both to fit and to test.

Key idea: Binned counts: use the Poisson likelihood

Minimise $\chi^2_\lambda = 2\sum_i[\mu_i - n_i + n_i \ln(n_i/\mu_i)]$. It is $-2\ln\mathcal{L}$ up to a constant, so it gives the maximum likelihood fit; the fitted total equals the observed total exactly; empty bins are harmless; and its minimum is approximately χ^2_{N-m}, so it is also the goodness-of-fit statistic.

3.11.3 What goes wrong with $\sqrt{n}$ error bars

Two least-squares recipes are common for histograms [Cowan §7.4]. The **Neyman** χ^2 takes the variance of each bin from the data, $\chi^2_N = \sum_i (n_i - \mu_i)^2/n_i$. The **Pearson** χ^2 takes it from the model, $\chi^2_P = \sum_i (n_i - \mu_i)^2/\mu_i$. Both treat the counts as Gaussian, and both get the normalisation wrong, in opposite directions. Let the model be $\mu_i = \nu p_i$ with $\sum_i p_i = 1$ and minimise over ν.

Neyman. Setting the derivative to zero,

$$\frac{\partial \chi^2_N}{\partial \nu} \overset{(1)}{=} -2\sum_i \frac{p_i(n_i - \nu p_i)}{n_i} = 0 \quad\Longrightarrow\quad 1 = \nu \sum_i \frac{p_i^2}{n_i},$$

$$\chi^2_{N,\min} \overset{(2)}{=} \sum_i n_i - 2\nu \sum_i p_i + \nu^2 \sum_i \frac{p_i^2}{n_i} \overset{(3)}{=} n - 2\nu + \nu \quad\Longrightarrow\quad \hat{\nu}_N = n - \chi^2_{N,\min}.$$

(1) Chain rule on $(n_i - \nu p_i)^2$, then $\sum_i p_i n_i/n_i = \sum_i p_i = 1$.
(2) Expand the square: $(n_i - \nu p_i)^2/n_i = n_i - 2\nu p_i + \nu^2 p_i^2/n_i$.
(3) $\sum_i n_i = n$, $\sum_i p_i = 1$, and the first line gives $\nu^2 \sum p_i^2/n_i = \nu$.

Pearson. In the same way,

$$\frac{\partial \chi^2_P}{\partial \nu} \overset{(1)}{=} -\sum_i \frac{n_i^2}{\nu^2 p_i} + 1 = 0 \quad\Longrightarrow\quad \nu^2 = \sum_i \frac{n_i^2}{p_i},$$

$$\chi^2_{P,\min} \overset{(2)}{=} \sum_i \frac{n_i^2}{\nu p_i} - 2n + \nu \overset{(3)}{=} \nu - 2n + \nu \quad\Longrightarrow\quad \hat{\nu}_P = n + \tfrac{1}{2}\chi^2_{P,\min}.$$

(1) Expand first: $(n_i - \nu p_i)^2/(\nu p_i) = n_i^2/(\nu p_i) - 2n_i + \nu p_i$, then differentiate in ν.
(2) Sum the expanded terms.
(3) The first line gives $\sum n_i^2/(\nu p_i) = \nu$.

These are Cowan's (7.22) and (7.21). They say that the Neyman fit undershoots the observed total by $\chi^2_{N,\min}$ and the Pearson fit overshoots it by $\frac{1}{2}\chi^2_{P,\min}$. Both relations hold exactly at the minimum, also when shape parameters are fitted at the same time, because the derivative with respect to ν must vanish there too. A correct model gives $\chi^2_{\min} \approx N - m$, so the Neyman fit is low by about the number of bins and the Pearson fit high by about half of it. The definitions show why. Neyman gives a bin that fluctuated down a small variance, hence a large weight, and the fit is dragged down. Pearson can lower its χ^2 by raising μ_i, which inflates the variances it divides by, and the fit drifts up. With $N = 20$ bins and $n \approx 120$ events the Neyman bias is about ten per cent, comparable to the statistical error $\sqrt{n} \approx 11$ on the total itself: a systematic error as large as the statistical one, introduced by the fitting recipe alone.

Example: two bins by hand. Suppose the model says the two halves of a range are equally likely, $p_1 = p_2 = \tfrac{1}{2}$, and we observe $n_1 = 4$, $n_2 = 16$, so $n = 20$. Then

$$\hat\nu_{\mathrm{ML}} = 20,$$
$$\hat\nu_N \overset{(1)}{=} \left(\frac{1/4}{4} + \frac{1/4}{16}\right)^{-1} = 12.8, \qquad \chi^2_{N,\min} \overset{(2)}{=} 20 - 12.8 = 7.2,$$
$$\hat\nu_P \overset{(3)}{=} \sqrt{\frac{16}{1/2} + \frac{256}{1/2}} = \sqrt{544} = 23.3, \qquad \chi^2_{P,\min} \overset{(4)}{=} 2(23.3 - 20) = 6.65.$$

(1) The Neyman condition $1 = \nu \sum p_i^2/n_i$.
(2) $\hat\nu_N = n - \chi^2_{N,\min}$.
(3) The Pearson condition $\nu^2 = \sum n_i^2/p_i$.
(4) $\hat\nu_P = n + \tfrac{1}{2}\chi^2_{P,\min}$.

Twenty events were seen. Only the likelihood says twenty.

3.11.4 A histogram fit in Python

The script `12_poisson_bins.py` histograms decay times into $N = 20$ bins on $0 \le t \le 5$ (in units of the true lifetime), with an expected total of $\nu = 120$, and fits (ν, τ) with the three statistics.

Listing 3.10: The three statistics for a histogram (`code/ch03/12_poisson_bins.py`)

```
def cash(par, n):          # -2 ln L up to a constant: 2 sum (mu - n ln mu)
    m = mu(par)
    return 2 * np.sum(m - n * np.log(m))


def baker_cousins(par, n):  # 2 sum (mu - n + n ln(n/mu)), with n ln n -> 0 for n = 0
    m = mu(par)
    return 2 * np.sum(m - n + special.xlogy(n, n) - special.xlogy(n, m))


def neyman(par, n):
    m = mu(par)
    return np.sum((n - m) ** 2 / np.maximum(n, 1))


def pearson(par, n):
    m = mu(par)
    return np.sum((n - m) ** 2 / m)
```

The Cash and Baker–Cousins functions differ only by terms in n_i, so they have the same minimum; `special.xlogy(n, n)` evaluates $n \ln n$ with the convention $0 \ln 0 = 0$. The Neyman function needs a patch for empty bins, `np.maximum(n, 1)`, which is itself a sign of trouble.

For the example histogram, with $n = 122$ events and 4 empty bins, the fitted totals are $\hat\nu = 122$ (Poisson), 110.7 (Neyman) and 129 (Pearson), with lifetimes 1.076, 0.9253 and 1.18 (Figure 3.38). The Baker–Cousins statistic at the minimum is 18.47 for 18 degrees of freedom, a normal-looking fit.

Over 3000 simulated histograms the Poisson fit returns $|\hat\nu - n| \le 3.996 \times 10^{-9}$ in every one of them, the exact conservation law of (3.94) up to the tolerance of the minimiser. The Pearson fit is high by 7.883 events on average, exactly $\tfrac{1}{2}\overline{\chi^2_{P,\min}} = 7.883$. The Neyman fit is low by -11.65; the relation $\hat\nu_N = n - \chi^2_{N,\min}$ predicts -13.73, and the difference is entirely due to the patch for empty bins, which breaks the identity $\sum_i p_i n_i/n_i = 1$ used in its derivation. The bias leaks into the shape: the mean fitted lifetimes are 1.002, 0.8992 and 1.114 for a true value of 1. Finally, the

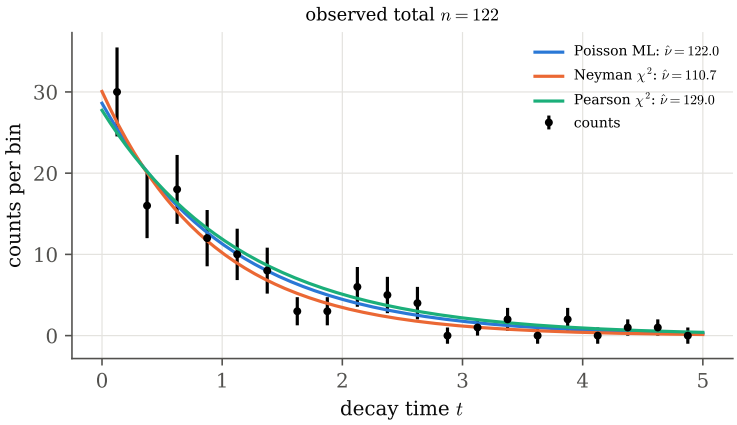


Figure 3.38: One histogram of decay times with few counts per bin (points, with $\sqrt{\max(n_i, 1)}$ bars for orientation only) and three fits of the same exponential shape. The Poisson likelihood fit reproduces the observed total. The Neyman χ^2 fit sits low, pulled down by the bins that fluctuated low; the Pearson χ^2 fit sits high.

Baker–Cousins $\chi^2_{\lambda,\min}$ averages 18.73 against 18 degrees of freedom and follows the χ^2_{18} curve even with this sparse histogram (Figure 3.39).

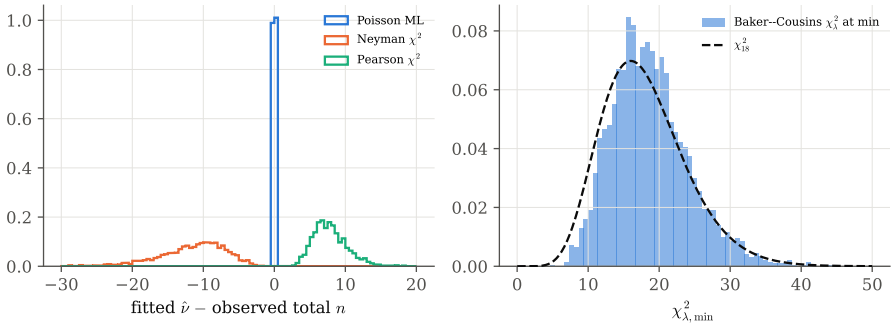


Figure 3.39: Three thousand simulated histograms. Left: fitted total minus observed total. The Poisson fit is a spike at zero; the Neyman fit is biased low by about ten events and the Pearson fit high by about eight. Right: the Baker–Cousins statistic at its minimum follows the χ^2 distribution for $20 - 2 = 18$ degrees of freedom, so it can be used as a goodness-of-fit test.

3.11.5 One idea, three implementations

The recipe is always the same: write the prediction, write the probability of the fluctuation needed to turn it into the data, and multiply over independent pieces. Only the noise model changes from one field to another. Three noise models cover most data in physics and cosmology: Poisson counts; Gaussian noise with a covariance that does not depend on the parameters (gravitational-wave strain is this case written in frequency); and a Gaussian field whose covariance is itself the signal, as for the CMB.

The last line needs a derivation, because it is the first case where the parameters sit in the covariance rather than in the mean. In a statistically isotropic Gaussian sky the spherical-harmonic coefficients $a_{\ell m}$ have mean zero and variance C_ℓ, the theory's power spectrum. The measured spectrum is $\widehat{C}_\ell = \frac{1}{2\ell+1}\sum_m |a_{\ell m}|^2$. Because the sky is real, $a_{\ell,-m}$ is fixed by $a_{\ell m}$, so at each ℓ there are $2\ell + 1$ independent real numbers: $a_{\ell 0}$, and $\sqrt{2}\,\mathrm{Re}\,a_{\ell m}$ and $\sqrt{2}\,\mathrm{Im}\,a_{\ell m}$ for $m = 1, \ldots, \ell$. Call them $y_{\ell k}$, $k = 1, \ldots, 2\ell + 1$. Each is Gaussian with variance C_ℓ, and

Taxonomy: Which likelihood when

data	noise model	$-2\ln\mathcal{L}$ (up to constants)
counts in bins (colliders, X-ray, decay histograms)	$n_i \sim \text{Poisson}(\mu_i(\boldsymbol{\theta}))$, independent bins	$2\sum_i[\mu_i - n_i \ln \mu_i]$ (3.93)
measurements with error bars; binned spectra	Gaussian, covariance $\mathbf{C}$ independent of $\boldsymbol{\theta}$	$\boldsymbol{r}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{r}$ (3.92)
GW strain	stationary Gaussian noise with PSD $S_n(f)$	$(D - h_{\boldsymbol{\theta}}\|D - h_{\boldsymbol{\theta}})$ (Section 3.10.5)
CMB map, full sky, no noise	$a_{\ell m}$ Gaussian with variance $C_\ell(\boldsymbol{\theta})$	$\sum_\ell (2\ell+1)\big[\widehat{C}_\ell/C_\ell + \ln C_\ell\big]$

$\sum_k y_{\ell k}^2 = \sum_m |a_{\ell m}|^2 = (2\ell+1)\widehat{C}_\ell$. Then

$$-2\ell(\boldsymbol{\theta}) \overset{(1)}{=} \sum_\ell \sum_{k=1}^{2\ell+1}\Big[\frac{y_{\ell k}^2}{C_\ell(\boldsymbol{\theta})} + \ln C_\ell(\boldsymbol{\theta}) + \ln 2\pi\Big] \overset{(2)}{=} \sum_\ell (2\ell+1)\Big[\frac{\widehat{C}_\ell}{C_\ell(\boldsymbol{\theta})} + \ln C_\ell(\boldsymbol{\theta})\Big] + \text{const.}$$

(1) (3.92) with a diagonal covariance whose entries are C_ℓ: $-2\ln$ of each Gaussian factor is $y^2/C_\ell + \ln C_\ell + \ln 2\pi$.

(2) Sum over k using $\sum_k y_{\ell k}^2 = (2\ell+1)\widehat{C}_\ell$; the $2\ell+1$ copies of $\ln C_\ell$ give the second term.

The $\ln C_\ell$ term is the $\ln\det\mathbf{C}$ of (3.92), and it cannot be dropped: without it the fit would push every C_ℓ to infinity. If the model sets each C_ℓ free, the score equation $-\widehat{C}_\ell/C_\ell^2 + 1/C_\ell = 0$ gives $\hat{C}_\ell = \widehat{C}_\ell$, so the familiar estimator is the maximum likelihood estimator. As a function of C_ℓ this likelihood is skewed, like the Poisson one above, and for the same reason: the spread of $\widehat{C}_\ell$ grows with its mean. A mask, instrument noise and a beam change this likelihood, as Chapters 7 and 8 show.

3.12 The width of the likelihood is the error bar

The width of the likelihood is the error bar of the estimate. To see why one would expect this, look at a likelihood with a sharp peak. Only a narrow range of parameter values can turn the prediction into the data with an ordinary fluctuation. Move a little away and the noise required becomes implausible. A broad likelihood says the opposite: many values would have done about equally well. For linear models with Gaussian noise we proved the claim exactly in Section 3.10.2: the log-likelihood is a parabola whose width is the standard deviation of $\hat{\theta}$. For any smooth likelihood with enough data it holds approximately, and a check on 10^4 simulated experiments shows how well. Three questions follow.

1. How do we read an error bar off the log-likelihood, from its curvature or from where it has dropped by $\frac{1}{2}$?
2. Why should the width of a function of θ, computed from one data set, equal the scatter of $\hat{\theta}$ over many data sets? The answer introduces the Fisher information.
3. How well does it work at a realistic sample size?

3.12.1 Curvature at the peak

Expand the log-likelihood about its maximum, where the first derivative vanishes [Padilla (29)–(31)]:

$$\ell(\theta) \overset{(1)}{=} \ell(\hat\theta) + \underbrace{\ell'(\hat\theta)}_{=0}(\theta-\hat\theta) + \tfrac{1}{2}\ell''(\hat\theta)(\theta-\hat\theta)^2 + \ldots \overset{(2)}{=} \ell_{\max} - \frac{(\theta-\hat\theta)^2}{2\hat\sigma^2} + \ldots, \qquad \hat\sigma^2 \equiv \frac{1}{-\ell''(\hat\theta)}. \tag{3.96}$$

(1) Taylor's theorem at the maximum; the score vanishes there.
(2) $\ell''(\hat\theta) < 0$ at a maximum, so $-1/\ell''$ is positive and we may call it $\hat\sigma^2$.

Near its peak the likelihood is therefore $\mathcal{L}(\theta) \approx \mathcal{L}_{\max}\exp[-(\theta-\hat\theta)^2/2\hat\sigma^2]$, a Gaussian bump of width $\hat\sigma$. The rule is: *the error bar is one over the square root of minus the second derivative of* $\ln\mathcal{L}$ *at its maximum* [Cowan §6.4]. For several parameters, minus the matrix of second derivatives (the Hessian with a minus sign) plays the role of $1/\hat\sigma^2$, and its inverse is the estimated covariance matrix. That is the line `V = np.linalg.inv(-H)` in the Newton–Raphson fit of the muon angular distribution (Section 3.9.3).

Example: the lifetime. For n decay times the maximum likelihood lifetime is $\hat\tau = \bar t$, and (3.82) gave $-\ell''(\hat\tau) = n/\hat\tau^2$. So $\hat\sigma_{\hat\tau} = \hat\tau/\sqrt{n}$. Fifty decays measure a lifetime to $1/\sqrt{50} \approx 14\%$. The true standard deviation of $\bar t$ is $\tau/\sqrt{n}$ (Section 3.9.1), and the curvature gives the same formula with the unknown τ replaced by its estimate.

Example: the detector efficiency. For the detector that fires S times in n trials, the log-likelihood is $\ell(p) = S\ln p + (n-S)\ln(1-p)$ and $\hat p = S/n$ (Section 3.9.1). Then

$$-\ell''(p) \overset{(1)}{=} \frac{S}{p^2} + \frac{n-S}{(1-p)^2}, \qquad -\ell''(\hat p) \overset{(2)}{=} \frac{n}{\hat p} + \frac{n}{1-\hat p} \overset{(3)}{=} \frac{n}{\hat p(1-\hat p)}, \qquad \hat\sigma_{\hat p} = \sqrt{\frac{\hat p(1-\hat p)}{n}}.$$

(1) Differentiate $S/p - (n-S)/(1-p)$ once more.
(2) At $\hat p = S/n$: $S/\hat p^2 = n/\hat p$ and $(n-S)/(1-\hat p)^2 = n/(1-\hat p)$.
(3) Common denominator: $1/\hat p + 1/(1-\hat p) = 1/[\hat p(1-\hat p)]$.

For $S = 12$ out of $n = 20$ this is $\sqrt{0.6\cdot 0.4/20} = 0.11$, the binomial standard error.

3.12.2 The graphical method: where $\ln\mathcal{L}$ has dropped by one half

If ℓ were exactly the parabola (3.96), then at $\theta = \hat\theta \pm \hat\sigma$ it would be exactly $\ell_{\max} - \tfrac12$. That suggests a second recipe that does not need derivatives: draw $\ell(\theta)$ and read off where it crosses $\ell_{\max} - \tfrac12$ [Cowan §6.7]. For Gaussian data $-2\ell = \chi^2 + \text{const}$, so this is the $\Delta\chi^2 = 1$ rule of Section 3.10.3. When the log-likelihood is not a parabola, the two crossings are at different distances from $\hat\theta$, and one quotes an asymmetric error $\hat\theta^{+\sigma_+}_{-\sigma_-}$.

Example: the lifetime interval exactly. Measure τ in units of the estimate, $u = \tau/\hat\tau$. Using $T = n\hat\tau$,

$$\ell(\tau) - \ell(\hat\tau) \overset{(1)}{=} -n\ln\tau - \frac{n\hat\tau}{\tau} + n\ln\hat\tau + n \overset{(2)}{=} -n\left[\ln u + \frac{1}{u} - 1\right].$$

(1) $\ell(\tau) = -n\ln\tau - T/\tau$ at τ and at $\hat\tau$, with $T/\hat\tau = n$.

(2) $\ln\tau - \ln\hat{\tau} = \ln u$ and $\hat{\tau}/\tau = 1/u$.

The $\Delta\ln\mathcal{L} = \frac{1}{2}$ interval is the pair of roots of $\ln u + 1/u - 1 = 1/(2n)$, the same for every data set once expressed in units of $\hat{\tau}$. For small $u - 1$, $\ln u + 1/u - 1 \approx \frac{1}{2}(u-1)^2$, which returns $u = 1 \pm 1/\sqrt{n}$, the curvature error. For $n = 50$ the exact roots are $u = 0.871$ and $u = 1.156$: the interval reaches further up than down, because decay times drawn with a longer lifetime also scatter more.

Why the graphical interval is the better one. The curvature error depends on how we name the parameter; the $\Delta\ln\mathcal{L}$ interval does not. The likelihood assigns the same number to a hypothesis whatever we call it (Section 3.8.3). So if we switch from τ to the rate $\lambda = 1/\tau$, the set of hypotheses with $\ell \geq \ell_{\max} - \frac{1}{2}$ is the same set, merely relabelled. If τ_- and τ_+ are the lower and upper ends of the τ interval, the λ interval is $[1/\tau_+, 1/\tau_-]$. The curvature, by contrast, changes: $-\partial^2\ell/\partial\lambda^2$ at $\hat{\lambda}$ is $n/\hat{\lambda}^2$, so the curvature interval in λ is $\hat{\lambda}(1 \pm 1/\sqrt{n})$, which is *not* the image of $\hat{\tau}(1 \pm 1/\sqrt{n})$. The two symmetric recipes disagree with each other, and each is right only to leading order. The graphical interval behaves as if we had found and used the parametrisation in which ℓ is most nearly a parabola, without our having to find it.

Watch out: A curvature error assumes a parabola

$\hat{\sigma} = (-\ell'')^{-1/2}$ describes the peak only. If ℓ is skewed, has a second maximum, or is cut off by a physical boundary (a mass that cannot be negative, a cliff like the uniform box), the curvature says nothing about the tails. Always plot $\ln\mathcal{L}$, or at least find the two $\Delta\ln\mathcal{L} = \frac{1}{2}$ crossings, before quoting $\pm\hat{\sigma}$.

Several parameters. With two parameters the level set $\ell = \ell_{\max} - \frac{1}{2}$ is a closed curve, an ellipse when ℓ is quadratic. In the right panel of Figure 3.34 it is drawn for the fit of the angular distribution $1 + \alpha x + \beta x^2$ (Section 3.9.3). The dotted lines tangent to it are at $\hat{\alpha} \pm \sigma_{\hat{\alpha}}$ and $\hat{\beta} \pm \sigma_{\hat{\beta}}$, with the errors from the inverse Hessian [Cowan §6.8]. Why tangent lines? Fix α and move β to make ℓ as large as possible. As α moves away from $\hat{\alpha}$ this best value falls, and it reaches $\ell_{\max} - \frac{1}{2}$ exactly where the vertical line at that α touches the contour. That best value as a function of α is the profile likelihood of Section 3.14. The tilt of the ellipse shows the correlation between the two estimates. For that fit the inverse Hessian gives $\sigma_{\hat{\alpha}} = 0.05099$, $\sigma_{\hat{\beta}} = 0.107$ and $\rho = 0.4237$; the 500 repetitions scatter with 0.05376, 0.1081 and $\rho = 0.4919$. The standard deviations agree to a few per cent. The correlation from 500 pairs is itself uncertain by $(1-\rho^2)/\sqrt{500} = 0.0339$, so the two values of ρ should be compared within that error. The Hessian's own ρ also changes from one experiment to the next: over the 500 repetitions it averages 0.4263 and scatters by 0.03868. Its average and the repetitions' ρ are consistent within about two of the repetitions' errors: 500 repetitions are not enough to tell them apart. The $\frac{1}{2}$-contour projects onto each axis as a 68% interval for that parameter alone. A region that should contain *both* true values with 68% probability is the larger contour $\Delta\ln\mathcal{L} = 2.30/2 = 1.15$, as in the $\Delta\chi^2$ table of Section 3.10.3.

3.12.3 Why the curvature is the error: score and Fisher information

We now have a recipe that uses one data set, a second derivative of a function of θ. The error bar we want is a statement about many data sets: the standard deviation of $\hat{\theta}$ over imagined repetitions. Why should they agree? The bridge is the score, $S(\theta) = \ell'(\theta)$, viewed as a random variable. For a fixed θ, each data set gives a different score. Two identities about its distribution do all the work. They hold whenever we may differentiate under the integral sign, which requires, in particular, that the range of the data does not depend on θ (it fails for the uniform

box).

The score has mean zero. For a single observation with density $f(x;\theta)$,

$$\mathbb{E}_\theta\Big[\frac{\partial \ln f(X;\theta)}{\partial\theta}\Big] \overset{(1)}{=} \int \frac{\partial_\theta f(x;\theta)}{f(x;\theta)}\, f(x;\theta)\,\mathrm{d}x \overset{(2)}{=} \frac{\partial}{\partial\theta}\int f(x;\theta)\,\mathrm{d}x \overset{(3)}{=} \frac{\partial}{\partial\theta}1 = 0. \tag{3.97}$$

(1) Expectation over the density of the data at the *same* θ; $\partial_\theta \ln f = \partial_\theta f / f$.
(2) The f cancels. Exchange the derivative and the integral (the regularity condition).
(3) Every density integrates to one, for every θ.

At the true parameter, the data pull neither up nor down on average.

The variance of the score is minus the mean curvature. Now differentiate the identity

$$\int (\partial_\theta \ln f)\, f\,\mathrm{d}x = 0$$

once more with respect to θ:

$$0 \overset{(1)}{=} \int \Big[(\partial_\theta^2 \ln f)\, f + (\partial_\theta \ln f)\,\partial_\theta f\Big]\mathrm{d}x \overset{(2)}{=} \int \Big[\partial_\theta^2 \ln f + (\partial_\theta \ln f)^2\Big] f\,\mathrm{d}x,$$
$$I(\theta) \equiv \mathbb{E}_\theta\big[(\partial_\theta \ln f)^2\big] \overset{(3)}{=} -\,\mathbb{E}_\theta\big[\partial_\theta^2 \ln f\big]. \tag{3.98}$$

(1) Product rule under the integral.
(2) $\partial_\theta f = (\partial_\theta \ln f)\, f$.
(3) Split the integral. Since the score has mean zero, the first form is its variance.

For n independent observations, $\ell = \sum_i \ln f(x_i;\theta)$ is a sum of independent terms, so the variance of the total score and the expected curvature both add: $I_n(\theta) = nI(\theta)$ [Wasserman p. 128, Thm. 9.17].

Working definition 3.13 (Fisher information)

The **Fisher information** in n observations is

$$I_n(\theta) = \mathrm{Var}_\theta\big[S(\theta)\big] = -\,\mathbb{E}_\theta\big[\ell''(\theta)\big] = nI(\theta),$$

the average curvature of the log-likelihood at the true parameter [Wasserman p. 128, Def. 9.16]. The **observed information** $-\ell''(\hat\theta)$ is the curvature of the one log-likelihood we have, at its peak. *Operationally:* differentiate $\ln f$ twice, take minus its average over the model, multiply by n. For several parameters, $I_{jk} = -\mathbb{E}[\partial_j \partial_k \ell]$ is the **Fisher matrix** (Chapter 10).

From the score to the scatter of $\hat\theta$. Now expand the score equation of one data set about the true value θ [Wasserman p. 136]:

$$0 \overset{(1)}{=} \ell'(\hat\theta) \overset{(2)}{=} \ell'(\theta) + (\hat\theta - \theta)\,\ell''(\theta) + \ldots \quad\Longrightarrow\quad \hat\theta - \theta \approx \frac{\ell'(\theta)}{-\ell''(\theta)} \overset{(3)}{=} \frac{\frac{1}{\sqrt n}\,\ell'(\theta)}{\frac{1}{\sqrt n}\,[-\ell''(\theta)]}$$
$$\sqrt{n}\,(\hat\theta - \theta) \approx \frac{\frac{1}{\sqrt n}\sum_i S_i}{\frac{1}{n}\sum_i [-\partial_\theta^2 \ln f(x_i;\theta)]} \overset{(4)}{\rightsquigarrow} \frac{\mathcal{N}\big(0, I(\theta)\big)}{I(\theta)} \overset{(5)}{=} \mathcal{N}\Big(0, \frac{1}{I(\theta)}\Big). \tag{3.99}$$

(1) $\hat\theta$ solves the score equation.

(2) Taylor expansion of ℓ' about the true value; for large n, $\hat{\theta}$ is close to θ (consistency, Section 3.13) and the dropped terms are smaller.
(3) Multiply numerator and denominator by $1/\sqrt{n}$ and move one $\sqrt{n}$ to the left.
(4) The numerator is $\sqrt{n}$ times the average of the n independent scores $S_i = \partial_\theta \ln f(x_i;\theta)$, each of mean 0 (3.97) and variance $I(\theta)$ (3.98), so by the central limit theorem it tends to $\mathcal{N}(0, I)$. The denominator is an average of n independent curvatures with mean $I(\theta)$, so by the law of large numbers it tends to the constant $I(\theta)$.
(5) Dividing a $\mathcal{N}(0, I)$ variable by the number I gives variance $I/I^2 = 1/I$.

So the scatter of $\hat{\theta}$ over repetitions is set by the Fisher information, the *average* curvature of the log-likelihood. The curvature $-\ell''(\hat{\theta})$ of the one data set we have is an average of n independent curvatures, so by the same law of large numbers it is close to I_n. That is why the width read off our one likelihood gives the error bar.

Key idea: The curvature of $\ln \mathcal{L}$ is the error bar

For a regular model and large n the MLE is approximately Gaussian,

$$\hat{\theta} \approx \mathcal{N}\Big(\theta,\ \frac{1}{I_n(\theta)}\Big), \qquad \widehat{\mathrm{se}} = \frac{1}{\sqrt{I_n(\hat{\theta})}} \approx \frac{1}{\sqrt{-\ell''(\hat{\theta})}},$$

and $\hat{\theta} \pm 1.96\,\widehat{\mathrm{se}}$ is an approximate 95% confidence interval [Wasserman pp. 129–130, Thms. 9.18–9.19]. With several parameters, $\mathrm{Cov}\,\hat{\boldsymbol{\theta}} \approx I_n^{-1}$.

Examples of Fisher information. The standard cases follow from (3.98) in one line each [Wasserman p. 130, Exs. 9.20–9.22].

- Gaussian mean, known σ: $\ln f = -(x-\mu)^2/2\sigma^2 + \ldots$, $\partial_\mu^2 \ln f = -1/\sigma^2$, so $I = 1/\sigma^2$ and $\mathrm{se} = \sigma/\sqrt{n}$. The curvature is the same for every data set, so observed and expected information coincide exactly.
- Poisson counts, mean λ: $\ln f = x \ln\lambda - \lambda - \ln x!$, $\partial_\lambda^2 \ln f = -x/\lambda^2$, and $\mathbb{E}\,X = \lambda$ gives $I = 1/\lambda$, $\mathrm{se} = \sqrt{\lambda/n}$.
- Bernoulli: $\partial_p^2 \ln f = -x/p^2 - (1-x)/(1-p)^2$, with $\mathbb{E}\,X = p$, so $I = 1/p + 1/(1-p) = 1/[p(1-p)]$.
- Exponential lifetime: $\partial_\tau^2(-\ln\tau - t/\tau) = 1/\tau^2 - 2t/\tau^3$, with $\mathbb{E}\,t = \tau$, so $I = 1/\tau^2$ and $\mathrm{se} = \tau/\sqrt{n}$.

The information per observation is large where the density changes quickly as θ changes, because then each observation discriminates sharply between nearby values: a Gaussian of small σ, a Poisson mean near zero, a coin whose p is near 0 or 1.

3.12.4 Testing the recipe on 10^4 experiments

The script `09_mle_decay.py` tests the whole argument on the lifetime, for $n = 50$ decays with true lifetime $\tau = 1$. It first analyses one experiment, the way an experimenter would.

Listing 3.11: One experiment: the MLE, the curvature error and the $\Delta \ln \mathcal{L} = \frac{1}{2}$ interval (`code/ch03/09_mle_decay.py`)

```
# ---------- one experiment ----------
```

```
t = rng.exponential(tau_true, n)
tau_hat = t.mean()                                          # analytic MLE
tau_num = optimize.minimize_scalar(lambda s: -lnL(s, t), bounds=(0.1, 5),
    method="bounded").x
sig_curv = tau_hat / np.sqrt(n)                             # 1/sqrt(-d2 lnL/dtau2) at tau_hat
lmax = lnL(tau_hat, t)
f = lambda s: lnL(s, t) - lmax + 0.5                        # zero where ln L has dropped by 1/2
lo = optimize.brentq(f, 0.3 * tau_hat, tau_hat)
hi = optimize.brentq(f, tau_hat, 3 * tau_hat)
```

The analytic MLE is the mean; `minimize_scalar` confirms it numerically (0.8831). The curvature error is $\hat{\tau}/\sqrt{n}$. The function `f` is zero where $\ln \mathcal{L}$ has dropped by $\frac{1}{2}$, and `brentq` finds its two roots on either side of the peak. The result is

$$\hat{\tau} = 0.8831 \pm 0.1249 \text{ (curvature)}, \qquad \hat{\tau} = 0.8831^{+0.1376}_{-0.114} \ (\Delta \ln \mathcal{L} = \tfrac{1}{2}).$$

Figure 3.40 shows the log-likelihood and its parabola: close near the peak, with the curve falling more steeply on the low side.

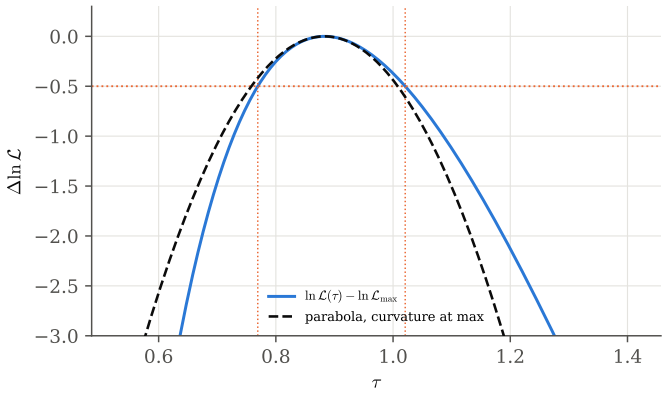


Figure 3.40: The log-likelihood of one experiment with fifty decays (solid) and the parabola with the same curvature at the peak (dashed). The dotted horizontal line is $\ln \mathcal{L}_{\max} - \frac{1}{2}$; its crossings with the curve (dotted verticals) give the asymmetric interval. The curve drops faster than the parabola on the left and slower on the right.

Then the script does what no experimenter can: it repeats the experiment 10^4 times and looks at the scatter.

Listing 3.12: Ten thousand experiments: scatter and coverage (`code/ch03/09_mle_decay.py`)

```
# ---------- 10^4 experiments ----------
T = rng.exponential(tau_true, (M, n))
th = T.mean(1)                                              # MLE of each experiment
sc = th / np.sqrt(n)                                        # its curvature error
sd_mc = th.std(ddof=1)
cover_curv = np.mean(np.abs(th - tau_true) <= sc)
# Delta lnL = 1/2 interval: solve n[ln(tau/th) + th/tau - 1] = 1/2 on both sides, per
    experiment
# in units u = tau/th the equation is u-independent: ln u + 1/u - 1 = 1/(2n)
g = lambda u: np.log(u) + 1 / u - 1 - 0.5 / n
ulo = optimize.brentq(g, 0.3, 1.0); uhi = optimize.brentq(g, 1.0, 3.0)
cover_dl = np.mean((ulo * th <= tau_true) & (tau_true <= uhi * th))
```

The numbers are

quantity	value
mean of $\hat{\tau}$ over 10^4 experiments	1.003
true scatter of $\hat{\tau}$ (Monte Carlo)	0.1406
$1/\sqrt{I_n(\tau)} = \tau/\sqrt{n}$	0.1414
average curvature error $\hat{\tau}/\sqrt{n}$	0.1419
coverage of $\hat{\tau} \pm \hat{\tau}/\sqrt{n}$	0.6913
coverage of the $\Delta \ln \mathcal{L} = \frac{1}{2}$ interval	0.6866

The scatter of the MLE over repetitions is the inverse square root of the Fisher information, to within the Monte Carlo error (about 0.7% for 10^4 experiments). The curvature error of an individual experiment fluctuates with $\hat{\tau}$, as the right panel of Figure 3.41 shows, but it is correct on average. Both intervals cover the true value in about 68% of the experiments, the fraction a $\pm 1\sigma$ Gaussian interval should cover (0.683). At $n = 50$ the asymptotic theory is already good. The left panel shows that the exact sampling distribution of $\hat{\tau}$, a Gamma distribution with shape n (the sum of n exponentials, Chapter 2), is slightly skewed compared with the Gaussian of (3.99); the skewness shrinks like $1/\sqrt{n}$.

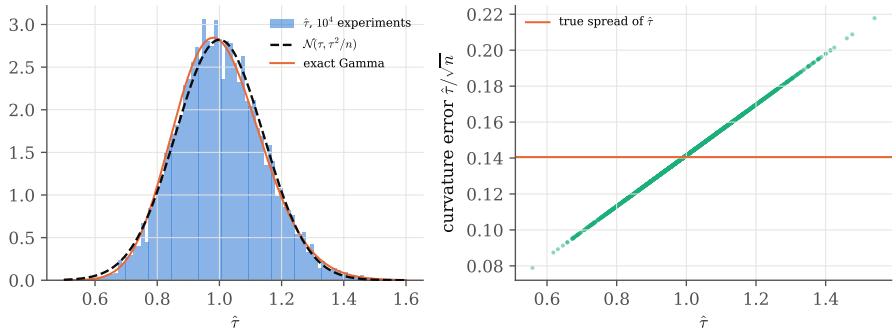


Figure 3.41: Ten thousand repetitions of the fifty-decay experiment. Left: the histogram of the maximum likelihood lifetimes, the Gaussian predicted by the Fisher information (dashed) and the exact Gamma distribution (solid). Right: each experiment's curvature error plotted against its estimate. The error grows with $\hat{\tau}$, because it is $\hat{\tau}/\sqrt{n}$, and it scatters around the true spread of the estimates (horizontal line).

3.13 What makes maximum likelihood good

We now have many ways to turn data into a number: the mean, the median, the method of moments, least squares, maximum likelihood. Why is maximum likelihood the default of experimental physics? Not because it is always unbiased. The Gaussian width $\hat{\sigma}^2$ is not (Section 3.9.1), and the decay rate $1/\hat{\tau}$ is not either (Section 3.13.2). The reason is a set of large-sample properties [Wasserman p. 125]. For a regular model (smooth in θ, with a support that does not move with θ, and a true value inside the parameter space) the MLE is

1. *consistent*: it converges to the true value;
2. *equivariant*: the MLE of $g(\theta)$ is $g(\hat{\theta})$;
3. *asymptotically Gaussian*, with variance $1/I_n(\theta)$ (Section 3.12.3);
4. *asymptotically efficient*: no unbiased estimator has a smaller variance than $1/I_n$, so for large n nothing beats it;
5. *approximately Bayesian*: for large n the posterior with any smooth prior peaks near $\hat{\theta}$ with width $1/\sqrt{I_n}$ (Chapter 5).

The third is (3.99). Each of the first, second and fourth has a short reason, each can be seen working on simulated data, and each fails when one of the regularity conditions fails.

3.13.1 Consistency: the log-likelihood peaks at the truth

Why should the peak of the likelihood of one finite data set sit near the true value $\theta_\star$? The answer is that, as n grows, the log-likelihood per observation settles down to a fixed curve, and that curve peaks at $\theta_\star$. To see it, compare each θ with the truth through the average log-likelihood ratio

$$M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ln \frac{f(x_i;\theta)}{f(x_i;\theta_\star)}.$$

It differs from $\ell(\theta)/n$ only by a term that does not depend on θ, so it peaks at $\hat\theta$. It is an average of n independent terms, so by the law of large numbers it approaches its expectation,

$$M_n(\theta) \;\to\; \mathbb{E}_{\theta_\star}\!\left[\ln \frac{f(X;\theta)}{f(X;\theta_\star)}\right] \overset{(1)}{=} -\int f(x;\theta_\star)\ln\frac{f(x;\theta_\star)}{f(x;\theta)}\,\mathrm{d}x \equiv -D(\theta_\star,\theta), \tag{3.100}$$

(1) Write the expectation as an integral over the true density and flip the ratio inside the logarithm, which flips the sign.

where $D(f,g) = \int f \ln(f/g)$ is the **Kullback–Leibler divergence** [how badly the density g describes data that really come from f, in units of log-probability per observation] [Wasserman p. 126]. It is never negative, and it vanishes only when the two densities are equal:

$$-D(f,g) \overset{(1)}{=} \mathbb{E}_f\left[\ln\frac{g(X)}{f(X)}\right] \overset{(2)}{\le} \ln \mathbb{E}_f\left[\frac{g(X)}{f(X)}\right] \overset{(3)}{=} \ln\int f(x)\frac{g(x)}{f(x)}\,\mathrm{d}x \overset{(4)}{=} \ln 1 = 0.$$

(1) Definition of D with the ratio inverted.
(2) Jensen's inequality: ln is concave, so the average of the logarithm is at most the logarithm of the average, with equality only if g/f is constant.
(3) Expectation over f.
(4) g is a density and integrates to one. A constant ratio g/f must therefore equal one.

Physicists know this as the Gibbs inequality of statistical mechanics: relative entropy is non-negative. So the limit of M_n is a function with a unique maximum, equal to zero, at $\theta = \theta_\star$, provided different parameter values give different densities (the model is **identifiable**). The M_n curves converge to that function, and their peaks, which are the MLEs, converge to its peak. That is Wasserman's consistency theorem [Wasserman p. 127, Thm. 9.13]; the proof needs the convergence to hold uniformly in θ, which is where regularity conditions enter.

Example: the lifetime. For exponential decays,

$$D(\tau_\star,\tau) \overset{(1)}{=} \mathbb{E}_{\tau_\star}\left[\ln\frac{\tau_\star^{-1}e^{-t/\tau_\star}}{\tau^{-1}e^{-t/\tau}}\right] \overset{(2)}{=} \ln\frac{\tau}{\tau_\star} + \mathbb{E}_{\tau_\star}[t]\left(\frac{1}{\tau}-\frac{1}{\tau_\star}\right) \overset{(3)}{=} \ln\frac{\tau}{\tau_\star} + \frac{\tau_\star}{\tau} - 1.$$

(1) Definition (3.100) with the exponential density.
(2) ln of the ratio is $\ln(\tau/\tau_\star) - t/\tau_\star + t/\tau$; take the expectation term by term.
(3) $\mathbb{E}_{\tau_\star} t = \tau_\star$.

This is the same function $\ln u + 1/u - 1$ that appeared in the exact $\Delta\ln\mathcal{L}$ interval, now with $u = \tau/\tau_\star$: the expected shape of the log-likelihood, per event, is minus the Kullback–Leibler divergence. The script `13_efficiency.py` draws $M_n(\tau)$ for one data set each of $n = 10$, 100 and

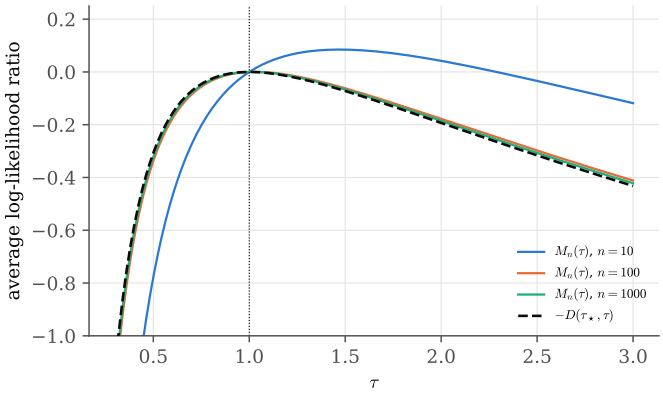


Figure 3.42: Why the MLE is consistent. The average log-likelihood ratio $M_n(\tau)$ of one simulated data set, for 10, 100 and 1000 decays, approaches minus the Kullback–Leibler divergence (dashed), which has its maximum, zero, exactly at the true lifetime (dotted). The peak of each M_n is that data set's MLE, and it is dragged towards the truth as the curves converge.

1000 decays (Figure 3.42). Their peaks, the MLEs, are at 1.47, 1.031 and 1.016, and the curves close in on $-D(\tau_\star, \tau)$ as n grows.

The same argument says what happens when the model is wrong. If the data come from a density g that is not in the family $f(x;\theta)$, M_n still converges, now to $-D(g, f(\cdot;\theta))$ plus a constant, and the MLE converges to the parameter whose density is closest to the truth in the Kullback–Leibler sense. The fit converges, but to the best wrong answer.

3.13.2 Invariance: the MLE does not care what we call the parameter

If $\eta = g(\theta)$ is a one-to-one reparametrisation, the likelihood in terms of η is $\mathcal{L}^*(\eta) = \mathcal{L}(g^{-1}(\eta))$: the same numbers attached to the same hypotheses (Section 3.8.3). Its maximum is therefore at the η that corresponds to $\hat\theta$,

$$\hat\eta = g(\hat\theta)$$

[Wasserman p. 127, Thm. 9.14], [CasellaBerger Thm. 7.2.10]. For a g that is not one-to-one (say $\eta = \theta^2$), Casella and Berger define the likelihood of η as the largest $\mathcal{L}(\theta)$ over all θ with $g(\theta) = \eta$, and the statement still holds. The MLE of the decay rate is $1/\hat\tau$, the MLE of σ is $\sqrt{\hat\sigma^2}$, and the MLE of the Hubble time is $1/\hat H_0$.

Unbiasedness does not survive such a change. Take $\hat\lambda = 1/\hat\tau = n/T$ with $T = \sum_i t_i$, a sum of n exponentials, which has the Gamma density $T^{n-1}e^{-T/\tau}/[\Gamma(n)\tau^n]$ (Chapter 2). Then

$$\mathbb{E}\Big[\frac{1}{T}\Big] \overset{(1)}{=} \int_0^\infty \frac{1}{T}\,\frac{T^{n-1}e^{-T/\tau}}{\Gamma(n)\tau^n}\,\mathrm{d}T \overset{(2)}{=} \frac{\Gamma(n-1)\,\tau^{n-1}}{\Gamma(n)\,\tau^n} \overset{(3)}{=} \frac{1}{(n-1)\tau}, \qquad \mathbb{E}\,\hat\lambda = \frac{n}{n-1}\,\lambda.$$

(1) Expectation of $1/T$ over the Gamma density.
(2) The remaining integrand is an unnormalised Gamma density of shape $n-1$, and $\int_0^\infty T^{n-2}e^{-T/\tau}\mathrm{d}T = \Gamma(n-1)\,\tau^{n-1}$.
(3) $\Gamma(n) = (n-1)\Gamma(n-1)$.

For $n = 50$ the bias factor is $50/49 = 1.02$; the 10^4 experiments of `09_mle_decay.py` give $\mathbb{E}\,\hat\lambda = 1.017$. Since $\hat\tau$ is exactly unbiased, $\hat\lambda$ cannot be. A nonlinear function of an unbiased estimator is biased, by Jensen's inequality, just as the sample standard deviation S is biased for σ even though S^2 is unbiased for σ^2 (Section 3.4.1). So an estimator can be unbiased in one parametrisation only. Maximum likelihood gives up unbiasedness and gains consistency in every parametrisation at once.

The error bar of $\hat\eta = g(\hat\theta)$ follows from the delta method (Section 3.6.3), now applied to the MLE [Wasserman p. 131, Thm. 9.24]: $\widehat{\mathrm{se}}(\hat\eta) \approx |g'(\hat\theta)|\,\widehat{\mathrm{se}}(\hat\theta)$. For $\lambda = 1/\tau$, $|g'| = 1/\tau^2$ and $\widehat{\mathrm{se}}(\hat\lambda) = \hat\tau^{-2}\cdot\hat\tau/\sqrt{n} = \hat\lambda/\sqrt{n}$, the same relative error, as it must be to first order. With several parameters the gradient replaces the derivative, $\widehat{\mathrm{se}}^2(\hat\eta) \approx \nabla g^{\mathsf T}\,\widehat{\mathrm{Cov}}(\hat{\boldsymbol\theta})\,\nabla g$, with the covariance from the inverse Fisher matrix [Wasserman p. 134, Thm. 9.28]. For example, take the coefficient of variation $\eta = \sigma/\mu$ of Gaussian data. One observation has $\ln f = -\ln\sigma - (x-\mu)^2/2\sigma^2 + \text{const}$. Its second derivatives are $\partial^2_\mu \ln f = -1/\sigma^2$, $\partial_\mu\partial_\sigma \ln f = -2(x-\mu)/\sigma^3$ and $\partial^2_\sigma \ln f = 1/\sigma^2 - 3(x-\mu)^2/\sigma^4$. Averaging with $\mathbb{E}(x-\mu) = 0$ and $\mathbb{E}(x-\mu)^2 = \sigma^2$ and multiplying by $-n$ gives a diagonal Fisher matrix with entries n/σ^2 and $2n/\sigma^2$, so $\mathrm{Var}\,\hat\mu \approx \sigma^2/n$ and $\mathrm{Var}\,\hat\sigma \approx \sigma^2/2n$. The gradient of η is $(\partial_\mu\eta, \partial_\sigma\eta) = (-\sigma/\mu^2,\ 1/\mu)$. Then $\widehat{\mathrm{se}}^2(\hat\eta) = \frac{\hat\sigma^2}{\hat\mu^4}\frac{\hat\sigma^2}{n} + \frac{1}{\hat\mu^2}\frac{\hat\sigma^2}{2n} = \hat\sigma^2/(n\hat\mu^2)\,[\hat\sigma^2/\hat\mu^2 + \tfrac12]$ [Wasserman p. 134, Ex. 9.29].

3.13.3 The Cramér–Rao bound: nobody can beat the information

The MLE has asymptotic variance $1/I_n(\theta)$. Could a cleverer estimator do better? The **Cramér–Rao inequality** (also called the Rao–Cramér–Fréchet bound) says no, for any unbiased estimator, at any sample size [CasellaBerger Thm. 7.3.9], [Cowan §6.6]. The proof is a single use of the Cauchy–Schwarz inequality for covariances, $\mathrm{Cov}(A,B)^2 \le \mathrm{Var}\,A\ \mathrm{Var}\,B$, applied to the estimator and the score.

Let $T(\boldsymbol d)$ be any estimator with mean $m(\theta) = \mathbb{E}_\theta\,T$, and $S(\theta) = \partial_\theta \ln p(\boldsymbol d\,|\,\theta)$ the score of the whole data set. Then

$$\mathrm{Cov}_\theta(T,S) \overset{(1)}{=} \mathbb{E}_\theta[T\,S] \overset{(2)}{=} \int T(\boldsymbol d)\,\frac{\partial_\theta p(\boldsymbol d\,|\,\theta)}{p(\boldsymbol d\,|\,\theta)}\,p(\boldsymbol d\,|\,\theta)\,\mathrm d\boldsymbol d \overset{(3)}{=} \frac{\partial}{\partial\theta}\int T(\boldsymbol d)\,p(\boldsymbol d\,|\,\theta)\,\mathrm d\boldsymbol d \overset{(4)}{=} m'(\theta),$$

$$m'(\theta)^2 \overset{(5)}{\le} \mathrm{Var}_\theta\,T\cdot\mathrm{Var}_\theta\,S \overset{(6)}{=} \mathrm{Var}_\theta\,T\cdot I_n(\theta) \quad\Longrightarrow\quad \mathrm{Var}_\theta\,T \ \ge\ \frac{m'(\theta)^2}{I_n(\theta)}. \tag{3.101}$$

(1) $\mathrm{Cov}(T,S) = \mathbb{E}[TS] - \mathbb{E}\,T\ \mathbb{E}\,S$, and $\mathbb{E}\,S = 0$ by (3.97).
(2) Write out the expectation; $S = \partial_\theta p/p$.
(3) Cancel p and move the derivative outside. T depends on the data only, not on θ. This exchange is the regularity condition.
(4) The integral is the definition of $m(\theta)$.
(5) Cauchy–Schwarz for covariances, with step (4) on the left.
(6) $\mathrm{Var}\,S = I_n$, (3.98).

For an unbiased estimator $m(\theta) = \theta$ and $m' = 1$, so that $\mathrm{Var}\,T \ge 1/I_n(\theta)$, which for independent, identically distributed data is $1/[nI(\theta)]$ [CasellaBerger Cor. 7.3.10]. If T has a bias $b(\theta)$, then $m = \theta + b$ and the bound is $(1+b')^2/I_n$, Cowan's form of the inequality.

Key idea: Cramér–Rao

For a regular model, every unbiased estimator satisfies

$$\mathrm{Var}_\theta\,\hat\theta \ \ge\ \frac{1}{I_n(\theta)} = \frac{1}{nI(\theta)}.$$

The Fisher information is the best precision the data can possibly give. The MLE reaches the bound as $n\to\infty$; it is **asymptotically efficient**. For several parameters, $\mathrm{Cov}\,\hat{\boldsymbol\theta} - \mathbf F^{-1}$ is positive semidefinite, where $\mathbf F$ is the Fisher matrix (Chapter 10).

When is the bound reached exactly? Cauchy–Schwarz is an equality only when the two variables are linearly related. The bound is attained by an unbiased T exactly when $S(\theta) =$

$a(\theta)\,[T(\boldsymbol{d}) - \theta]$ for some function a [CasellaBerger Cor. 7.3.15]. This happens for the exponential families of distributions [Wasserman p. 141]. The lifetime is an example: its score, $-n/\tau + T/\tau^2 = (n/\tau^2)(\bar{t} - \tau)$, has exactly this form, so $\bar{t}$ attains the bound τ^2/n at every n, not only asymptotically. So does $\bar{x}$ for a Gaussian mean. For the Gaussian variance with unknown mean the bound is $2\sigma^4/n$, while the best unbiased estimator, S^2, has variance $2\sigma^4/(n-1)$: close, but the bound is not attained [CasellaBerger Ex. 7.3.14].

When the bound does not apply: the uniform box again. For Uniform$(0, \theta)$, a blind application of the recipe gives $\partial_\theta \ln(1/\theta) = -1/\theta$, "information" $1/\theta^2$ per point, and a "bound" θ^2/n. But the unbiased estimator $\frac{n+1}{n} x_{(n)}$ does far better. Since $\mathrm{P}(x_{(n)} \le y) = (y/\theta)^n$, the scaled maximum $y = x_{(n)}/\theta$ has density $n y^{n-1}$ on $[0, 1]$, and

$$\mathbb{E}\, y \overset{(1)}{=} \frac{n}{n+1}, \qquad \mathbb{E}\, y^2 \overset{(2)}{=} \frac{n}{n+2},$$
$$\operatorname{Var}\Big[\frac{n+1}{n} x_{(n)}\Big] \overset{(3)}{=} \frac{(n+1)^2}{n^2}\,\theta^2\Big[\frac{n}{n+2} - \frac{n^2}{(n+1)^2}\Big] \overset{(4)}{=} \frac{\theta^2}{n(n+2)}.$$

(1) $\int_0^1 y \cdot n y^{n-1} \mathrm{d}y = n/(n+1)$.
(2) $\int_0^1 y^2 \cdot n y^{n-1} \mathrm{d}y = n/(n+2)$.
(3) $\operatorname{Var}(cX) = c^2 \operatorname{Var} X$ with $x_{(n)} = \theta y$, and $\operatorname{Var} y = \mathbb{E}\, y^2 - (\mathbb{E}\, y)^2$.
(4) Common denominator: $n(n+1)^2 - n^2(n+2) = n$, so the bracket is $n/[(n+1)^2(n+2)]$.

The variance falls like $1/n^2$, not $1/n$. The simulation in `13_efficiency.py` gives $n \operatorname{Var}/\theta^2 = 0.01877$ at $n = 51$, against the exact $1/(n+2) = 0.01887$ and the "bound" 1. There is no contradiction. Step (3) of (3.101) moved a derivative through an integral. Here the upper limit of that integral is θ itself, so differentiating it produces an extra boundary term and the exchange is not allowed. Indeed the "score" $-n/\theta$ does not have mean zero: it is not random at all [CasellaBerger Ex. 7.3.13]. The lesson is that a sharp edge in the data is worth more than any smooth feature, because a single point near the edge pins it down. Physicists use this all the time: the endpoint of a beta spectrum or of a kinematic distribution measures a mass far more sharply than the bulk of the distribution does.

3.13.4 Efficiency in practice: the mean against the median

The ratio of the variances of two consistent estimators, at large n, is their **asymptotic relative efficiency** [Wasserman p. 131, Thm. 9.23]. For Gaussian data the mean is the MLE of the centre and attains the bound σ^2/n. The sample median also estimates the centre, and it is popular because a single wild point cannot move it much. How much precision does it cost?

The asymptotic variance of the median follows from a counting argument. Let m be the true median and $f(m)$ the density there. The sample median lies below $m + \delta$ exactly when at least half the points do. The number of points below $m + \delta$ is binomial with n trials and success probability $F(m + \delta) \approx \frac{1}{2} + f(m)\delta$, so

$$\mathrm{P}(\hat{m} \le m + \delta) \overset{(1)}{=} \mathrm{P}\Big(\mathrm{Bin}\big(n, \tfrac{1}{2} + f\delta\big) \ge \tfrac{n}{2}\Big) \overset{(2)}{=} \mathrm{P}\Big(Z \ge \frac{n/2 - n/2 - nf\delta}{\sqrt{n}/2}\Big) \overset{(3)}{=} \Phi\big(2\sqrt{n}\, f(m)\, \delta\big),$$
$$\operatorname{Var} \hat{m} \overset{(4)}{=} \frac{1}{4n f(m)^2} \overset{(5)}{=} \frac{\pi\sigma^2}{2n} \quad \text{(Gaussian data)}.$$

(1) The median is below $m + \delta$ if at least $n/2$ points are. Near m, $F(m + \delta) \approx F(m) + f(m)\delta$ with $F(m) = \frac{1}{2}$.

(2) Gaussian approximation of the binomial (Chapter 2): mean $n(\frac{1}{2} + f\delta)$, variance $\approx n/4$ for small δ.
(3) $P(Z \geq -a) = \Phi(a)$ for a standard Gaussian Z.
(4) Step (3) says $\hat{m} - m$ is Gaussian with standard deviation $1/(2\sqrt{n}f)$.
(5) For $\mathcal{N}(\mu, \sigma^2)$, $f(m) = 1/(\sqrt{2\pi}\sigma)$ and $1/(4f^2) = 2\pi\sigma^2/4$.

The median's relative efficiency is $(\sigma^2/n)/(\pi\sigma^2/2n) = 2/\pi = 0.64$: using the median on Gaussian data is like throwing away 36% of the events [Wasserman p. 131]. The simulation, 60 000 data sets at each of several n, confirms it (Figure 3.43): at $n = 201$, n Var is 0.9981 σ^2 for the mean and 1.578 σ^2 for the median ($\pi/2 = 1.571$), an efficiency of 0.6327 against $2/\pi = 0.6366$.

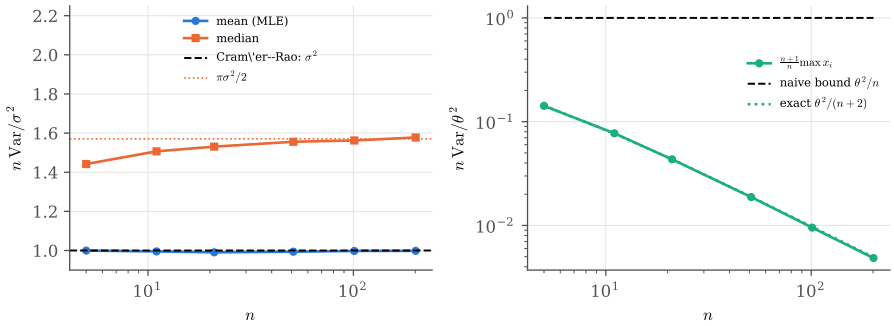


Figure 3.43: Efficiency, simulated. Left: n times the variance of the mean (the MLE) and of the median for Gaussian data. The mean sits on the Cramér–Rao bound at every n; the median approaches $\pi\sigma^2/2$ (dotted). Right: for the uniform box the estimator $\frac{n+1}{n}\max x_i$ falls far below the naive "bound" θ^2/n (dashed) and follows the exact $\theta^2/(n+2)$ (dotted), because the regularity conditions fail.

The lost precision buys robustness. For the Cauchy distribution, whose tails are so heavy that the sample mean does not converge at all, the median is consistent, with variance $\pi^2\gamma^2/(4n)$ for half-width γ (the same formula with $f(m) = 1/(\pi\gamma)$). The MLE is optimal only *within the model*. When the real data have heavier tails than the model allows, an estimator that is slightly inefficient under the model but robust to outliers can be far better in practice.

Watch out: The fine print of "the MLE is optimal"

Every property above is asymptotic and assumes a regular model with the true value inside the parameter space. At small n the MLE can be biased ($\hat{\sigma}^2$, $\hat{\lambda}$). At a boundary (a signal strength fitted to zero) the sampling distribution is not Gaussian. With a moving support (the uniform box) the Cramér–Rao bound does not apply. When the number of nuisance parameters grows with the data (one unknown mean per pair of measurements, say), the MLE of the common parameter need not even be consistent. Check with a simulation whenever any of these apply.

3.14 Nuisance parameters and the profile likelihood

Almost no real likelihood has only the parameters we care about. A search for a new particle wants the signal rate, but the background rate is also unknown. A cosmological fit wants Ω_m and σ_8, but the galaxy bias, the calibration of the telescope and the amplitude of a foreground come along. These are **nuisance parameters** [parameters that the likelihood depends on and the data constrain, but whose values we do not want to report] [Wasserman p. 120]. They cannot be ignored. Fixing them at a guessed value pretends we know them, and that makes the error on the interesting parameter too small. Three questions need answers.

1. How do we reduce a likelihood of several parameters to a function of the one we care about, without pretending the others are known?
2. Why does the honest answer have a wider interval, and by how much?
3. How does this compare with the Bayesian way of removing a nuisance parameter, by integrating it out?

3.14.1 A signal on top of a background

A standard counting experiment has two regions. In the *signal region* we count n_{on} events, expected to be $s + b$: signal plus background. In a *control region*, chosen so that it contains no signal, we count n_{off} events, expected to be kb, where the known factor k is the ratio of exposures (here $k = 1$). Both counts are Poisson, and they are independent:

$$\ell(s, b) = n_{\text{on}} \ln(s + b) - (s + b) + n_{\text{off}} \ln(kb) - kb.$$

The global maximum is easy. Setting the derivatives with respect to s and to b to zero gives the two score equations $n_{\text{on}}/(s + b) = 1$ and $n_{\text{on}}/(s + b) - 1 + n_{\text{off}}/b - k = 0$. By the first, the first two terms of the second cancel, which leaves $n_{\text{off}}/b = k$. So $\hat{b} = n_{\text{off}}/k$, and then the first gives $\hat{s} = n_{\text{on}} - n_{\text{off}}/k$. With $n_{\text{on}} = 20$ and $n_{\text{off}} = 8$, $\hat{s} = 12$ and $\hat{b} = 8$. The question is the error on $\hat{s}$.

Two ways to get a one-dimensional likelihood. To get the error on $\hat{s}$ we need a likelihood of s alone, and there are two ways to remove b. The lazy way is to fix b at its best value $\hat{b}$ and look at $\ell(s, \hat{b})$, the **conditional** likelihood. It pretends that the control region measured b perfectly. The honest way is to let b adjust: for each trial value of s, find the background that makes the data most probable, and use that.

Working definition 3.14 (profile likelihood)

For parameters $(\theta, \boldsymbol{\nu})$, with θ of interest and $\boldsymbol{\nu}$ nuisance, the **profile likelihood** is

$$\mathcal{L}_p(\theta) = \max_{\boldsymbol{\nu}} \mathcal{L}(\theta, \boldsymbol{\nu}) = \mathcal{L}\big(\theta, \hat{\hat{\boldsymbol{\nu}}}(\theta)\big),$$

where $\hat{\hat{\boldsymbol{\nu}}}(\theta)$ is the best value of the nuisance parameters at fixed θ [the ridge line of the likelihood, one point for every θ]. *Operationally:* on a grid of θ, maximise over everything else, record the maximum; then treat $\mathcal{L}_p$ like a one-parameter likelihood, with its maximum at the global $\hat{\theta}$.

For the on/off problem the inner maximisation can be done by hand. At fixed s,

$$\begin{aligned} 0 &\overset{(1)}{=} \frac{n_{\text{on}}}{s + b} - 1 + \frac{n_{\text{off}}}{b} - k \\ &\overset{(2)}{\Longrightarrow} n_{\text{on}}\, b - b(s + b) + n_{\text{off}}(s + b) - kb(s + b) = 0 \\ &\overset{(3)}{\Longrightarrow} (1 + k)\, b^2 + \big[(1 + k)s - n_{\text{on}} - n_{\text{off}}\big]\, b - n_{\text{off}}\, s = 0, \end{aligned}$$

(1) $\partial\ell/\partial b$ at fixed s.
(2) Multiply by $b(s + b) > 0$.
(3) Collect powers of b and multiply by -1.

and $\hat{\hat{b}}(s)$ is the positive root of this quadratic (the constant term is negative for $s > 0$, so there is exactly one). As s increases, the best background decreases: more of the n_{on} events are already explained by signal. That anticorrelation is what the conditional likelihood ignores.

3.14.2 Why profiling widens the interval

Near the maximum every smooth log-likelihood is quadratic, so it is enough to understand the Gaussian case. Take one parameter of interest θ and one nuisance parameter ν. Write $\boldsymbol{x} = (\theta - \hat\theta, \nu - \hat\nu)$ and $\ell = \ell_{\max} - \tfrac12 \boldsymbol{x}^{\mathsf T}\mathbf{F}\boldsymbol{x}$ with the information matrix $\mathbf{F} = \begin{pmatrix} F_{\theta\theta} & F_{\theta\nu} \\ F_{\theta\nu} & F_{\nu\nu} \end{pmatrix}$, whose inverse $V = \mathbf{F}^{-1}$ is the covariance matrix. Maximise over the nuisance direction:

$$
\begin{aligned}
\ell - \ell_{\max} &\overset{(1)}{=} -\tfrac12 \left[F_{\theta\theta} x_\theta^2 + 2F_{\theta\nu} x_\theta x_\nu + F_{\nu\nu} x_\nu^2 \right], \\
\frac{\partial \ell}{\partial x_\nu} = 0 &\overset{(2)}{\Longrightarrow} \; x_\nu = -\frac{F_{\theta\nu}}{F_{\nu\nu}}\, x_\theta, \\
\ell_p - \ell_{\max} &\overset{(3)}{=} -\tfrac12 \left(F_{\theta\theta} - \frac{F_{\theta\nu}^2}{F_{\nu\nu}} \right) x_\theta^2 \overset{(4)}{=} -\frac{x_\theta^2}{2V_{\theta\theta}}.
\end{aligned}
\tag{3.102}
$$

(1) Write out the quadratic form.
(2) Derivative in x_ν: $-(F_{\theta\nu}x_\theta + F_{\nu\nu}x_\nu) = 0$. This is the profile path, a straight line through the maximum.
(3) Substitute: $2F_{\theta\nu}x_\theta x_\nu + F_{\nu\nu}x_\nu^2 = -F_{\theta\nu}^2 x_\theta^2 / F_{\nu\nu}$.
(4) The (θ,θ) element of the inverse of a symmetric 2×2 matrix is $V_{\theta\theta} = F_{\nu\nu}/\det\mathbf{F}$, and $F_{\theta\theta} - F_{\theta\nu}^2/F_{\nu\nu} = \det\mathbf{F}/F_{\nu\nu} = 1/V_{\theta\theta}$.

The profile likelihood is a Gaussian of variance $V_{\theta\theta}$: the full marginal variance of $\hat\theta$, including the uncertainty it inherits from the nuisance parameter. The conditional likelihood at $\nu = \hat\nu$ has curvature $F_{\theta\theta}$, hence variance $1/F_{\theta\theta}$, and

$$
\frac{1}{F_{\theta\theta}} = V_{\theta\theta}\,(1 - \rho^2) \;\le\; V_{\theta\theta},
$$

where ρ is the correlation coefficient of $\hat\theta$ and $\hat\nu$. The identity follows from the same 2×2 inverse, whose entries are $V_{\theta\theta} = F_{\nu\nu}/\det\mathbf{F}$, $V_{\nu\nu} = F_{\theta\theta}/\det\mathbf{F}$ and $V_{\theta\nu} = -F_{\theta\nu}/\det\mathbf{F}$:

$$
\rho^2 = \frac{V_{\theta\nu}^2}{V_{\theta\theta}V_{\nu\nu}} = \frac{F_{\theta\nu}^2}{F_{\theta\theta}F_{\nu\nu}}, \qquad V_{\theta\theta}(1-\rho^2) = \frac{F_{\nu\nu}}{\det\mathbf{F}} \cdot \frac{F_{\theta\theta}F_{\nu\nu} - F_{\theta\nu}^2}{F_{\theta\theta}F_{\nu\nu}} = \frac{1}{F_{\theta\theta}}.
$$

So the conditional interval is too short by the factor $\sqrt{1-\rho^2}$. The stronger the correlation, the more it understates the error. Figure 3.44 shows the geometry.

Example: the on/off errors in the Gaussian limit. For large counts, $\hat s = n_{\text{on}} - n_{\text{off}}/k$ is a difference of independent Poisson counts. A Poisson count has variance equal to its mean, estimated by the count itself, so the variances are about n_{on} and n_{off}. By linear error propagation (Section 3.6.1), $\sigma_{\hat s}^2 \approx n_{\text{on}} + n_{\text{off}}/k^2$. For $(20, 8)$ with $k = 1$ that is $\sqrt{28} = 5.292$. Fixing b leaves only the fluctuation of n_{on}, $\sqrt{20} = 4.472$. The profile likelihood reproduces the first number, the conditional likelihood the second.

3.14.3 The on/off problem in Python

Listing 3.13: Profile and conditional intervals for the on/off problem (`code/ch03/14_profile.py`)

```
def bhat_of_s(s, non, noff):
    """maximise over b at fixed s: (1+k) b^2 + ((1+k) s - non - noff) b - noff s = 0"""
    A = 1 + k; B = (1 + k) * s - non - noff; C = -noff * s
    return (-B + np.sqrt(B * B - 4 * A * C)) / (2 * A)
```

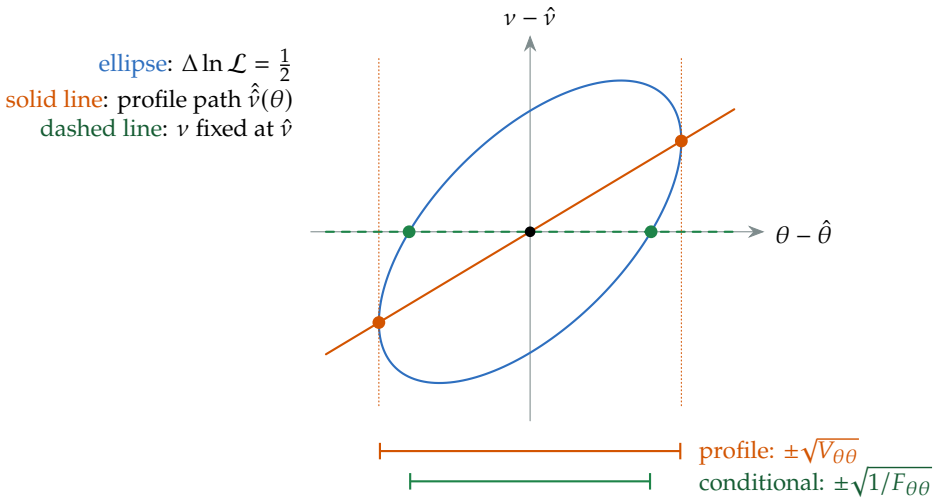


Figure 3.44: Profile versus conditional likelihood for a Gaussian likelihood with correlation $\rho = 0.6$ between the parameter of interest and the nuisance parameter. The profile path runs through the points where the ellipse has vertical tangents, so the profile likelihood drops by $\frac{1}{2}$ at the full width of the ellipse (orange bar). Holding the nuisance parameter fixed cuts the ellipse along the horizontal line and gives an interval shorter by the factor $\sqrt{1-\rho^2} = 0.8$ (green bar).

```
5
6
7  def intervals(non, noff):
8      b0 = noff / k; s0 = non - b0                        # the global MLE
9      Lmax = lnL(s0, b0, non, noff)
10     prof = lambda s: lnL(s, bhat_of_s(s, non, noff), non, noff) - Lmax + 0.5
11     cond = lambda s: lnL(s, b0, non, noff) - Lmax + 0.5
12     lo_s = -b0 + 1e-9                                    # s + b must stay positive
13     out = []
14     for f in (prof, cond):
15         lo = optimize.brentq(f, max(lo_s, s0 - 20 * np.sqrt(non + 1)), s0)
16         hi = optimize.brentq(f, s0, s0 + 20 * np.sqrt(non + 1))
17         out.append((lo, hi))
18     return s0, b0, out
```

The function `bhat_of_s` is the positive root of the quadratic for $\hat{\hat{b}}(s)$ in Section 3.14.1. The functions `prof` and `cond` are the two one-dimensional log-likelihoods shifted so that their zeros are the $\Delta \ln \mathcal{L} = \frac{1}{2}$ crossings, which `brentq` finds on either side of $\hat{s}$. For the observed counts the result is

$$\text{profile: } \hat{s} = 12^{+5.47}_{-5.192}, \qquad \text{conditional: } \hat{s} = 12^{+4.811}_{-4.145}.$$

The profile interval is about 20% wider, as the Gaussian estimates 5.292 and 4.472 said, and both are asymmetric because the counts are small. Figure 3.45 shows the two-dimensional contours, the profile path and the two curves.

Which interval is right? The script simulates 10^4 experiments with $s = 12$ and $b = 8$ and counts how often each interval contains the true signal. (In 9995 of them both intervals exist; the rest have $n_{\text{off}} = 0$ or a similar degenerate outcome.) The profile interval covers the truth in 0.6849 of the experiments, the 68% a $\pm 1\sigma$ interval promises. The conditional interval covers it in only 0.6088. The reason the profile works is **Wilks' theorem** again [22]: $-2\ln[\mathcal{L}_p(\theta)/\mathcal{L}_{\max}]$ at the true θ is asymptotically χ^2_1, whatever the number of nuisance parameters that were maximised over, so the $\Delta \ln \mathcal{L}_p = \frac{1}{2}$ interval has 68% coverage for large samples. This is what the `MINOS` errors of `MINUIT` compute, and it is the basis of the profile-likelihood test statistics used in every LHC search (Chapter 4).

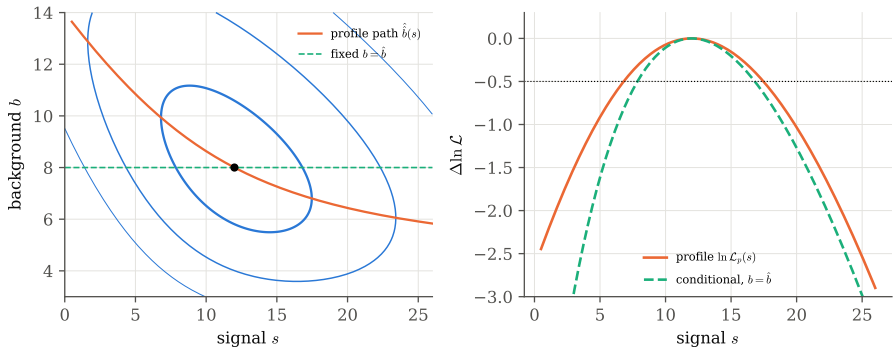


Figure 3.45: A signal on top of a background, from 20 counts in the signal region and 8 in the control region. Left: contours of $\ln \mathcal{L}(s, b)$ at $\Delta \ln \mathcal{L} = \frac{1}{2}$, 2 and 4.5, the profile path (the best background for each signal value, falling as s rises) and the line of fixed background. Right: the profile log-likelihood is wider than the conditional one; the dotted line at $-\frac{1}{2}$ marks the interval endpoints.

Key idea: Profile out nuisance parameters

Maximise over the nuisance parameters at each value of the parameter of interest, and read the interval from $\Delta \ln \mathcal{L}_p = \frac{1}{2}$. For a Gaussian likelihood this gives exactly the marginal error $\sqrt{V_{\theta\theta}}$; fixing the nuisance parameters instead gives $\sqrt{1/F_{\theta\theta}}$, too small by $\sqrt{1-\rho^2}$.

3.14.4 Profiling and marginalising

The Bayesian removes a nuisance parameter differently: by integrating it out of the posterior, $p(\theta \mid \boldsymbol{d}) = \int p(\theta, \nu \mid \boldsymbol{d})\, \mathrm{d}\nu$ [Padilla (38)], [HobsonBMC p. 49]. This is the marginalisation of a joint distribution that we have used since Chapter 1: average over the variable we do not want. For a Gaussian likelihood and a flat prior the two operations give the same answer, because the integral of a two-dimensional Gaussian over one variable is a Gaussian in the other with variance $V_{\theta\theta}$, the same as (3.102). They differ when the likelihood is not Gaussian, and the difference is itself a warning that the asymptotic theory is being stretched.

Example: an unknown noise level. Hobson et al. consider a fit where every datum has Gaussian noise of the same, unknown size σ [HobsonBMC p. 50, (2.25)–(2.26)]. With $Q(\theta) = \sum_i (D_i - d_i(\theta))^2$ the likelihood is $\mathcal{L}(\theta, \sigma) = (2\pi\sigma^2)^{-N/2} \exp[-Q(\theta)/2\sigma^2]$. Profile over σ:

$$\frac{\partial \ell}{\partial \sigma} \overset{(1)}{=} -\frac{N}{\sigma} + \frac{Q}{\sigma^3} = 0 \;\Rightarrow\; \hat{\sigma}^2(\theta) = \frac{Q(\theta)}{N}, \qquad \mathcal{L}_p(\theta) \overset{(2)}{=} \left(\frac{2\pi Q(\theta)}{N}\right)^{-N/2} e^{-N/2} \;\propto\; Q(\theta)^{-N/2}.$$

(1) Differentiate $-N \ln \sigma - Q/2\sigma^2$.
(2) Substitute: the exponent becomes $-Q/(2Q/N) = -N/2$.

Now marginalise instead, with the prior $p(\sigma) \propto 1/\sigma$ (flat in $\ln \sigma$, the prior Hobson et al. use):

$$\int_0^\infty \sigma^{-N} e^{-Q/2\sigma^2} \frac{\mathrm{d}\sigma}{\sigma} \overset{(1)}{=} \int_0^\infty \left(\frac{2u}{Q}\right)^{N/2} e^{-u} \frac{\mathrm{d}u}{2u} \overset{(2)}{=} \frac{1}{2}\left(\frac{2}{Q}\right)^{N/2} \Gamma\!\left(\frac{N}{2}\right) \;\propto\; Q(\theta)^{-N/2}.$$

(1) Substitute $u = Q/2\sigma^2$, so $\sigma^{-2} = 2u/Q$ and $\mathrm{d}\sigma/\sigma = -\mathrm{d}u/(2u)$ (the sign is absorbed by swapping the limits).
(2) $\int_0^\infty u^{N/2-1} e^{-u} \mathrm{d}u = \Gamma(N/2)$.

The two routes give the same function of θ, which is Hobson's (2.26) up to a constant. The unknown noise level turns the Gaussian $e^{-\chi^2/2}$ into the heavier-tailed $Q^{-N/2}$, a Student-t shape in θ (Chapter 2): not knowing the size of the error bars costs precision, and the likelihood knows it. The same device, an unknown scale factor on all quoted error bars, is Hobson's "safety valve" for data with underestimated errors [HobsonBMC p. 52].

3.15 The bootstrap: the sampling distribution from one data set

Every error bar so far came from a formula: $\sigma/\sqrt{n}$ for the mean, the curvature of $\ln \mathcal{L}$ for the MLE, the delta method for a function of estimates. To check those formulas we also used a luxury no experimenter has: we simulated 10^4 fresh experiments from the true distribution and looked at the scatter. Now suppose there is no formula (the median of lifetimes, a ratio of two fitted amplitudes, the peak position of a smoothed histogram) and no known truth to simulate from. All we have is one data set. The **bootstrap** lets that one data set stand in for the unknown distribution, and then simulates from it [23]. Three questions need answers.

1. What does it mean to "simulate from the data", and why does it estimate the sampling distribution?
2. How do we turn the simulated replicas into a standard error and a confidence interval?
3. When does it fail?

3.15.1 Two worlds

Let the data $x_1, \dots, x_n$ come independently from an unknown distribution F, and let $T_n = g(x_1, \dots, x_n)$ be any statistic. Its standard error is $\sqrt{\mathrm{Var}_F(T_n)}$, which we could compute by simulation if we knew F: draw many samples of size n from F, compute T_n on each, take the standard deviation. We do not know F. But we have an estimate of it, the **empirical distribution** $\hat{F}_n$ [the distribution that puts probability $1/n$ on each observed data point, and nothing elsewhere]. The bootstrap replaces F by $\hat{F}_n$ and then does the simulation [Wasserman pp. 108–109]:

$$\mathrm{Var}_F(T_n) \overset{\text{(i)}}{\approx} \mathrm{Var}_{\hat{F}_n}(T_n) \overset{\text{(ii)}}{\approx} v_{\text{boot}} = \frac{1}{B-1}\sum_{b=1}^{B}\left(T^*_{n,b} - \overline{T^*}\right)^2 .$$

Approximation (i) is the statistical one: it is good when $\hat{F}_n$ is close to F, which requires n to be reasonably large, and it cannot be improved without more data. Approximation (ii) is pure Monte Carlo: its error falls as $1/\sqrt{B}$, and we make it small by running the computer longer.

Drawing a sample of size n from $\hat{F}_n$ means picking n of the observed points at random, *with replacement*: some points appear twice or three times, others not at all. Each such **bootstrap sample** $x^*_1, \dots, x^*_n$ gives a **replica** T^*_n of the statistic. Figure 3.46 lays out the analogy.

Working definition 3.15 (nonparametric bootstrap)

To estimate the standard error of a statistic T_n from one data set of size n: for $b = 1, \dots, B$, draw n indices uniformly from $\{1, \dots, n\}$ with replacement, form the bootstrap sample, and compute the replica $T^*_{n,b}$. The bootstrap standard error is the sample standard deviation of the B replicas [Wasserman p. 109, (8.1)]. B of a few hundred suffices for a standard error; intervals need a few thousand.

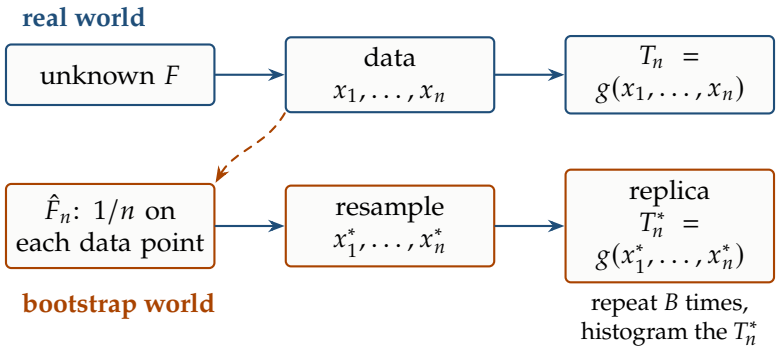


Figure 3.46: The two worlds of the bootstrap. In the real world an unknown distribution produced one data set and one value of the statistic, and we cannot repeat it. In the bootstrap world the empirical distribution (built from the data, dashed arrow) plays the role of the truth. We can draw from it as often as we like, and the scatter of the replicas stands in for the scatter of the statistic in the real world.

Example: the bootstrap of the mean, exactly. For the mean we can do the bootstrap world by hand. A single resampled point x^* equals each x_i with probability $1/n$, so

$$\mathbb{E}_{\hat{F}}\, x^* \overset{(1)}{=} \frac{1}{n}\sum_i x_i = \bar{x}, \qquad \mathrm{Var}_{\hat{F}}\, x^* \overset{(2)}{=} \frac{1}{n}\sum_i (x_i - \bar{x})^2 \equiv \hat{\sigma}^2,$$

$$\mathrm{Var}_{\hat{F}}\, \bar{x}^* \overset{(3)}{=} \frac{\hat{\sigma}^2}{n} \overset{(4)}{=} \frac{n-1}{n}\,\frac{S^2}{n}.$$

(1) The expectation over the empirical distribution is the plain average of the data.
(2) Its variance is the average squared deviation from that average, with divisor n.
(3) The n resampled points are independent draws from $\hat{F}$, so the variance of their mean is the single-point variance over n.
(4) $S^2 = \sum_i (x_i - \bar{x})^2/(n-1)$ is the unbiased sample variance.

The bootstrap reproduces the textbook standard error $S/\sqrt{n}$, up to the factor $\sqrt{(n-1)/n}$ that is negligible for n of a few tens. For the data $\{1, 2, 6\}$, $\bar{x} = 3$ and $\hat{\sigma}^2 = (4 + 1 + 9)/3 = 14/3$, so the bootstrap variance of the mean is $14/9 = 1.56$, against $S^2/n = 7/3 = 2.33$ from the unbiased formula: with three points the factor $(n-1)/n = 2/3$ matters. There are only $\binom{2n-1}{n} = 10$ distinct bootstrap samples here, so for tiny n the bootstrap distribution is coarse; for $n = 50$ there are about 5×10^{28}.

The parametric bootstrap. If we trust a parametric model $f(x;\theta)$, there is a better stand-in for F than $\hat{F}_n$: the fitted model $f(x;\hat{\theta})$. The **parametric bootstrap** simulates B data sets of size n from $f(x;\hat{\theta})$ and refits each one [Wasserman p. 134]. This is exactly what we did with the 10^4 simulated lifetime experiments of Section 3.12.4, except that the fitted value replaces the truth. It is also the standard way to calibrate a complicated analysis in physics: fit, simulate from the fit, analyse the simulations the same way, look at the scatter.

3.15.2 The bootstrap in Python

The script `15_bootstrap.py` takes one data set of $n = 50$ decay times and asks for the standard errors of two statistics: the mean $\hat{\tau}$, where we know the answer, and the median, where we have no simple formula (its asymptotic variance is $1/[4nf(m)^2]$ from Section 3.13.4, which needs the unknown density at the unknown median).

Listing 3.14: Nonparametric and parametric bootstrap of the mean and the median (code/ch03/15_bootstrap.py)

```
t = rng.exponential(tau_true, n)                        # the one data set we have
tau_hat, med_hat = t.mean(), np.median(t)

# the truth (available only because this is a simulation)
T = rng.exponential(tau_true, (B, n))
true_tau, true_med = T.mean(1), np.median(T, 1)

# nonparametric bootstrap: draw n points WITH replacement from the data
idx = rng.integers(0, n, (B, n))
Tb = t[idx]
np_tau, np_med = Tb.mean(1), np.median(Tb, 1)
# parametric bootstrap: simulate from the fitted model f(t; tau_hat)
Tp = rng.exponential(tau_hat, (B, n))
p_tau, p_med = Tp.mean(1), np.median(Tp, 1)

se = lambda a: a.std(ddof=1)
q = lambda a, p: np.quantile(a, p)
ci_norm = (med_hat - 1.96 * se(np_med), med_hat + 1.96 * se(np_med))
ci_piv = (2 * med_hat - q(np_med, 0.975), 2 * med_hat - q(np_med, 0.025))
ci_pct = (q(np_med, 0.025), q(np_med, 0.975))
```

Line 34 is the whole nonparametric bootstrap: a $B \times n$ array of random indices, used to pick data points with replacement, so that each row is one bootstrap sample. Line 38 is the parametric bootstrap, fresh exponential samples with the fitted lifetime. Lines 30–31 compute what no experimenter can, the true sampling distribution from 10^4 fresh experiments, as a referee.

standard error	true (fresh experiments)	nonparametric	parametric
mean $\hat{\tau}$	0.1425	0.1116	0.1162
median	0.1418	0.1352	0.1155

In this run the data set has $\hat{\tau} = 0.8176$ and median 0.5147, for true values 1 and $\ln 2 = 0.6931$. The bootstrap estimates the standard error of the distribution it simulates from, and that distribution is built from these fifty numbers, so it inherits their luck. The parametric bootstrap simulates from an exponential with mean $\hat{\tau}$, so its error for the mean follows $\hat{\tau}$, as does the curvature error $\hat{\tau}/\sqrt{n} = 0.1156$. The nonparametric bootstrap follows the spread of the data instead. A data set that comes out low or narrow gives error bars that are too small, one that comes out high or wide gives error bars that are too large, and no method that uses only these fifty numbers can tell which kind of draw it has. Dividing by the estimate removes the overall scale: the *relative* error here is $0.1116/0.8176 = 0.1365$ against the true $0.1425/1$. That too depends on the draw. Computed on each of the 10^4 fresh experiments, the same relative error averages 0.1375 but scatters by 0.01819 from one data set to the next, so a single data set can easily miss it by that much. For the median the bootstrap gives an error bar where we had no formula at all, with the same caveat: it comes from one data set and shares its luck. Figure 3.47 compares the shapes of the bootstrap distributions, each centred on its own estimate, with the true sampling distribution.

3.15.3 Bootstrap confidence intervals

A standard error gives the simplest interval, the **Normal interval** $T_n \pm z_{\alpha/2}\,\widehat{\mathrm{se}}_{\mathrm{boot}}$ [Wasserman p. 110, (8.2)], which assumes a Gaussian sampling distribution. The bootstrap distribution contains more than a width, and two other constructions use its shape.

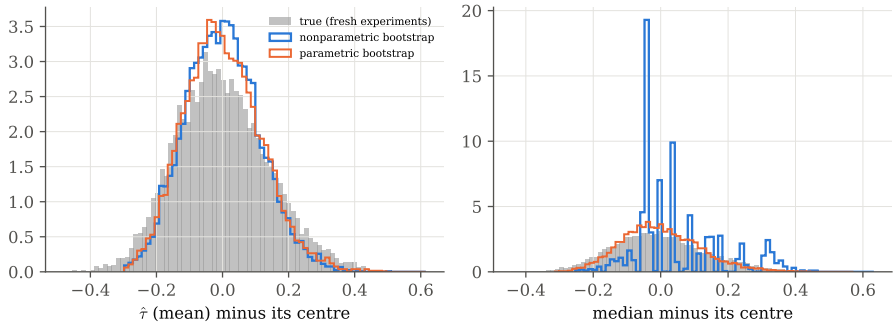


Figure 3.47: Bootstrap distributions from one data set of fifty decay times, each centred on its own centre so the shapes can be compared. Grey: the true sampling distribution from ten thousand fresh experiments. Blue: resampling the data with replacement. Orange: simulating from the fitted exponential. Left: the mean. Right: the median, whose nonparametric bootstrap distribution is lumpy because a resampled median can only take a few values near the middle of the data.

The pivotal interval. Let $R_n = \hat{\theta} - \theta$ be the error of the estimate, with unknown distribution function $H(r) = \mathrm{P}(R_n \le r)$. Choose $a = \hat{\theta} - H^{-1}(1-\alpha/2)$ and $b = \hat{\theta} - H^{-1}(\alpha/2)$. Then

$$\mathrm{P}(a \le \theta \le b) \overset{(1)}{=} \mathrm{P}\big(\hat{\theta} - b \le R_n \le \hat{\theta} - a\big) \overset{(2)}{=} H\big(\hat{\theta} - a\big) - H\big(\hat{\theta} - b\big)$$
$$\overset{(3)}{=} H\big(H^{-1}(1-\tfrac{\alpha}{2})\big) - H\big(H^{-1}(\tfrac{\alpha}{2})\big) = 1 - \alpha.$$

(1) $a \le \theta$ is the same as $\hat{\theta} - \theta \le \hat{\theta} - a$, and $\theta \le b$ is $\hat{\theta} - b \le \hat{\theta} - \theta$.
(2) Probability of an interval from the distribution function (for a continuous H).
(3) Insert the chosen a and b.

We do not know H, but the bootstrap gives its analogue: the distribution of $R^* = \theta^* - \hat{\theta}$, whose quantiles are $\theta^*_\beta - \hat{\theta}$ with θ^*_β the β-quantile of the replicas. Substituting gives the **pivotal interval** [Wasserman p. 111, (8.6)]

$$C_n = \big(2\hat{\theta} - \theta^*_{1-\alpha/2},\; 2\hat{\theta} - \theta^*_{\alpha/2}\big).$$

The interval is reflected. Suppose the replicas scatter further above $\hat{\theta}$ than below it. In the bootstrap world, then, the estimate overshoots its truth by larger amounts than it undershoots. Carried back to the real world, our $\hat{\theta}$ may sit well above the true θ, so the interval reaches further *down*.

The percentile interval. The simplest recipe of all takes the quantiles of the replicas directly, $(\theta^*_{\alpha/2}, \theta^*_{1-\alpha/2})$ [Wasserman p. 111]. It is exact when some monotone transformation makes the sampling distribution symmetric and Gaussian, and it respects the parametrisation in the way the $\Delta \ln \mathcal{L}$ interval does. All three intervals have coverage tending to $1-\alpha$ under weak conditions [Wasserman p. 111, Thm. 8.3].

Example: the median of the lifetimes. For the median of our fifty decay times, 95% intervals from 10^4 nonparametric replicas are

Normal: $(0.2496,\ 0.7798)$, pivotal: $(0.1678,\ 0.6632)$, percentile: $(0.3661,\ 0.8616)$,

for a true median of 0.6931. The replicas of a median are lumpy and lopsided, and the three recipes treat the lopsidedness differently: the percentile interval follows it, the pivotal interval reflects it about the estimate, and the Normal interval ignores it. So the three intervals differ,

and for one data set one of them can miss the true median while another contains it. Any 95% procedure misses in one experiment in twenty. Which of them misses less often for a given problem can be settled only by the same kind of simulation over many data sets, as in Chapter 4.

3.15.4 Where the bootstrap fails

The bootstrap works when the statistic depends smoothly on the whole distribution. It fails when the statistic is controlled by a few extreme points, because $\hat{F}_n$ has no probability beyond the largest observation. The cleanest example is the uniform box again, with $T_n = x_{(n)}$. The true sampling distribution of the maximum is continuous, with density $n x^{n-1}/\theta^n$ below θ. A bootstrap replica, however, is the maximum of n points drawn from the data, and it equals the observed maximum whenever that point is drawn at least once:

$$\mathrm{P}\big(x^*_{(n)} = x_{(n)}\big) \overset{(1)}{=} 1 - \mathrm{P}\big(x_{(n)} \text{ is never drawn}\big) \overset{(2)}{=} 1 - \left(1 - \frac{1}{n}\right)^n \xrightarrow{\;n\to\infty\;} 1 - e^{-1} = 0.632.$$

(1) The replica's maximum is $x_{(n)}$ as soon as $x_{(n)}$ is among the resampled points; otherwise it is smaller.

(2) Each of the n independent draws misses that point with probability $1 - 1/n$.

For $n = 50$ this is 0.6358, and the simulation gives 0.6386: almost two thirds of the replicas sit on a single value, a spike that the continuous true distribution does not have (Figure 3.48). No amount of resampling can put a replica above $x_{(n)}$, yet the true θ is always there. The same caution applies to any statistic that lives in the tails: extreme values, endpoint fits, the largest peak in a noisy map. For such statistics the parametric bootstrap, or a simulation from a physical model, is the right tool.

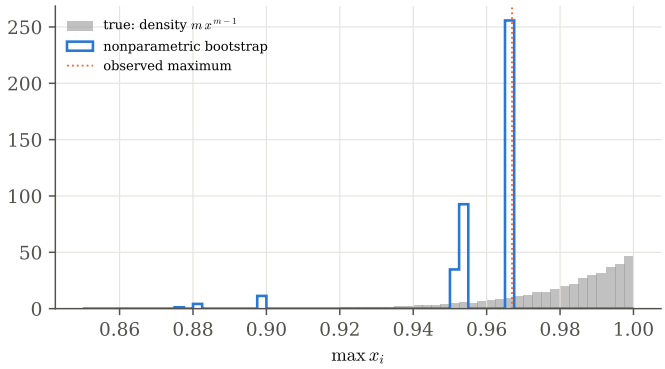


Figure 3.48: A statistic the bootstrap cannot handle: the maximum of fifty uniform points. Grey: its true sampling distribution, smooth and piled against the end of the box. Blue: the bootstrap replicas, about 63% of which equal the observed maximum (dotted line), with the rest spread over a few lower data values. None can exceed the observed maximum.

How this is used in real research: Verification versus research

Most of the simulations above *verified* results already derived: that $\chi^2_{\min}$ follows χ^2_{N-m}, that the curvature of $\ln \mathcal{L}$ equals the scatter of the MLE, that the Poisson fit conserves the total. Their job was to make the formulas believable and to show how large n must be before the asymptotic statements hold. In research the same machinery is the instrument itself, used exactly where no formula exists. A cosmological analysis estimates the covariance matrix of a masked, noisy power spectrum from hundreds of simulated skies, a parametric bootstrap from the fitted model (Chapters 7 and 8); an LHC search calibrates its profile-likelihood test statistic on pseudo-experiments when Wilks' theorem is doubtful; large-scale-structure

analyses resample patches of the sky (a jackknife or block bootstrap) to capture correlations no one can model analytically. The logic is always that of Figure 3.46: build the best stand-in for the truth, simulate from it, analyse each simulation exactly as the data, read the error off the scatter.

3.16 From a model to an estimator: how to find the right one

A theorist hands us a model: a prediction $d(\theta)$, a noise level, and a parameter θ to be measured. Nothing says which function of the data to quote as the answer. The sample mean, the median, the midrange, the value that minimises some χ^2: each is an estimator, each has a sampling distribution, and they differ in width. So the question is: which function of the data should we quote? The tools for answering it are already in hand. The likelihood gives the probability of the data. The score equation turns it into an estimate. The curvature gives an error bar. The Cramér–Rao bound says how narrow the error bar can ever be. Put in order, these tools become a procedure that starts from the model and ends with an estimator whose sampling distribution we know and cannot improve (Figure 3.49). The procedure has five steps, and every one has already appeared in a special case.

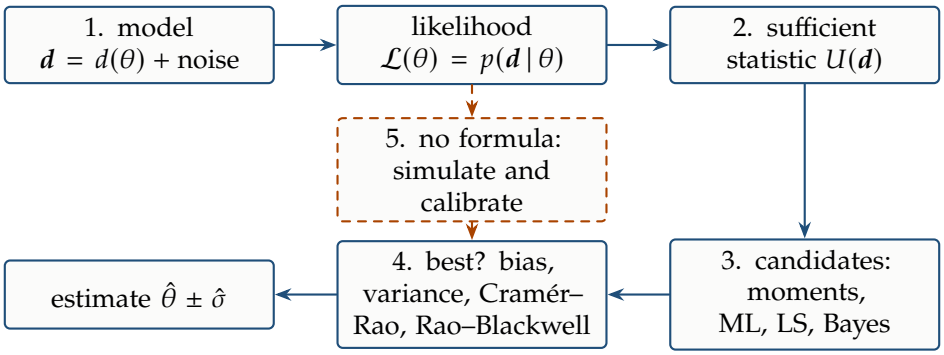


Figure 3.49: From a model to an estimator. The solid path is the procedure when the likelihood can be written down. If it cannot (dashed path), a simple statistic is computed and its bias and scatter are read off simulated data sets.

Four examples run through every step: independent Gaussian readings $x_i \sim \mathcal{N}(\mu, \sigma^2)$ with known σ; Poisson counts $X_i \sim \text{Poisson}(\lambda)$, the numbers of events in n equal time slots; exponential decay times t_i with mean lifetime τ; and the multipole coefficients $a_{\ell m}$ of a full-sky Gaussian CMB map. The CMB example is then followed through the whole chain in Section 3.16.7.

3.16.1 One measurement, five jobs

Each tool met so far has one job. One measurement model, covering most of what a physicist meets, shows all of them:

$$x_i = a_i\theta + b_i + \epsilon_i, \qquad i = 1, \dots, N. \tag{3.103}$$

Here θ is the parameter we want; a_i is the known sensitivity of the instrument in reading i, how much that reading changes per unit change of θ (a gain, a filter transmission, a template value); b_i is any other contribution to the reading (a background, a foreground, an offset); and ϵ_i is noise. Theory and calibration tell us how much we know about each piece, and what we know decides which tool does which job.

Job 1: build a rule from a predicted average. Suppose we know only that the noise averages to zero. Then $\mathbb{E}\,x_i = a_i\theta + b_i$, and that is already enough to estimate θ, with no statement about the shape of the noise. Add the readings, replace the expected sum by the observed one, and solve:

$$\sum_i \mathbb{E}\,x_i = \theta \sum_i a_i + \sum_i b_i \quad\Longrightarrow\quad \hat\theta_{\text{mom}} = \frac{\sum_i (x_i - b_i)}{\sum_i a_i},$$

which for $a_i = 1$, $b_i = 0$ is the sample mean. This is the method of moments (Section 3.9.2): put the measured counterpart in place of a predicted expectation. Second moments work the same way. If theory predicts a variance, the average of squared deviations estimates it (Section 3.4.1); if it predicts a correlation, $\mathbb{E}[x_i y_i] = \theta$, the average of the products $x_i y_i$ estimates it, and that estimator is *quadratic* in the readings, the starting point of the quadratic estimators of Section T5.6. Probability is still involved, since an expectation is defined by a probability model; the method simply uses less of that model than a likelihood does. Not every simple rule is of this kind. The median, for instance, matches no expectation: it matches a predicted *quantile* (half the probability on each side), or equivalently it minimises the sum of absolute deviations, by the argument that makes the posterior median the answer for absolute loss in Section 3.16.4.

Job 2: build a rule from the whole law of the fluctuations. Suppose we know more: the noise is Gaussian with a known covariance $\mathbf{C}$. Now the probability of the entire data vector can be written, and the likelihood asks which θ makes the observed readings least surprising. With $\boldsymbol{r} = \boldsymbol{x} - \boldsymbol{a}\theta - \boldsymbol{b}$,

$$-2\ln \mathcal{L}(\theta) \overset{(1)}{=} \boldsymbol{r}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{r} + \text{const}, \qquad 0 \overset{(2)}{=} -2\boldsymbol{a}^{\mathsf{T}}\mathbf{C}^{-1}(\boldsymbol{x} - \boldsymbol{a}\hat\theta - \boldsymbol{b}), \qquad \hat\theta_{\text{ML}} \overset{(3)}{=} \frac{\boldsymbol{a}^{\mathsf{T}}\mathbf{C}^{-1}(\boldsymbol{x} - \boldsymbol{b})}{\boldsymbol{a}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{a}}. \tag{3.104}$$

(1) The multivariate Gaussian density of Section 3.10.5; the normalisation does not depend on θ because $\mathbf{C}$ does not.
(2) Set the derivative with respect to θ to zero; $\partial\boldsymbol{r}/\partial\theta = -\boldsymbol{a}$, and $\mathbf{C}^{-1}$ is symmetric.
(3) Solve the linear equation for $\hat\theta$.

Moments and likelihood are two ways of using the same physical relation (3.103), with different amounts of knowledge about the noise. Sometimes they give the same rule (independent, equal-variance readings with $a_i = 1$: both give the mean); sometimes not (unequal variances: the moment rule weights all readings equally, the likelihood does not).

Job 3: choose how to combine the readings. Many combinations of the readings respond to θ. Restrict to linear ones, $\hat\theta = \boldsymbol{w}^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{b})$. Its average is $\boldsymbol{w}^{\mathsf{T}}\boldsymbol{a}\,\theta$, so it is unbiased exactly when $\boldsymbol{w}^{\mathsf{T}}\boldsymbol{a} = 1$; its variance is $\boldsymbol{w}^{\mathsf{T}}\mathbf{C}\boldsymbol{w}$. Which allowed combination fluctuates least?

$$\frac{\partial}{\partial \boldsymbol{w}}\Big[\boldsymbol{w}^{\mathsf{T}}\mathbf{C}\boldsymbol{w} - 2\lambda(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{a} - 1)\Big] \overset{(1)}{=} 2\mathbf{C}\boldsymbol{w} - 2\lambda\boldsymbol{a} = 0, \qquad \boldsymbol{w} \overset{(2)}{=} \lambda\mathbf{C}^{-1}\boldsymbol{a} \overset{(3)}{=} \frac{\mathbf{C}^{-1}\boldsymbol{a}}{\boldsymbol{a}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{a}}. \tag{3.105}$$

(1) Minimise the variance with the unbiasedness condition attached by a Lagrange multiplier λ (the factor 2 is for convenience).
(2) Multiply by $\mathbf{C}^{-1}/2$.
(3) Fix λ from the condition $\boldsymbol{w}^{\mathsf{T}}\boldsymbol{a} = 1$: $\lambda\,\boldsymbol{a}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{a} = 1$.

Inserted into $\boldsymbol{w}^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{b})$ this is exactly (3.104). For independent readings with $a_i = 1$ it is the inverse-variance weighted mean of Section 3.7. Variance optimisation is not a method for the special occasions when noise matters; every method above has to account for the noise. It answers one particular question, which allowed combination scatters least, and under the Gaussian model the likelihood gives the same answer.

Job 4: decide what the rule must keep and what it must remove. The term b_i is not noise; it is a signal we do not want. What to do depends on what we know about it. If b_i is known, subtract

it, as above. If its shape is known but not its size, $b_i = \beta c_i$ with known c_i, either fit β together with θ and profile it away (Section 3.14), or build a combination that cancels it, adding the condition $\boldsymbol{w}^\mathsf{T}\boldsymbol{c} = 0$ to Job 3, the noise-cancelling trick of Section 3.7. And if the unwanted signal responds exactly like the wanted one, $\boldsymbol{c} \propto \boldsymbol{a}$, no combination of these data can tell them apart: the condition $\boldsymbol{w}^\mathsf{T}\boldsymbol{a} = 1$ and $\boldsymbol{w}^\mathsf{T}\boldsymbol{c} = 0$ contradict each other, and only outside information (another frequency, another experiment, a prior) can separate them. These are background subtraction, nuisance parameters and component separation. None of them is what statisticians call bias correction.

Job 5: check what the finished rule does over many repetitions. Even with every contaminant handled correctly, the rule itself may miss on average, $\mathbb{E}_\theta\,\hat\theta \neq \theta$. Estimating a Gaussian variance with the mean also estimated from the same readings divides by N and comes out low by the factor $(N-1)/N$; correcting it gives the divisor $N-1$ (Section 3.4.1). No unwanted signal was present: the bias came from the procedure. Removing a contaminant can remove a source of bias, but the two are different things, and a rule that is free of contaminants still has to be checked for its own bias and scatter, by a formula where one exists and by simulated repetitions where it does not.

So the chain from physics to an answer has five links. What does the theory predict? What do we know about the fluctuations and the contaminants? How can the parameter be extracted? Which extraction rule serves the purpose? And does that rule recover the truth reliably over repeated experiments? Moments and likelihood build a rule; variance optimisation chooses a combination; the treatment of contaminants fixes what must be kept and removed; and bias correction concerns how the finished rule behaves when the experiment is repeated.

3.16.2 Step 1: write the likelihood

Everything starts from the statement that the data are the model plus noise (Section 3.8.1): the noise realisation that the data require has a known probability, and that probability, read as a function of the parameter, is $\mathcal{L}(\theta)$. For independent data the likelihood is the product (3.80). For the four examples, keeping only the factors that depend on the parameter,

$$\begin{aligned}
\text{Gaussian mean:}\quad & \mathcal{L}(\mu) \propto \exp\Big[-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2\Big],\\
\text{Poisson rate:}\quad & \mathcal{L}(\lambda) = \prod_i \frac{e^{-\lambda}\lambda^{X_i}}{X_i!} \propto e^{-n\lambda}\lambda^{K}, \qquad K \equiv \sum_i X_i,\\
\text{exponential lifetime:}\quad & \mathcal{L}(\tau) = \tau^{-n}e^{-T/\tau}, \qquad T \equiv \sum_i t_i,\\
\text{full-sky CMB:}\quad & \mathcal{L}(C_\ell) \propto C_\ell^{-(2\ell+1)/2}\exp\Big[-\frac{(2\ell+1)\widehat{C}_\ell}{2C_\ell}\Big].
\end{aligned} \tag{3.106}$$

The last line is the isotropic Gaussian sky of Section 3.11.5. Recall its ingredients. At multipole ℓ there are $2\ell+1$ independent real numbers $y_{\ell k}$: $a_{\ell 0}$, and $\sqrt{2}\,\mathrm{Re}\,a_{\ell m}$ and $\sqrt{2}\,\mathrm{Im}\,a_{\ell m}$ for $m = 1,\dots,\ell$. (For $m > 0$ the real and imaginary parts of $a_{\ell m}$ each have variance $C_\ell/2$, and $a_{\ell,-m}$ is the complex conjugate of $a_{\ell m}$ up to a sign, so it adds nothing new; the factor $\sqrt{2}$ brings each variance up to C_ℓ.) Each $y_{\ell k}$ has density $(2\pi C_\ell)^{-1/2}e^{-y_{\ell k}^2/2C_\ell}$. Their product over the $2\ell+1$ values of k gives the last line, with $(2\ell+1)\widehat{C}_\ell = \sum_k y_{\ell k}^2 = \sum_m |a_{\ell m}|^2$.

3.16.3 Step 2: what can the data tell us? Sufficient statistics

Look at (3.106). The Gaussian likelihood contains the n readings, yet after expanding the square it depends on them only through $\bar{x}$ (and through a constant). The Poisson likelihood depends on the counts only through their total K, the exponential one only through T, the CMB one only through $\widehat{C}_\ell$. Every data set with the same K gives exactly the same likelihood curve, so no estimate, error bar or confidence interval computed from the likelihood can distinguish them. The remaining information in the n numbers, which ones were large, which came first, says nothing about θ.

Definition. A statistic $U(\boldsymbol{d})$ is **sufficient** for θ if the conditional distribution of the data given the value of U does not involve θ [Wasserman p. 137, Def. 9.32]. Given U, the data set is one of the points on the "surface" $\{\boldsymbol{d}' : U(\boldsymbol{d}') = U(\boldsymbol{d})\}$, and sufficiency says that θ has no influence on which point of the surface we land on: θ only decides how probable the surface is. Statistical mechanics has met this before. In the canonical ensemble $p(\text{microstate} \mid \beta) = e^{-\beta E}/Z(\beta)$, and given the energy E all microstates are equally likely, whatever β is. The energy is therefore a sufficient statistic for the inverse temperature β: a thermometer that reads only E loses nothing.

The coarsest statistic with this property is called **minimal sufficient**: two data sets have the same value of it exactly when the ratio of their likelihoods $p(\boldsymbol{d} \mid \theta)/p(\boldsymbol{d}' \mid \theta)$ does not depend on θ [Wasserman p. 138, Def. 9.35]. In all the examples below the statistic found is minimal, so no further compression is possible without losing information.

The factorisation theorem. Checking the definition directly needs the conditional distribution. The **Neyman–Fisher factorisation theorem** [Wasserman p. 140, Thm. 9.40], [CasellaBerger Thm. 6.2.6] replaces that with a test on the likelihood:

$$U \text{ is sufficient} \iff p(\boldsymbol{d} \mid \theta) = g\big(U(\boldsymbol{d}), \theta\big)\, h(\boldsymbol{d}) \quad \text{for some functions } g, h.$$

All the θ-dependence must pass through U, and whatever else depends on the data (the h) must not know about θ. The proof, for discrete data (sums become integrals for densities), has two halves. Write $\mathcal{A}_u$ for the set of all data sets with $U = u$.

Factorised form implies sufficiency. For a data set $\boldsymbol{d}$ with $U(\boldsymbol{d}) = u$,

$$\mathrm{P}(\boldsymbol{d} \mid U = u, \theta) \overset{(1)}{=} \frac{\mathrm{P}(\boldsymbol{d} \mid \theta)}{\mathrm{P}(U = u \mid \theta)} \overset{(2)}{=} \frac{g(u,\theta)\, h(\boldsymbol{d})}{\sum_{\boldsymbol{d}' \in \mathcal{A}_u} g(u,\theta)\, h(\boldsymbol{d}')} \overset{(3)}{=} \frac{h(\boldsymbol{d})}{\sum_{\boldsymbol{d}' \in \mathcal{A}_u} h(\boldsymbol{d}')}. \tag{3.107}$$

(1) The event $\{\boldsymbol{d}\}$ lies inside $\{U = u\}$, so $\mathrm{P}(\boldsymbol{d}, U = u \mid \theta) = \mathrm{P}(\boldsymbol{d} \mid \theta)$.
(2) Numerator by the factorisation. The denominator is the sum of the probabilities of all data sets in A_u, and $g(U(\boldsymbol{d}'), \theta) = g(u, \theta)$ for each of them.
(3) $g(u, \theta)$ is a common factor and cancels.

The result contains no θ, which is the definition.

Sufficiency implies the factorised form. If $\mathrm{P}(\boldsymbol{d} \mid U = u, \theta) = c(\boldsymbol{d})$ does not depend on θ, then

$$\mathrm{P}(\boldsymbol{d} \mid \theta) = \mathrm{P}(\boldsymbol{d} \mid U = U(\boldsymbol{d}), \theta)\ \mathrm{P}(U = U(\boldsymbol{d}) \mid \theta) = c(\boldsymbol{d})\ q\big(U(\boldsymbol{d}), \theta\big),$$

which is the factorised form with $h = c$ and $g = q$, the sampling distribution of U itself.

Examples. Expanding the square, $\sum_i (x_i - \mu)^2 = \sum_i (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2$, so the Gaussian likelihood factorises with $g = \exp[-n(\bar{x} - \mu)^2/2\sigma^2]$ and $h = \exp[-\sum_i (x_i - \bar{x})^2/2\sigma^2]$: $\bar{x}$ is sufficient for μ. Reading (3.106), the Poisson total K is sufficient for λ with $g = e^{-n\lambda}\lambda^K$ and

$h = \prod_i 1/X_i!$; the sum of decay times T is sufficient for τ with $h = 1$; and $\widehat{C}_\ell$, equivalently $\sum_m |a_{\ell m}|^2$, is sufficient for C_ℓ with $h = 1$. The last statement is the precise form of the everyday remark that for a Gaussian isotropic sky the power spectrum contains everything there is to know about the parameters: the phases of the $a_{\ell m}$ carry no information. It is a statement about the model. If the field is not Gaussian, h is no longer free of C_ℓ and the phases carry information that $\widehat{C}_\ell$ throws away, a point that returns when topological summaries of fields appear (Chapter 12).

The exponential family: sufficient statistics and moment matching. All four examples, and most of the distributions of Chapter 2, have the form

$$p(x \mid \theta) = h(x) \exp\big[\eta(\theta)\, \phi(x) - A(\theta)\big], \tag{3.108}$$

with $\phi(x)$ a single function of one observation (the Poisson has $\phi = x$, $\eta = \ln\lambda$, $A = \lambda$; the exponential lifetime has $\phi = t$, $\eta = -1/\theta$, $A = \ln\theta$ with θ the mean lifetime; the CMB coefficient has $\phi = y^2$, $\eta = -1/2C_\ell$, $A = \tfrac{1}{2}\ln 2\pi C_\ell$). For n independent observations, the product has $\exp[\eta \sum_i \phi(x_i) - nA] \prod_i h(x_i)$, so $\sum_i \phi(x_i)$ is sufficient by the factorisation theorem. More: the MLE is the solution of a moment-matching equation. Differentiating the normalisation $\int p\, \mathrm{d}x = 1$ with respect to θ gives

$$0 = \int (\eta' \phi - A')\, p\, \mathrm{d}x \quad \Longrightarrow \quad \mathbb{E}_\theta\, \phi = \frac{A'(\theta)}{\eta'(\theta)},$$

and then

$$\frac{\partial \ell}{\partial \theta} \overset{(1)}{=} \eta' \sum_i \phi(x_i) - nA' \overset{(2)}{=} \eta'(\theta) \Big[\sum_i \phi(x_i) - n\ \mathbb{E}_\theta\, \phi \Big]. \tag{3.109}$$

(1) Logarithm of the product, then differentiate; h does not contain θ.
(2) Insert $A' = \eta'\ \mathbb{E}_\theta\, \phi$.

With $\eta' \neq 0$ the score equation is

$$\mathbb{E}_{\hat\theta}\ \phi = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i), \tag{3.110}$$

the model's mean of the sufficient statistic equals its observed mean. That is why the maximum-likelihood lifetime is the mean decay time, the Poisson rate estimate is the mean count $\hat\lambda = \bar X$, the CMB estimate is the mean of y^2, and in each case the method of moments (Section 3.9.2) gives the same answer as maximum likelihood. Outside the exponential family, for instance for the angular distribution (3.83), whose parameters also sit in the normalisation, the two methods give different estimates and the moment estimates are generally less precise.

3.16.4 Step 3: constructing candidate estimators

Having found what the data can tell us, we build estimators from it. The families met so far are the following.

- *Method of moments* (Section 3.9.2): set model moments equal to sample moments. No likelihood is needed, only moments of the model. Quick, consistent, often a good starting point for Newton–Raphson.
- *Maximum likelihood* (Section 3.9): maximise $\mathcal{L}(\theta)$. Always available when the likelihood is, invariant under reparametrisation, asymptotically efficient.

- *Least squares* (Section 3.10.1): the maximum likelihood estimator for independent Gaussian noise, $-2\ell = \chi^2$, and for correlated Gaussian noise with the weight matrix $\mathbf{C}^{-1}$. Not a separate principle.
- *Bayes estimators*, which use a prior and a cost of being wrong. They complete the list and are derived below; Chapter 5 develops them.

The EM algorithm: maximum likelihood with hidden variables. Sometimes the likelihood is a sum over quantities we did not observe, such as the unknown class of each event in a mixture of signal and background, $p(\boldsymbol{d} \mid \theta) = \sum_z p(\boldsymbol{d}, z \mid \theta)$, and the logarithm of a sum has no closed-form maximum. Given a current guess θ_t, let $q(z) = p(z \mid \boldsymbol{d}, \theta_t)$ be the posterior probability of each hidden assignment. Since the logarithm is concave, Jensen's inequality gives

$$\ln p(\boldsymbol{d} \mid \theta) \overset{(1)}{=} \ln \sum_z q(z) \frac{p(\boldsymbol{d}, z \mid \theta)}{q(z)} \overset{(2)}{\geq} \sum_z q(z) \ln p(\boldsymbol{d}, z \mid \theta) - \sum_z q(z) \ln q(z), \tag{3.111}$$

(1) Multiply and divide by $q(z)$, which sums to one.
(2) Jensen: the log of an average is at least the average of the log.

with equality at $\theta = \theta_t$. (There $p(\boldsymbol{d}, z \mid \theta_t)/q(z) = p(\boldsymbol{d} \mid \theta_t)$ is the same for every z, and Jensen's inequality is an equality for an average of equal numbers.) The right side differs from $Q(\theta \mid \theta_t) = \sum_z q(z) \ln p(\boldsymbol{d}, z \mid \theta)$ by a constant, and Q involves the logarithm of a single term, usually easy to maximise. The algorithm alternates the **E step** (compute $q(z)$, hence Q) with the **M step** ($\theta_{t+1} = \arg\max_\theta Q$). No cycle can decrease the likelihood, because the right side touches $\ln p(\boldsymbol{d} \mid \theta)$ at θ_t and lies below it elsewhere [Wasserman pp. 143–144]. Raising the right side therefore raises $\ln p$ at least as much. It converges to a stationary point, which can be a local maximum.

Bayes estimators: choose the answer that minimises the expected loss. Suppose we must announce one number a for θ, and being off by $\theta - a$ costs us $L(\theta, a)$. We do not know θ, but after the experiment we have the posterior $p(\theta \mid \boldsymbol{d}) \propto \mathcal{L}(\theta) p(\theta)$, the probability density of the value of θ given the data and a prior $p(\theta)$ (Eq. (3.81)). Announce the a that minimises the posterior expected loss

$$\rho(a) = \int L(\theta, a)\, p(\theta \mid \boldsymbol{d})\, \mathrm{d}\theta.$$

The expectation is over the plausible values of θ at fixed data, not over repeated data sets. Three choices of loss give three familiar summaries.

Squared loss, $L = (\theta - a)^2$. Let $m = \mathbb{E}[\theta \mid \boldsymbol{d}]$ be the posterior mean. Then

$$\rho(a) = \mathbb{E}\big[(\theta - a)^2 \mid \boldsymbol{d}\big] \tag{3.112}$$

$$\overset{(1)}{=} \mathbb{E}\big[(\theta - m)^2 \mid \boldsymbol{d}\big] + 2(m - a)\, \mathbb{E}\big[\theta - m \mid \boldsymbol{d}\big] + (m - a)^2 \tag{3.113}$$

$$\overset{(2)}{=} \mathrm{Var}[\theta \mid \boldsymbol{d}] + (m - a)^2. \tag{3.114}$$

(1) Write $\theta - a = (\theta - m) + (m - a)$ and expand the square. The number $m - a$ is a constant for the expectation.
(2) $\mathbb{E}[\theta - m \mid \boldsymbol{d}] = 0$ by definition of m, which kills the cross term.

The first term does not depend on a, so ρ is smallest at $a = m$: the *posterior mean* minimises the expected squared error.

Absolute loss, $L = |\theta - a|$. With $F(a) = \int_{-\infty}^{a} p\,\mathrm{d}\theta$,

$$\rho(a) = \int_{-\infty}^{a} (a-\theta)\,p\,\mathrm{d}\theta + \int_{a}^{\infty} (\theta - a)\,p\,\mathrm{d}\theta,$$
$$\frac{\mathrm{d}\rho}{\mathrm{d}a} \overset{(1)}{=} \int_{-\infty}^{a} p\,\mathrm{d}\theta - \int_{a}^{\infty} p\,\mathrm{d}\theta = 2F(a) - 1.$$

(1) Differentiate the integrals with respect to a in the integrand and at the limits; the limit terms contain $(a-\theta)p$ at $\theta = a$, which is zero.

The derivative vanishes at $F(a) = \frac{1}{2}$: the *posterior median*.

Zero–one loss, $L = 0$ if $|\theta - a| < h$ and 1 otherwise. Then $\rho(a) = 1 - \int_{a-h}^{a+h} p\,\mathrm{d}\theta \approx 1 - 2h\,p(a \mid \boldsymbol{d})$, smallest where the posterior is largest: the *posterior mode* (MAP, maximum *a posteriori*), obtained as $h \to 0$.

For a symmetric, single-peaked posterior the three coincide. For a skewed one they do not. The Poisson rate with a single observation $K = 2$ and a flat prior has posterior $p(\lambda \mid K) \propto \lambda^2 e^{-\lambda}$, a Gamma density with shape 3. Its mode is 2 (equal to the MLE, because the prior is flat, Section 3.8.4), its median is 2.67 and its mean is 3. Evaluating the expected losses on a grid of announced values (Figure 3.50, middle) reproduces the three minima at 2, 2.677 and 3.005. Which number is "the estimate" depends on what a mistake costs. A flat prior makes the maximum-likelihood estimate the Bayes estimator under zero–one loss, and only under it.

3.16.5 Step 4: choosing the best

Several candidates, several sampling distributions. Two tools rank them. The first is the Cramér–Rao bound (Section 3.13.3), a floor on the variance of an unbiased estimator. If a candidate sits on the floor it cannot be improved, and the search ends. The second tool says how to improve a candidate that is above it.

The Rao–Blackwell theorem. Let $W(\boldsymbol{d})$ be any unbiased estimator of $g(\theta)$ and U a sufficient statistic. Define

$$\tilde{W}(U) = \mathbb{E}[W \mid U],$$

the average of W over all data sets with the observed value of U. Three facts follow.

First, $\tilde{W}$ is a legitimate statistic. The average uses the conditional distribution of the data given U, which by sufficiency does not contain θ. Without sufficiency the average would need the unknown θ and could not be computed.

Second, $\tilde{W}$ is unbiased, by the law of total expectation, $\mathbb{E}\,\tilde{W} = \mathbb{E}[\mathbb{E}(W \mid U)] = \mathbb{E}\,W = g(\theta)$.

Third, its variance is never larger:

$$\operatorname{Var} W \overset{(1)}{=} \mathbb{E}\big[\operatorname{Var}(W \mid U)\big] + \operatorname{Var}\big[\mathbb{E}(W \mid U)\big] \overset{(2)}{=} \mathbb{E}\big[\operatorname{Var}(W \mid U)\big] + \operatorname{Var}\tilde{W} \overset{(3)}{\geq} \operatorname{Var}\tilde{W}. \tag{3.115}$$

(1) The law of total variance: split the scatter of W into the average scatter inside each slice $U = u$ and the scatter of the slice averages between slices.
(2) Definition of $\tilde{W}$.
(3) The first term is an average of variances, each ≥ 0. It vanishes only if W is already a function of U.

[Wasserman p. 140, Thm. 9.42], [CasellaBerger Thm. 7.3.17]. The reduction is exactly the average variance that W wasted on the part of the data that U does not contain. The theorem gives an algorithm: start with any crude unbiased estimator, however silly, and average it over the sufficient statistic.

Example: estimating $e^{-\lambda}$, the chance of an empty slot. In n time slots we count $X_i \sim \text{Poisson}(\lambda)$ events, and we want $g(\lambda) = e^{-\lambda} = \mathrm{P}(X = 0)$, the probability that a slot is empty. A crude unbiased estimator uses the first slot only: $W = \mathbf{1}\{X_1 = 0\}$, since $\mathbb{E}\,W = \mathrm{P}(X_1 = 0) = e^{-\lambda}$. It wastes $n - 1$ slots. Its variance is that of a 0/1 variable, $e^{-\lambda}(1 - e^{-\lambda})$. The total $K = \sum_i X_i$ is sufficient. The conditional mean is the probability that slot 1 is empty given the total $K = s$. Independent Poisson variables add to a Poisson variable, so

$$\mathrm{P}(X_1 = 0 \mid K = s) \overset{(1)}{=} \frac{\mathrm{P}(X_1 = 0)\ \mathrm{P}\big(\sum_{i\ge 2} X_i = s\big)}{\mathrm{P}(K = s)} \overset{(2)}{=} \frac{e^{-\lambda}\, e^{-(n-1)\lambda}\,[(n-1)\lambda]^s/s!}{e^{-n\lambda}\,(n\lambda)^s/s!} \overset{(3)}{=} \Big(\frac{n-1}{n}\Big)^s. \tag{3.116}$$

(1) Definition of conditional probability; $X_1 = 0$ and $K = s$ together mean $X_1 = 0$ and the other $n - 1$ slots add up to s, and the slots are independent.
(2) The sum of the other $n - 1$ slots is Poisson with mean $(n - 1)\lambda$, the total is Poisson with mean $n\lambda$.
(3) The exponentials and factorials cancel, and so does λ^s, leaving $(n - 1)^s/n^s$. This is the last place that λ could have entered.

So $\tilde{W} = (1 - 1/n)^K$. Conditional on the total, each of the K events fell in slot 1 with probability $1/n$, independently, so slot 1 is empty with probability $(1 - 1/n)^K$. The variance is found from the generating function of the Poisson distribution, $\mathbb{E}\,z^K = \exp[n\lambda(z - 1)]$: with $z = 1 - 1/n$ this gives $\mathbb{E}\,\tilde{W} = e^{-\lambda}$ (unbiased, as promised), and with $z^2 = 1 - 2/n + 1/n^2$ it gives $\mathbb{E}\,\tilde{W}^2 = e^{-2\lambda+\lambda/n}$, so

$$\operatorname{Var}\tilde{W} = e^{-2\lambda}\big(e^{\lambda/n} - 1\big).$$

For $n = 10$ and $\lambda = 1.5$ the crude estimator has variance 0.1733 and the Rao–Blackwellised one 0.008057, a reduction by a factor of more than twenty. The simulation of 20 000 experiments (Section 3.16.8) gives 0.1734 and 0.00799, both means at 0.2232 and 0.2231 against the truth 0.2231. The Cramér–Rao floor for estimating $g(\lambda)$ is $g'(\lambda)^2/I_n = \lambda e^{-2\lambda}/n = 0.007468$ with $I_n = n/\lambda$. The improved estimator sits just above it. Expanding, $\operatorname{Var}\tilde{W} \approx (\lambda e^{-2\lambda}/n)(1 + \lambda/2n)$, so the excess over the floor is a factor $1 + \lambda/2n$, which tends to one as n grows.

Best unbiased estimators. Can averaging a second time, over a different crude starting estimator, improve it again? For the Poisson total, no. The total K is **complete** [the only function of K whose mean is zero for every λ is the zero function]. The Lehmann–Scheffé theorem [CasellaBerger Thm. 7.3.23] says that for a complete sufficient statistic U, any unbiased function of U is the unique minimum-variance unbiased estimator (UMVUE). Two unbiased functions of U could be different only by a function of U with zero mean, and there is none. The exponential family (3.108) with a rich enough parameter has complete statistics, so $\tilde{W}$ above is the best unbiased estimator of $e^{-\lambda}$, even though it does not reach the Cramér–Rao floor. The bound is attained only when the score is a linear function of the estimator, which for $e^{-\lambda}$ it is not.

When a biased estimator beats an unbiased one. UMVUEs are the best among unbiased estimators. The mean squared error is bias squared plus variance (Section 3.2.1), and a small bias that buys a larger drop in variance can win. The maximum-likelihood estimate of $e^{-\lambda}$ is $e^{-\hat{\lambda}} = e^{-K/n}$. It is biased, but by $O(1/n)$, so its squared bias is $O(1/n^2)$ and the variance is $O(1/n)$: the MLE and $\tilde{W}$ have nearly the same variance, and which has the lower MSE depends on n and λ. The earlier example with the Gaussian variance went further: the divisor $N + 1$ beats both $N - 1$ and N (Section 3.4.2). The posterior mean with a prior centred away from the data is another example: it shrinks the estimate toward the prior, adds bias and removes variance (Chapter 5). Unbiasedness is a convenient property, not a goal. The MLE's claim to

be good is asymptotic efficiency, and the Bayes and shrinkage estimates add to it when prior knowledge exists.

3.16.6 Step 5: when the likelihood cannot be written down

The procedure needs $p(\boldsymbol{d} \mid \theta)$. For a masked, noisy, beamed CMB map the field is still Gaussian, but the pixels are correlated, and the likelihood with a huge covariance matrix is too expensive to evaluate at every point of a Markov chain. For the large-scale-structure correlation function, the number of independent modes is far larger than the number of simulations available to estimate the covariance. For a topological summary of a field (Chapter 12) there is no likelihood at all. The procedure still works if we give up on optimality and compute a simple summary statistic, then replace the missing likelihood by simulation:

1. choose a statistic $\hat{S}(\boldsymbol{d})$ that is cheap and sensitive to θ (a pseudo-C_ℓ, a counts-in-cells histogram);
2. simulate many data sets from the model at several θ, through the same mask, noise and beam, and compute $\hat{S}$ for each (Chapter 6);
3. read off the mean $\langle \hat{S} \rangle_\theta$ (the bias and the transfer function of the estimator) and the covariance of $\hat{S}$ from the scatter over the simulations, with the Hartlap correction of Section 3.5.4;
4. use a Gaussian likelihood for $\hat{S}$ with that mean and covariance, or quote the parameter at which $\langle \hat{S} \rangle_\theta$ matches the observed $\hat{S}$, with an error bar from the simulated spread.

This works even if $\hat{S}$ is not sufficient: its bias is calibrated out and its error bar is honest, but the error bar is wider than the Cramér–Rao floor of the full data. A mask removing a fraction of the sky, for example, inflates the variance of the pseudo-spectrum roughly by one over the observed sky fraction and correlates neighbouring multipoles; the simulations measure both exactly. The same machinery, used for verification where a formula exists and as the research instrument where none does, is developed in Chapters 6 and 8.

3.16.7 The full chain on one example: the CMB power spectrum

Take a full-sky, noiseless, Gaussian isotropic temperature map, with the coefficients $a_{\ell m}$ of variance C_ℓ (the power spectrum of the theory), and ask for the best estimate of C_ℓ at one multipole. Write $\nu = 2\ell + 1$.

Step 1: likelihood. From (3.106),

$$\ln \mathcal{L}(C_\ell) = -\frac{\nu}{2}\left[\frac{\widehat{C}_\ell}{C_\ell} + \ln C_\ell\right] + \text{const}, \qquad \widehat{C}_\ell = \frac{1}{\nu}\sum_{m=-\ell}^{\ell} |a_{\ell m}|^2. \tag{3.117}$$

Step 2: sufficiency. $\widehat{C}_\ell$ enters alone, and the factorisation is complete with $h = 1$. Nothing about the phases or the sign of m is needed.

Step 3: the candidate. The score equation:

$$\frac{\partial \ln \mathcal{L}}{\partial C_\ell} \overset{(1)}{=} -\frac{\nu}{2C_\ell} + \frac{\nu \widehat{C}_\ell}{2C_\ell^2} \overset{(2)}{=} \frac{\nu}{2C_\ell^2}\left(\widehat{C}_\ell - C_\ell\right) = 0 \quad \Longrightarrow \quad \hat{C}_\ell^{\mathrm{ML}} = \widehat{C}_\ell. \tag{3.118}$$

(1) Differentiate (3.117): $\partial_C(\widehat{C}_\ell/C) = -\widehat{C}_\ell/C^2$ and $\partial_C \ln C = 1/C$.
(2) Common denominator $2C_\ell^2$.

The second derivative is $\nu/2C_\ell^2 - \nu\widehat{C}_\ell/C_\ell^3$. At the solution $C_\ell = \widehat{C}_\ell$ it equals $\nu/2\widehat{C}_\ell^2 - \nu/\widehat{C}_\ell^2 = -\nu/2\widehat{C}_\ell^2 < 0$, so this is a maximum. The familiar estimator, the average of $|a_{\ell m}|^2$ over the $2\ell+1$ values of m, is the maximum-likelihood estimator, the moment-matching solution (3.110) with $\phi = y^2$, and (with flat prior) the posterior mode.

Step 4: is it the best? Bias: $\mathbb{E}\, y^2 = C_\ell$ for each of the ν variables, so $\mathbb{E}\,\widehat{C}_\ell = C_\ell$, unbiased. Variance: $\nu\widehat{C}_\ell/C_\ell = \sum_k (y_k/\sqrt{C_\ell})^2$ is a sum of ν squares of unit Gaussians, which is χ^2_ν with mean ν and variance 2ν, so

$$\operatorname{Var}\widehat{C}_\ell = \frac{C_\ell^2}{\nu^2}\,2\nu = \frac{2C_\ell^2}{2\ell+1}. \tag{3.119}$$

Fisher information, the negative mean of the second derivative at the truth:

$$I(C_\ell) \overset{(1)}{=} -\mathbb{E}\Big[\frac{\nu}{2C_\ell^2} - \frac{\nu\widehat{C}_\ell}{C_\ell^3}\Big] \overset{(2)}{=} -\frac{\nu}{2C_\ell^2} + \frac{\nu}{C_\ell^2} = \frac{2\ell+1}{2C_\ell^2}. \tag{3.120}$$

(1) The identity $I = -\mathbb{E}\,\partial^2 \ln \mathcal{L}/\partial\theta^2$ (3.98).
(2) Only $\widehat{C}_\ell$ is random, and $\mathbb{E}\,\widehat{C}_\ell = C_\ell$.

The Cramér–Rao bound is $1/I = 2C_\ell^2/(2\ell+1)$, which is exactly (3.119). The estimator attains the bound: the score (3.118) is proportional to $\widehat{C}_\ell - C_\ell$, the exact condition for equality in Cauchy–Schwarz in the derivation of (3.101). No unbiased estimator of C_ℓ built from one sky can have a smaller variance. This number is **cosmic variance**: a fractional error of $\sqrt{2/(2\ell+1)}$, which is 63% at the quadrupole and 18% at $\ell = 30$, caused by nothing but the fact that we have one sky, with $2\ell+1$ independent numbers at each ℓ. No instrument improves it. In Section 3.4.3 the same number appeared as the variance of a variance. Here it is the Cramér–Rao bound of the model.

3.16.8 The recipe in Python

Listing 3.15: Rao–Blackwell for the empty-slot probability (`code/ch03/16_estimator_recipe.py`)

```
# ---------- (i) Rao-Blackwell: estimate g(lambda) = exp(-lambda) from n Poisson counts ----------
lam, n, M = 1.5, 10, 20000
X = rng.poisson(lam, size=(M, n))
T_crude = (X[:, 0] == 0).astype(float)            # unbiased: E 1{X_1 = 0} = exp(-lambda)
S = X.sum(axis=1)                                  # sufficient statistic
T_rb = ((n - 1) / n) ** S                          # E[T_crude | S] = ((n-1)/n)^S
g = np.exp(-lam)
out["RBTruth"] = g
out["RBCrudeMean"], out["RBCrudeVar"] = T_crude.mean(), T_crude.var()
out["RBRBMean"], out["RBRBVar"] = T_rb.mean(), T_rb.var()
out["RBRatio"] = T_rb.var() / T_crude.var()
out["RBCR"] = lam * np.exp(-2 * lam) / n           # Cramer-Rao bound for g: g'^2 / I_n
out["RBCrudeVarTh"] = g * (1 - g)                  # variance of a 0/1 variable
out["RBRBVarTh"] = np.exp(-2 * lam) * (np.exp(lam / n) - 1)
out["RBn"], out["RBlam"] = n, lam
```

The array `X` holds 20 000 experiments of $n = 10$ Poisson slots each. The crude estimator looks only at the first slot, and the sufficient statistic `S` is the row sum (the K of the text). The

Rao–Blackwellised estimator is the function $((n-1)/n)^K$ of (3.116). The printed variances are the numbers quoted above: 0.1734 against 0.00799, a ratio of 0.04608, and both means agree with $e^{-\lambda} = 0.2231$.

Listing 3.16: Expected loss of three point estimates under a skewed posterior (code/ch03/16_estimator_recipe.py)

```
# ---------- (ii) Bayes estimators: minimise the expected loss under a skewed posterior ----------
post = rng.gamma(3.0, 1.0, size=400000)            # a posterior sample, Gamma(3,1)
grid = np.linspace(0.05, 8, 800)
sq = np.array([np.mean((post - a) ** 2) for a in grid])
ab = np.array([np.mean(np.abs(post - a)) for a in grid])
from scipy.stats import gamma as gamma_dist
h = 0.05                                           # 0-1 loss with a tolerance h: P(|theta-a| > h),
z1 = 1 - 2 * h * gamma_dist.pdf(grid, 3.0)         # evaluated with the exact posterior density
out["BayesSqArg"], out["BayesAbsArg"], out["BayesZOArg"] = grid[sq.argmin()], grid[ab.argmin()], grid[z1.argmin()]
out["BayesMean"], out["BayesMedian"], out["BayesMode"] = post.mean(), np.median(post), 2.0
```

A posterior sample of 400 000 points from the Gamma density is used to evaluate $\rho(a)$ on a grid, for squared and absolute loss. For the zero–one loss the exact density is used in $1 - 2h\,p(a)$. The grids have their minima at 3.005 (mean 3), 2.677 (median 2.672) and 2 (mode 2).

Listing 3.17: A full-sky Gaussian CMB: the maximum-likelihood C_ℓ and its variance (code/ch03/16_estimator_recipe.py)

```
# ---------- (iii) full-sky C_ell: a_lm Gaussian, variance C ----------
def sky(ell, C=1.0, M=20000):
    a = rng.normal(0, np.sqrt(C), size=(M, 2 * ell + 1))    # 2l+1 independent real degrees of freedom
    return (a ** 2).sum(axis=1)                             # sum_m |a_lm|^2
res = {}
for ell in (2, 30):
    Ssum = sky(ell)
    Chat = Ssum / (2 * ell + 1)
    w = {2: "Two", 30: "Thirty"}[ell]
    out[f"CMBMean{w}"], out[f"CMBVar{w}"] = Chat.mean(), Chat.var()
    out[f"CMBBound{w}"] = 2 / (2 * ell + 1)
    out[f"CMBRelErr{w}"] = Chat.std()
    res[ell] = Chat
# Fisher information from the score at the truth: S = -(2l+1)/(2C) + sum|a|^2/(2C^2), Var S = (2l+1)/(2C^2)
ell = 30; Ssum = sky(ell); score = -(2 * ell + 1) / 2 + Ssum / 2
out["CMBScoreVarThirty"], out["CMBFishThirty"] = score.var(), (2 * ell + 1) / 2
out["CMBScoreMeanThirty"] = score.mean()
```

For $\ell = 2$ and $\ell = 30$, 20 000 simulated skies supply $\sum_m |a_{\ell m}|^2$; dividing by $2\ell+1$ gives $\widehat{C}_\ell$ for $C = 1$. The means are 0.9962 and 1.002 (unbiased), the variances 0.401 and 0.03236 against the bounds $2/(2\ell+1) = 0.4$ and 0.03279, so the simulated estimator sits on the Cramér–Rao bound at both multipoles within sampling noise. The fractional errors are 0.6332 and 0.1799. The last lines evaluate the score (3.118) at the true C: its mean is 0.008949 (zero, (3.97)) and its variance is 30.16 against the Fisher information $\nu/2 = 30.5$ at $\ell = 30$, as (3.98) requires.

3.17 A worked problem: finding the estimator of the Hubble constant

A white dwarf in a binary system accretes matter from its companion until it approaches the Chandrasekhar mass, and then it explodes. The explosion, a type Ia supernova, is nearly the

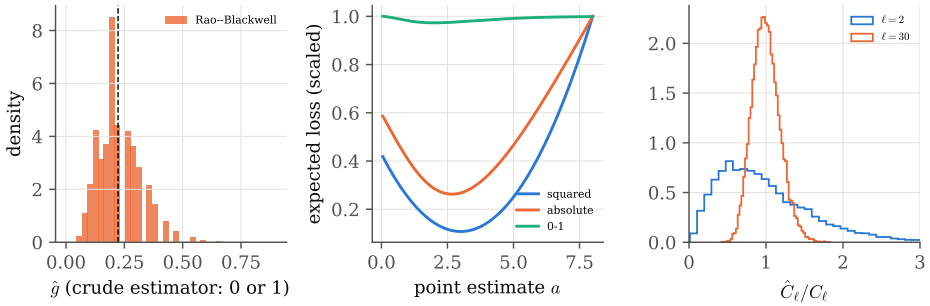


Figure 3.50: Three outcomes of the recipe. Left: the Rao–Blackwellised estimate of $e^{-\lambda}$ over 20 000 experiments; the crude estimator can only answer 0 or 1, while the averaged one is tightly peaked on the truth (dashed). Middle: expected loss for a skewed posterior as a function of the announced value, for squared, absolute and zero–one loss. Their minima are at the mean, the median and the mode, from right to left. Right: the maximum-likelihood $\widehat{C}_\ell / C_\ell$ from 20 000 skies at $\ell = 2$ and $\ell = 30$, centred on the truth with widths set by cosmic variance.

Taxonomy: Which estimator when

situation	estimator	why
exponential family (Poisson, exponential, Gaussian, χ^2_ν-type)	maximum likelihood = match the mean of the sufficient statistic	sufficient, efficient, on the Cramér–Rao floor for the mean parameter
Gaussian noise, known $\mathbf{C}$	least squares (weights $\mathbf{C}^{-1}$)	it is the ML estimator; exact for linear models
counts in bins	Poisson ML $(2\sum[\mu - n\ln\mu])$, not $\sqrt{n}$-weighted χ^2	no bias on the normalisation, valid for small counts
Gaussian isotropic sky	$\widehat{C}_\ell = \frac{1}{2\ell+1}\sum_m \lvert a_{\ell m}\rvert^2$	ML, unbiased, attains cosmic variance
a crude unbiased estimator exists	average it over a sufficient statistic	never increases the variance (Rao–Blackwell)
prior information, or a cost of error	Bayes: posterior mean, median or mode	minimises the expected squared, absolute, 0–1 loss
intractable likelihood, masks, non-Gaussian summaries	simple statistic, bias and covariance calibrated by simulation	honest error bar, not optimal

same every time: after a correction for how fast its light curve declines, its peak luminosity varies from one supernova to the next by only about 15%. That makes it a **standard candle** [an object whose true brightness is known, so its apparent brightness tells us its distance]. A spectrum of the host galaxy gives its redshift z to a few parts in 10^5. The peak apparent brightness, measured as a magnitude m, is known far less well: the intrinsic spread and the photometry together scatter it by about 0.15 magnitudes.

A survey collects N pairs (z_i, m_i). Nearby galaxies recede with velocities proportional to their distances, and the constant of proportionality is the Hubble constant H_0. So the pairs contain H_0. The obvious way to get it out is the one Hubble used: turn each magnitude into a distance, plot velocity against distance, and fit a straight line through the origin. That obvious estimator is wrong. For a survey of 100 supernovae its slope is too low by about half a km/s/Mpc, as large as its own error bar, and the offset does not shrink when the survey grows to 400 or 4000 supernovae. A different function of the same pairs has no offset and the smallest error bar any estimator can have. It comes out of the five-step procedure

of Section 3.16: write the likelihood, find the sufficient statistic, build candidates, compare them with the Cramér–Rao bound, and simulate when formulas run out. The same constant returns several times later, each time reached by a different route: the distance ladder that supplies the supernova luminosity these mock surveys take as known (Section T2.12) and its disagreement with the CMB (Section T2.12.3), the baryon acoustic ruler read backwards as an inverse ladder (Section T2.15), a CMB chain reweighted by the ladder value (Section 9.23.1), the CMB forecast in which H_0 is a derived parameter (Section 10.6.3), cosmic chronometers that need no calibration at all (Sections 3.17.6 and 11.19), and the peculiar velocities of the local flow (Sections 11.16 and 11.17); Section 11.20 brings them together.

The question. Given redshifts z_i and apparent magnitudes m_i of N supernovae,

1. where does the noise live, and what is the likelihood?
2. which function of the data is an unbiased estimator of H_0 with the smallest scatter, and what is its error bar?
3. what can supernovae alone *not* tell us about H_0?
4. what changes when the supernovae are far enough away that the Hubble law is no longer a straight line?

3.17.1 Step 1: the model, and where the noise lives

Two standard results of cosmology are the physical inputs. The first is the **Hubble law**: at low redshift the recession velocity cz of a galaxy is proportional to its distance d,

$$cz = H_0\, d.$$

A galaxy also moves through space with a **peculiar velocity** [its velocity relative to the smooth expansion, caused by the pull of nearby structure]. Along the line of sight this adds a random velocity v_i to the recession velocity. Its spread is $\sigma_v \approx 300$ km/s, so that for galaxy i

$$cz_i = H_0\, d_i + v_i, \qquad v_i \sim \mathcal{N}(0, \sigma_v^2). \tag{3.121}$$

The second piece is the **distance modulus**: the apparent magnitude m and the absolute magnitude M (the magnitude the object would have at 10 pc) differ by

$$\mu \equiv m - M = 5\log_{10}\frac{D_L}{10\,\mathrm{pc}} = 5\log_{10}\frac{D_L}{\mathrm{Mpc}} + 25,$$

where D_L is the luminosity distance [the distance for which the observed flux is $L/4\pi D_L^2$]. At low redshift $D_L = d$. The +25 is $5\log_{10}(10^6\,\mathrm{pc}/10\,\mathrm{pc})$, the price of measuring distances in Mpc. Magnitudes are logarithmic and run backwards: 5 magnitudes fainter means 100 times less flux, and a larger m means a larger distance. From now on d is in Mpc, velocities are in km/s and H_0 is in km/s/Mpc.

The noise is Gaussian in the magnitude. The luminosity of a supernova scatters by a fixed *fraction* from one explosion to the next, and the photometry measures fluxes with a fixed fractional error. On a logarithmic scale a fixed fraction is a fixed additive amount, so the scatter is an additive Gaussian noise in the magnitude:

$$m_i = M + 5\log_{10} d_i + 25 + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma_{\mathrm{int}}^2), \tag{3.122}$$

with $\sigma_{\text{int}} = 0.15$ mag, the combined intrinsic and measurement scatter. The noise ϵ_i is independent of the peculiar velocity v_i. The distance "measured" from the magnitude is

$$\tilde{d}_i \equiv 10^{(m_i - M - 25)/5} = d_i \, 10^{\epsilon_i/5} = d_i \, e^{\kappa \epsilon_i}, \qquad \kappa \equiv \frac{\ln 10}{5} \approx 0.4605. \tag{3.123}$$

A Gaussian in the exponent makes $\tilde{d}_i$ **log-normal** [its logarithm, not itself, is Gaussian]. Its median is the true d_i, since ϵ_i is positive or negative with equal probability, but its mean is not. All the moments follow from one Gaussian integral, which we shall use repeatedly. For $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and any real t,

$$\mathbb{E}\, e^{t\epsilon} = \int_{-\infty}^{\infty} \frac{\mathrm{d}\epsilon}{\sqrt{2\pi}\sigma} \, e^{t\epsilon - \epsilon^2/2\sigma^2} \overset{(1)}{=} \int_{-\infty}^{\infty} \frac{\mathrm{d}\epsilon}{\sqrt{2\pi}\sigma} \, e^{-(\epsilon - t\sigma^2)^2/2\sigma^2 + t^2\sigma^2/2} \overset{(2)}{=} e^{t^2\sigma^2/2}. \tag{3.124}$$

(1) Complete the square: $t\epsilon - \epsilon^2/2\sigma^2 = -(\epsilon - t\sigma^2)^2/2\sigma^2 + t^2\sigma^2/2$.
(2) The remaining integrand is a normalised Gaussian centred on $t\sigma^2$, which integrates to one.

With $t = \kappa$ and $t = 2\kappa$,

$$\mathbb{E}\, \tilde{d}_i = d_i \, e^{\kappa^2 \sigma_{\text{int}}^2/2}, \qquad \mathbb{E}\, \tilde{d}_i^{\,2} = d_i^2 \, e^{2\kappa^2 \sigma_{\text{int}}^2}, \qquad \mathbb{E}\, \tilde{d}_i^{\,-1} = d_i^{-1} \, e^{\kappa^2 \sigma_{\text{int}}^2/2}. \tag{3.125}$$

For $\sigma_{\text{int}} = 0.15$ mag, $\kappa^2 \sigma_{\text{int}}^2 = 0.004772$: the average distance built from magnitudes is too large by a quarter of a percent, and the average of its square by one percent. Both the distance and its inverse are biased *upwards*. These small factors are the whole story of the first two candidates below.

The likelihood, in the variable where the noise lives. The redshifts are measured almost exactly, so we condition on them and let the noise sit in m_i. Write $x_i \equiv v_i/cz_i$ for the fractional peculiar velocity. Solving (3.121) for d_i and putting it into (3.122),

$$\begin{aligned} m_i &\overset{(1)}{=} M + 25 + 5\log_{10} \frac{cz_i(1 - x_i)}{H_0} + \epsilon_i \\ &\overset{(2)}{=} M + 25 + 5\log_{10}(cz_i) - 5\log_{10} H_0 + \frac{5}{\ln 10} \ln(1 - x_i) + \epsilon_i \\ &\overset{(3)}{=} M + 25 + 5\log_{10}(cz_i) - a - \frac{5}{\ln 10} \, x_i + \epsilon_i + O(x_i^2), \qquad a \equiv 5\log_{10} H_0. \end{aligned} \tag{3.126}$$

(1) $d_i = (cz_i - v_i)/H_0 = cz_i(1 - x_i)/H_0$.
(2) The logarithm of a product is a sum of logarithms, and $\log_{10} y = \ln y / \ln 10$.
(3) $\ln(1 - x) = -x + O(x^2)$. Even at $z = 0.01$, $cz = 3000$ km/s and $\sigma_v/cz = 0.1$, so the neglected term is small.

The parameter H_0 enters only through $a = 5\log_{10} H_0$, and it enters *additively*: in the variable where the noise is Gaussian, the model is a constant shift. The two noise terms are independent Gaussians, so their sum is Gaussian with the sum of the variances. This is linear error propagation (Section 3.6.1) of the velocity v_i through the factor $\partial m_i/\partial v_i = -(5/\ln 10)/cz_i$:

$$\sigma_i^2 = \sigma_{\text{int}}^2 + \left(\frac{5}{\ln 10} \, \frac{\sigma_v}{cz_i} \right)^2. \tag{3.127}$$

The peculiar-velocity term is 0.2173 mag at $z = 0.01$, larger than the intrinsic scatter, and 0.02173 mag at $z = 0.1$, where it is negligible. The very nearest supernovae are the least useful for H_0, because their own motions are a large fraction of their recession velocities. The data are

the model plus noise, as in (3.75), and the likelihood is the probability of the noise realisation $m_i - \text{prediction}_i$ that the data require:

$$\ln \mathcal{L}(a, M) = -\frac{1}{2} \sum_{i=1}^{N} \frac{\left[m_i - M - 25 - 5 \log_{10}(cz_i) + a\right]^2}{\sigma_i^2} + \text{const.} \tag{3.128}$$

Figure 3.51 shows one simulated survey of 100 supernovae with $0.01 < z < 0.1$, drawn from $H_0 = 70$ and $M = -19.25$ in a universe where (3.121) is exact. In the velocity–distance plane the points fan out as the distance grows, because a fixed magnitude error is a fixed *fractional* distance error. In the magnitude residuals the scatter is the same at all redshifts except the nearest, where the peculiar velocities widen the band (3.127).

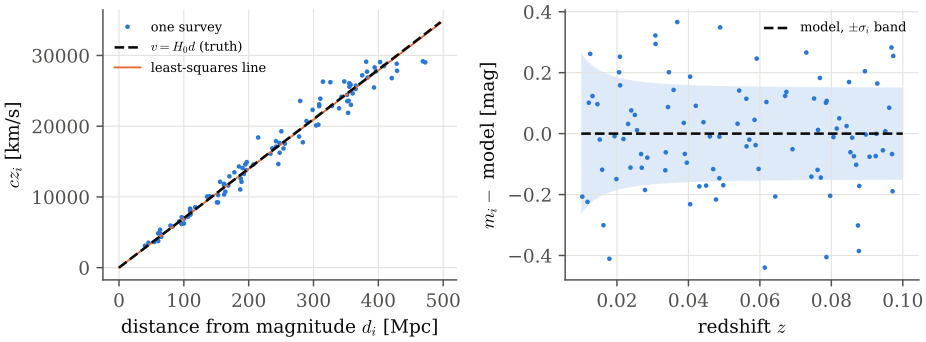


Figure 3.51: One mock survey of 100 supernovae. Left: recession velocity against the distance inferred from each magnitude, with the true Hubble line (dashed) and the least-squares line through the origin, which lies slightly below it. The scatter grows in proportion to the distance. Right: the magnitudes minus the model prediction at the true parameters. The noise is Gaussian in this variable, with a width (shaded, $\pm\sigma_i$) that is constant except at the lowest redshifts, where peculiar velocities add to it.

3.17.2 Steps 2 and 3: the sufficient statistic and three candidates

For the moment, pretend that the absolute magnitude M of a type Ia supernova is known exactly. (It is not, and Section 3.17.4 shows that this changes everything about the error bar.) Each supernova then gives its own estimate of $a = 5 \log_{10} H_0$,

$$y_i \equiv M + 25 + 5 \log_{10}(cz_i) - m_i, \qquad \mathbb{E}\, y_i = a, \quad \operatorname{Var} y_i = \sigma_i^2, \tag{3.129}$$

by reading (3.126) backwards. The likelihood (3.128) is $\ln \mathcal{L} = -\frac{1}{2} \sum_i w_i (y_i - a)^2$ with the weights $w_i \equiv 1/\sigma_i^2$.

Step 2: the sufficient statistic. Let $W \equiv \sum_i w_i$ and $\bar{y}_w \equiv \sum_i w_i y_i / W$, the inverse-variance weighted mean of the y_i. Then

$$\begin{aligned} \sum_i w_i (y_i - a)^2 &\overset{(1)}{=} \sum_i w_i \left[(y_i - \bar{y}_w) + (\bar{y}_w - a) \right]^2 \\ &\overset{(2)}{=} \sum_i w_i (y_i - \bar{y}_w)^2 + 2(\bar{y}_w - a) \sum_i w_i (y_i - \bar{y}_w) + W(\bar{y}_w - a)^2 \\ &\overset{(3)}{=} \sum_i w_i (y_i - \bar{y}_w)^2 + W(\bar{y}_w - a)^2. \end{aligned} \tag{3.130}$$

(1) Add and subtract $\bar{y}_w$.
(2) Expand the square; $\bar{y}_w - a$ does not depend on i.
(3) $\sum_i w_i(y_i - \bar{y}_w) = W\bar{y}_w - W\bar{y}_w = 0$.

So $\mathcal{L}(a) = \exp[-W(\bar{y}_w - a)^2/2] \times \exp[-\sum_i w_i(y_i - \bar{y}_w)^2/2]$. The first factor is the $g(U, a)$ of the factorisation theorem (Section 3.16.3) with $U = \bar{y}_w$, and the second is an $h(\boldsymbol{d})$ that knows nothing about a. The weighted mean $\bar{y}_w$ is sufficient for a, hence for $H_0 = 10^{a/5}$. Every estimator that is not a function of $\bar{y}_w$ uses part of the data that carries no information about H_0, and pays for it in variance or bias.

Candidate (a): fit the Hubble line to distances. Put $v_i \equiv cz_i$ on the vertical axis and the magnitude distances $\tilde{d}_i$ of (3.123) on the horizontal one, and minimise $\sum_i (v_i - H\tilde{d}_i)^2$. Setting the derivative $-2\sum_i \tilde{d}_i(v_i - H\tilde{d}_i)$ to zero,

$$\hat{H}_{\text{naive}} = \frac{\sum_i v_i \tilde{d}_i}{\sum_i \tilde{d}_i^2}. \tag{3.131}$$

This is least squares, but for a model in which the noise sits in v and the horizontal coordinate is exact. Here the horizontal coordinate carries the larger noise. For a large survey the numerator and the denominator, divided by N, each converge to their means by the law of large numbers, so $\hat{H}_{\text{naive}}$ converges to the ratio of the means. With $d_i = cz_i(1 - x_i)/H_0$ and ϵ_i independent of x_i,

$$\begin{aligned}
\mathbb{E}\big[v_i\tilde{d}_i\big] &\overset{(1)}{=} cz_i\ \mathbb{E}[d_i]\ \mathbb{E}\big[e^{\kappa\epsilon_i}\big] \overset{(2)}{=} \frac{(cz_i)^2}{H_0}\, e^{\kappa^2\sigma_{\text{int}}^2/2}, \\
\mathbb{E}\big[\tilde{d}_i^2\big] &\overset{(3)}{=} \mathbb{E}[d_i^2]\ \mathbb{E}\big[e^{2\kappa\epsilon_i}\big] \overset{(4)}{=} \frac{(cz_i)^2 + \sigma_v^2}{H_0^2}\, e^{2\kappa^2\sigma_{\text{int}}^2}, \\
\hat{H}_{\text{naive}} &\xrightarrow{N\to\infty} \frac{\sum_i \mathbb{E}[v_i\tilde{d}_i]}{\sum_i \mathbb{E}[\tilde{d}_i^2]} \overset{(5)}{=} H_0\, e^{-3\kappa^2\sigma_{\text{int}}^2/2}\, \frac{\sum_i (cz_i)^2}{\sum_i \big[(cz_i)^2 + \sigma_v^2\big]}.
\end{aligned} \tag{3.132}$$

(1) $v_i = cz_i$ is fixed; $\tilde{d}_i = d_i e^{\kappa\epsilon_i}$ and the two factors are independent.
(2) $\mathbb{E}\, d_i = cz_i/H_0$ since $\mathbb{E}\, x_i = 0$, and (3.124) with $t = \kappa$.
(3) As in (1), with $\tilde{d}_i^2 = d_i^2 e^{2\kappa\epsilon_i}$.
(4) $\mathbb{E}\, d_i^2 = (cz_i)^2\, \mathbb{E}(1 - x_i)^2/H_0^2 = [(cz_i)^2 + \sigma_v^2]/H_0^2$, and (3.124) with $t = 2\kappa$.
(5) Divide; $e^{\kappa^2\sigma^2/2}/e^{2\kappa^2\sigma^2} = e^{-3\kappa^2\sigma^2/2}$.

The limit is not H_0. The dominant factor, $e^{-3\kappa^2\sigma_{\text{int}}^2/2}$, gives a bias of -0.4992 km/s/Mpc for $\sigma_{\text{int}} = 0.15$; the peculiar velocities lower it a little further, to -0.5181 for the redshift range of the survey. The mechanism is the textbook one for noise in the horizontal variable ("regression dilution"): the scatter of $\tilde{d}_i$ inflates $\sum \tilde{d}_i^2$ more than it inflates $\sum v_i\tilde{d}_i$, and the slope is pulled down. Because the limit (3.132) does not involve N, the estimator is *inconsistent*: more supernovae make it converge faster to the wrong number. Only the $O(1/N)$ corrections to the ratio of means depend on the survey size.

Candidate (b): average the individual ratios. Each supernova gives its own H_0 as $v_i/\tilde{d}_i$, and we average them: $\hat{H}_{\text{ratio}} = \frac{1}{N}\sum_i v_i/\tilde{d}_i$. Each term is

$$\frac{v_i}{\tilde{d}_i} = \frac{cz_i\, e^{-\kappa\epsilon_i}}{d_i} = H_0\, \frac{e^{-\kappa\epsilon_i}}{1 - x_i},$$

and its mean follows from (3.124) with $t = -\kappa$ and $1/(1-x) = 1 + x + x^2 + \dots$, so that $\mathbb{E}(1-x_i)^{-1} \approx 1 + \sigma_v^2/(cz_i)^2$:

$$\mathbb{E}\,\hat{H}_{\text{ratio}} \approx H_0\, e^{\kappa^2\sigma_{\text{int}}^2/2}\,\frac{1}{N}\sum_i\Big[1 + \frac{\sigma_v^2}{(cz_i)^2}\Big]. \tag{3.133}$$

The bias now has the opposite sign, 0.1672 km/s/Mpc from the magnitude scatter and 0.2376 in total, because the nearest supernovae have large x_i. It is again independent of N. The mean of ratios is not the ratio of means, and averaging a nonlinear function of noisy numbers averages its bias too.

Candidate (c): maximum likelihood. The score of (3.128) is $\partial \ln\mathcal{L}/\partial a = \sum_i w_i(y_i - a) = W(\bar{y}_w - a)$, which vanishes at

$$\hat{a} = \bar{y}_w = \frac{\sum_i w_i\big[M + 25 + 5\log_{10}(cz_i) - m_i\big]}{\sum_i w_i}, \qquad w_i = \frac{1}{\sigma_i^2}, \qquad \hat{H}_0 = 10^{\hat{a}/5}. \tag{3.134}$$

It is the weighted mean of Section 3.7, applied to the individual estimates y_i of a in the variable where their noise is Gaussian, and it is a function of the sufficient statistic only.

Bias. $\mathbb{E}\,\hat{a} = \sum_i w_i\,\mathbb{E}\,y_i/W = a$: unbiased, to the order of (3.126). The neglected term, $-(5/\ln 10)\ln(1-x_i) = (5/\ln 10)(x_i + x_i^2/2 + \dots)$, has mean $(5/\ln 10)\,\sigma_v^2/2(cz_i)^2$, which after weighting is 7.927×10^{-4} mag, about a twentieth of the error bar.

Variance and efficiency. The y_i are independent, so

$$\operatorname{Var}\hat{a} \overset{(1)}{=} \frac{\sum_i w_i^2\,\operatorname{Var} y_i}{W^2} \overset{(2)}{=} \frac{\sum_i w_i}{W^2} = \frac{1}{W}, \qquad I(a) \overset{(3)}{=} -\,\mathbb{E}\,\frac{\partial^2\ln\mathcal{L}}{\partial a^2} = \sum_i w_i = W. \tag{3.135}$$

(1) The variance of a sum of independent terms is the sum of the variances, and a constant factor comes out squared.
(2) $w_i^2\operatorname{Var} y_i = w_i^2\sigma_i^2 = w_i$.
(3) Differentiate the score $\sum_i w_i(y_i - a)$ once more; the result $-W$ is not random, so the expectation does nothing.

The variance equals $1/I(a)$: the estimator sits on the Cramér–Rao bound (3.101). The score $W(\bar{y}_w - a)$ is a linear function of $\hat{a}$, the condition for equality. The error bar is $\sigma_a = W^{-1/2}$, and by linear propagation through $H_0 = e^{\kappa a}$, $\sigma_{H_0} = \kappa H_0\sigma_a$.

From a to H_0. The invariance of maximum likelihood (Section 3.13.2) makes $\hat{H}_0 = 10^{\hat{a}/5}$ the maximum-likelihood estimator of H_0. It is not exactly unbiased, because the exponential is convex. Expand $g(\hat{a}) = e^{\kappa\hat{a}}$ around the truth to second order (the delta method of Section 3.6.3, one order further):

$$\mathbb{E}\,\hat{H}_0 \overset{(1)}{=} \mathbb{E}\Big[g(a) + g'(a)(\hat{a}-a) + \tfrac{1}{2}g''(a)(\hat{a}-a)^2\Big] \overset{(2)}{=} H_0\Big(1 + \tfrac{1}{2}\kappa^2\sigma_a^2\Big) \overset{(3)}{\approx} H_0\, e^{\kappa^2\sigma_a^2/2}. \tag{3.136}$$

(1) Taylor expansion of g; higher orders are smaller by further powers of σ_a.
(2) $g' = \kappa H_0$, $g'' = \kappa^2 H_0$; $\mathbb{E}(\hat{a}-a) = 0$ and $\mathbb{E}(\hat{a}-a)^2 = \sigma_a^2$.
(3) Because $\hat{a}$ is exactly Gaussian, (3.124) with $t = \kappa$ gives the exponential exactly; the delta method gives its first two terms.

The bias-corrected estimator is $\hat{H}_0\, e^{-\kappa^2\sigma_a^2/2}$. Since $\sigma_a \to 0$ as N grows, the uncorrected $\hat{H}_0$ is consistent. For 100 supernovae $\sigma_a \approx 0.01611$ and the bias is 0.001926 km/s/Mpc, utterly negligible. It matters when σ_a is large: an H_0 from a single supernova ($\sigma_a = 0.15$) is biased by

0.24%, and averaging many such single-object values accumulates the bias coherently, which is exactly what went wrong with candidate (b). The cure is always the same: average in the variable where the noise is Gaussian and transform once at the end. The median of $\hat{H}_0$ is exactly H_0 (the exponential is monotonic), and the interval $[10^{(\hat{a}-\sigma_a)/5}, 10^{(\hat{a}+\sigma_a)/5}]$ has exactly the coverage of $\hat{a} \pm \sigma_a$.

Key idea: The maximum-likelihood estimator of H_0

Work in the variable where the noise is Gaussian. For low-redshift supernovae with calibrated M that is $a = 5\log_{10} H_0$, and the estimator is the inverse-variance weighted mean (3.134) of the single-object values $M + 25 + 5\log_{10}(cz_i) - m_i$, with σ_i^2 from (3.127). It is sufficient, unbiased for a, on the Cramér–Rao bound $\sigma_a = (\sum_i w_i)^{-1/2}$, and $10^{\hat{a}/5}$ is consistent for H_0. Converting magnitudes to distances first gives inconsistent estimators.

3.17.3 Step 4: comparing the candidates on simulated surveys

The three biases (3.132), (3.133), (3.136) are large-N approximations, and the claim that the maximum-likelihood error bar is right is a statement about repeated surveys. Both can be checked by simulating many surveys from a known truth, which is the verification use of Monte Carlo (Chapter 6).

Listing 3.18: Mock supernova surveys and the three estimators of H_0 (code/ch03/17_hubble_estimator.py)

```
def sigma_mag(z):
    """Total magnitude scatter: intrinsic plus the peculiar velocity propagated into magnitudes."""
    return np.sqrt(SIG_INT**2 + (5 / np.log(10) * SIG_V / (C * z))**2)

def mock_lowz(shape):
    """A universe where the Hubble law is exact: cz = H0 d + v_pec, m = M + 5 log10(d/10pc) + eps."""
    z = rng.uniform(0.01, 0.1, shape)
    d = (C * z - rng.normal(0, SIG_V, shape)) / H0                  # true distance, Mpc
    m = M_TRUE + 25 + 5 * np.log10(d) + rng.normal(0, SIG_INT, shape)
    return z, m

# ---------- the three candidate estimators ----------
def naive(z, m, M=M_TRUE):
    d = 10 ** ((m - M - 25) / 5)                                    # distance "measured" from each magnitude
    v = C * z
    return (v * d).sum(-1) / (d * d).sum(-1)                        # least squares of v = H0 d through the origin

def ratio(z, m, M=M_TRUE):
    d = 10 ** ((m - M - 25) / 5)
    return (C * z / d).mean(-1)

def ml(z, m, M=M_TRUE):
    w = 1 / sigma_mag(z)**2
    y = M + 25 + 5 * np.log10(C * z) - m                            # each SN's own estimate of a = 5 log10 H0
    a = (w * y).sum(-1) / w.sum(-1)                                 # inverse-variance weighted mean
    sa = 1 / np.sqrt(w.sum(-1))                                     # 1/sqrt(Fisher information)
    return 10 ** (a / 5), a, sa
```

The function `sigma_mag` is (3.127). The function `mock_lowz` draws redshifts uniformly in $0.01 < z < 0.1$, a peculiar velocity for each galaxy, the true distance from (3.121) and a magnitude from (3.122); it generates the data from the physics, not from the linearised likelihood, so the approximation in (3.126) is tested too. The arrays have one row per survey, and every estimator sums over the last axis, so 20000 surveys of 100 supernovae are processed at once. `naive` is (3.131), `ratio` the mean of $v_i/\tilde{d}_i$, and `ml` returns $10^{\hat{a}/5}$, $\hat{a}$ and $\sigma_a = W^{-1/2}$ from (3.134) and (3.135).

estimator	mean bias [km/s/Mpc]	predicted bias [km/s/Mpc]	scatter [km/s/Mpc]	rms error [km/s/Mpc]
(a) least-squares line (3.131)	−0.5076	−0.5181	0.6265	0.8063
(b) mean of ratios	0.2388	0.2376	0.5362	0.587
(c) maximum likelihood (3.134)	0.02722	0.001926	0.5191	0.5198

The predicted biases are (3.132) and (3.133) averaged over the simulated redshifts, and for (c) the convexity term of (3.136). The first two agree with the simulation to within the $O(1/N)$ corrections: for surveys of 400 supernovae the measured biases are −0.5158 and 0.2403, no smaller, while the scatter of the maximum-likelihood estimator halves to 0.2588. The least-squares line is off by about one standard deviation for 100 supernovae and by two for 400. The small residual bias of (c) is not the convexity term but the second-order peculiar-velocity term: the simulated mean of $\hat{a}$ is too large by 7.845×10^{-4} mag, against the predicted 7.927×10^{-4}. The scatter of $\hat{a}$ is 0.0161 against the Cramér–Rao value 0.01611, and the interval $\hat{a}\pm\sigma_a$ covers the truth in a fraction 0.6853 of the surveys ($\pm1.96\sigma_a$: 0.948), as a 68% and a 95% interval should. Figure 3.52 shows the three sampling distributions.

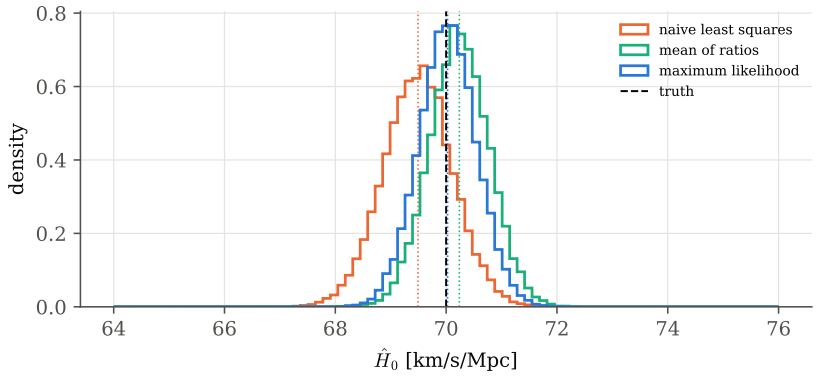


Figure 3.52: Sampling distributions of the three estimators of H_0 over 20 000 mock surveys of 100 supernovae with $H_0 = 70$ (dashed). The dotted lines mark the means. The least-squares line through the magnitude distances is shifted low and is also the widest; the mean of the individual ratios is shifted high; the maximum-likelihood estimator is centred on the truth and is the narrowest.

3.17.4 What supernovae alone cannot tell us: the H_0–M degeneracy

Look again at the likelihood (3.128). The absolute magnitude M and $a = 5\log_{10} H_0$ appear only in the combination $M - a$. Increasing both by the same amount leaves every prediction, and so the likelihood, unchanged. The only thing the supernovae measure is the combination

$$\mathcal{M} \equiv M - 5\log_{10} H_0 + 25,$$

which is a statement about the product "luminosity times H_0^{-2}": brighter candles in a faster-expanding universe look exactly like fainter candles in a slower one. In the Fisher matrix of

(3.128) for the two parameters (M, a),

$$\mathbf{F} \overset{(1)}{=} \sum_i w_i \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} = W \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \qquad \det \mathbf{F} \overset{(2)}{=} W^2 - W^2 = 0, \tag{3.137}$$

(1) $F_{jk} = -\mathbb{E}\,\partial^2 \ln \mathcal{L}/\partial\theta_j \partial\theta_k$; the residual is $m_i - M - 25 - 5\log_{10}(cz_i) + a$, whose derivatives are -1 for M and $+1$ for a, so each supernova contributes w_i times the product of the two derivatives.

(2) Determinant of the 2×2 matrix.

A singular Fisher matrix has a zero mode: the direction $(\delta M, \delta a) \propto (1, 1)$ costs no likelihood at all, and the variance along it is infinite. Figure 3.53 shows this for the survey of Figure 3.51. The $\Delta\chi^2$ contours in the (M, H_0) plane are an open band along $H_0 \propto 10^{M/5}$, and the χ^2 as a function of H_0 at three fixed values of M has three different minima of exactly the same depth. Evaluated numerically for that survey, $\det \mathbf{F} = 0$ to rounding.

The absolute magnitude must therefore come from somewhere else: from supernovae in galaxies whose distances are measured independently, with Cepheid variables or the tip of the red giant branch, which are in turn calibrated geometrically with parallaxes. This is the **distance ladder**. The distance-ladder value of H_0 is 73.04 ± 1.04 km/s/Mpc [24], while the CMB with ΛCDM gives 67.36 ± 0.54 [25], 4.8 standard deviations apart: the **Hubble tension**. Since supernovae only fix $M - 5\log_{10} H_0$, the same tension can be stated as a disagreement of $5\log_{10}(73.04/67.36) = 0.18$ mag in the absolute magnitude of type Ia supernovae. The ladder that supplies M is built as one linear model, with its error budget rung by rung, in Section T2.12, where the tension is also computed (Section T2.12.3).

A calibrated M as a nuisance parameter. Suppose the calibrators give $M = M_{\rm cal} \pm \sigma_M$ with Gaussian errors. That is one more measurement, independent of the Hubble-flow supernovae, and its log-likelihood $-(M - M_{\rm cal})^2/2\sigma_M^2$ adds to (3.128). The Fisher matrix gains $1/\sigma_M^2$ in its MM element. The variance of $\hat{a}$, with M treated as a nuisance parameter, is the aa element of the inverse:

$$\mathbf{F}' = \begin{pmatrix} W + \sigma_M^{-2} & -W \\ -W & W \end{pmatrix}, \qquad \det \mathbf{F}' \overset{(1)}{=} \frac{W}{\sigma_M^2},$$

$$\sigma_a^2 = (\mathbf{F}'^{-1})_{aa} \overset{(2)}{=} \frac{W + \sigma_M^{-2}}{W/\sigma_M^2} \overset{(3)}{=} \frac{1}{W} + \sigma_M^2. \tag{3.138}$$

(1) $(W + \sigma_M^{-2})W - W^2 = W/\sigma_M^2$.

(2) The $(2,2)$ element of the inverse of a 2×2 matrix is its $(1,1)$ element divided by the determinant.

(3) Divide term by term.

This is the profile-likelihood variance $V_{\theta\theta}$ of (3.102) with $\theta = a$ and $\nu = M$: maximising the likelihood over M at each a gives a parabola in a of exactly this width, and for a Gaussian likelihood marginalising over M gives the same (Section 3.14.4). The two errors add in quadrature: the statistical error of the Hubble-flow sample and the calibration error. The estimator itself is (3.134) with $M = M_{\rm cal}$, because the Hubble-flow supernovae cannot move M at all.

For the simulated surveys $W^{-1/2} = 0.01618$ mag, and with a calibration of $\sigma_M = 0.03$ mag, typical of current ladders, the total is $\sigma_a = 0.03409$ mag: the calibration dominates, and adding more Hubble-flow supernovae hardly helps. In the simulation each survey gets its own calibration $M_{\rm cal}$ drawn around the truth. The scatter of $\hat{H}_0$ is then 1.104 km/s/Mpc against

the predicted $\kappa H_0 \sigma_a = 1.098$, and the interval with the full σ_a covers the truth in 0.6798 of the surveys. Quoting only the statistical error $W^{-1/2}$ would give intervals that cover the truth in 0.3605 of the surveys, instead of 68%.

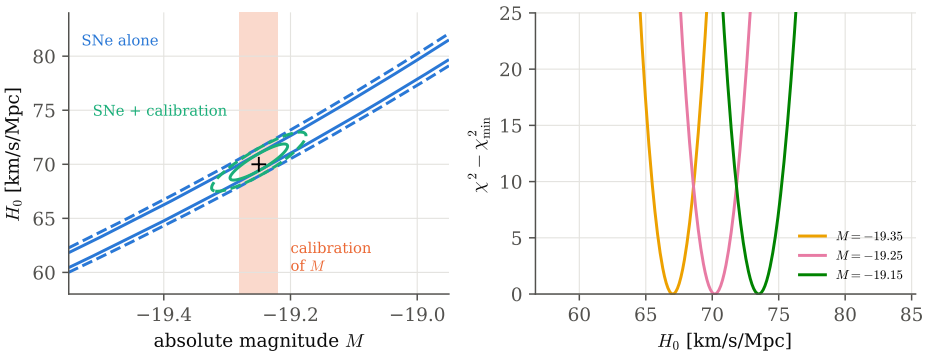


Figure 3.53: The degeneracy between H_0 and the absolute magnitude M for one survey of 100 supernovae. Left: $\Delta\chi^2 = 2.30$ (solid) and 6.18 (dashed) contours. The supernovae alone allow an open band; a calibration of M to ± 0.03 mag (vertical strip) closes it into a small ellipse, whose height is set mostly by the width of the strip. Right: χ^2 against H_0 at three fixed values of M. Each has a perfectly good minimum of the same depth, at a different H_0.

Watch out: An error bar that forgets the calibration

Fixing M at a calibrated value and reporting the curvature error of the Hubble-flow fit, $\kappa H_0 W^{-1/2}$, conditions on the nuisance parameter instead of profiling it. The interval is too short by the factor $(1 + W\sigma_M^2)^{-1/2}$, here about one half, and its coverage falls from 68% to about a third. The calibration error must be carried into every number derived from the calibrated supernovae.

3.17.5 Beyond low redshift: the luminosity distance and a profile over Ω_{m0}

At larger redshift the Hubble law bends. In a flat universe with matter density parameter Ω_{m0} and a cosmological constant, neglecting radiation, the luminosity distance is a standard result of cosmology, taken here as an input:

$$D_L(z) = \frac{c}{H_0}\,(1+z)\int_0^z \frac{\mathrm{d}z'}{E(z')}, \qquad E(z) = \sqrt{\Omega_{m0}(1+z)^3 + 1 - \Omega_{m0}}, \tag{3.139}$$

the distance travelled by light through an expanding universe whose expansion rate is $H_0 E(z)$. The model for the magnitudes becomes $m_i = M + 25 + 5\log_{10}(D_L(z_i)/\mathrm{Mpc}) + \text{noise}$. Write $\bar{D}_L \equiv H_0 D_L / c$ for the dimensionless integral, which depends on Ω_{m0} but not on H_0. Then

$$m_i = M + 25 + 5\log_{10}\!\left[c\,\bar{D}_L(z_i;\Omega_{m0})\right] - a + \text{noise}, \tag{3.140}$$

so a still enters additively, and M and a are still degenerate: the whole of Section 3.17.4 carries over unchanged. What is new is that the model depends on Ω_{m0} through an integral, nonlinearly. The maximum-likelihood estimator is the minimum of $\chi^2(a, \Omega_{m0}) = \sum_i w_i [y_i(\Omega_{m0}) - a]^2$, with $y_i(\Omega_{m0}) = M + 25 + 5\log_{10}[c\bar{D}_L(z_i;\Omega_{m0})] - m_i$. At fixed Ω_{m0} the minimum over a is the weighted mean $\hat{a}(\Omega_{m0}) = \sum_i w_i y_i(\Omega_{m0})/W$ as before, and substituting it leaves the profile $\chi^2_p(\Omega_{m0}) = \sum_i w_i [y_i - \hat{a}(\Omega_{m0})]^2$, a function of one variable that a one-dimensional minimiser handles.

Listing 3.19: The flat-ΛCDM luminosity distance (`code/ch03/17_hubble_estimator.py`)

```
def dl_bar(zv, Om):
    """H0 D_L / c for flat LambdaCDM (radiation neglected): (1+z) int_0^z dz'/E(z')."""
    E = np.sqrt(Om * (1 + ZG)**3 + 1 - Om)
    chi = cumulative_trapezoid(1 / E, ZG, initial=0)
    return (1 + zv) * np.interp(zv, ZG, chi)
```

`dl_bar` evaluates $\bar{D}_L$ by integrating $1/E$ with the trapezoidal rule on a fine grid and interpolating at the supernova redshifts. At the Planck 2018 best-fit parameters, it agrees with the luminosity distance computed by CAMB to a fractional 8.326×10^{-5} out to $z = 1$; CAMB includes radiation and massive neutrinos, which is the size of the difference.

Listing 3.20: Profiling $a = 5\log_{10} H_0$ analytically and minimising over Ω_{m0} (`code/ch03/17_hubble_estimator.py`)

```
def profile_chi2(Om, zv, mv, M=M_TRUE):
    """chi^2 minimised analytically over a = 5 log10 H0 at fixed Omega_m; returns (chi2_p,
    a_hat)."""
    w = 1 / sigma_mag(zv)**2
    y = M + 25 + 5 * np.log10(C * dl_bar(zv, Om)) - mv
    ah = (w * y).sum() / w.sum()
    return (w * (y - ah)**2).sum(), ah

Hfit, Omfit = np.empty(NMH), np.empty(NMH)
for k in range(NMH):
    r = optimize.minimize_scalar(lambda o: profile_chi2(o, zh[k], mh[k])[0], bounds=(0.02,
    0.9),
                                 method="bounded", options={"xatol": 1e-5})
    Omfit[k] = r.x
    Hfit[k] = 10 ** (profile_chi2(r.x, zh[k], mh[k])[1] / 5)
```

The simulated surveys have 300 supernovae uniform in $0.01 < z < 1$, with $H_0 = 70$ and $\Omega_{m0} = 0.3$, and M known. `profile_chi2` returns χ^2_p and $\hat{a}$ at one Ω_{m0}, and `minimize_scalar` finds the minimum over Ω_{m0}. For the first survey a direct two-dimensional minimisation of $\chi^2(H_0, \Omega_{m0})$ with Nelder–Mead gives $H_0 = 70.79$, $\Omega_{m0} = 0.2878$, the same as the profile (70.79, 0.2878): analytic profiling over the linear parameter loses nothing.

Over 1000 surveys the estimates of H_0 average 69.99 with a scatter of 0.6258 km/s/Mpc, and those of Ω_{m0} average 0.3006 with a scatter of 0.02365: both are unbiased. The Fisher matrix in (a, Ω_{m0}), computed with a numerical derivative of $5\log_{10}\bar{D}_L$ with respect to Ω_{m0}, predicts 0.6333 and 0.02399. With Ω_{m0} held fixed the H_0 error would be only 0.2811: the two estimates are correlated with $\rho = -0.8957$, and profiling over Ω_{m0} widens the error by $1/\sqrt{1-\rho^2}$, exactly the mechanism of Section 3.14.2. A higher Ω_{m0} decelerates the expansion, makes distant supernovae nearer and brighter, and is compensated by a lower H_0.

When does the straight Hubble law fail? Expanding (3.139) in z gives the distance modulus $\mu = 25 + 5\log_{10}(cz/H_0) + 1.086\,(1 - q_0)\,z + O(z^2)$, where $1.086 = 5/(2\ln 10)$ and q_0 is the deceleration parameter, $q_0 = \frac{3}{2}\Omega_{m0} - 1 = -0.55$ for flat ΛCDM. The pure Hubble law drops the linear term. Applying the estimator (3.134) to noiseless ΛCDM magnitudes with redshifts up to $z_{\max}$ gives its bias directly, since the estimator is linear in the data (Figure 3.54, right). For $z_{\max} = 0.05$ the bias is −1.687 km/s/Mpc, for $z_{\max} = 0.1$ it is −2.965, about six times the statistical error 0.5192 of 100 supernovae. The Monte Carlo surveys confirm it: −2.988 at $z_{\max} = 0.1$ and −7.16 at 0.3 (predicted −7.171). The bias exceeds the statistical error of 100 supernovae already for $z_{\max} \approx 0.02$. Keeping the q_0 term moves this to $z_{\max} \approx 0.192$, with a bias of 0.1555 km/s/Mpc at $z_{\max} = 0.1$ but 2.873 at 0.5; beyond that only the full (3.139) is safe. In practice the "low-redshift" estimator means the Hubble law with the q_0 term (and often the next one) applied to $0.02 \lesssim z \lesssim 0.15$: the lower cut removes the supernovae whose

peculiar velocities dominate, the upper cut keeps the expansion accurate. Like the bias of the least-squares line, this is a bias of a wrong model, and it does not average away.

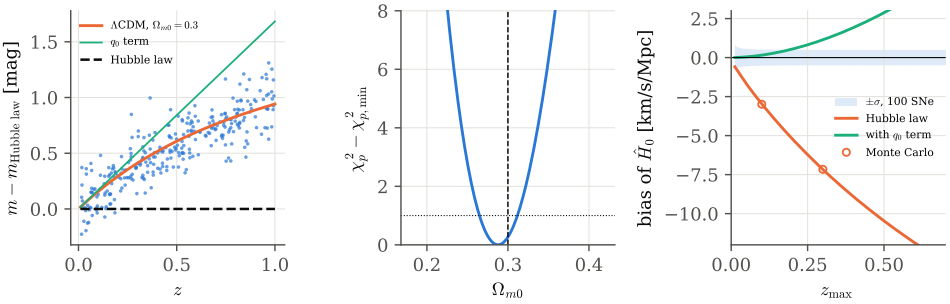


Figure 3.54: Supernovae beyond the linear Hubble law. Left: one survey of 300 supernovae to $z = 1$, as magnitudes minus the Hubble-law prediction; the ΛCDM curve follows the points, the first-order q_0 term follows them only at low redshift. Middle: the profile $\chi^2_p(\Omega_{m0})$ of that survey, with a minimised analytically at each Ω_{m0}; the dotted line is $\Delta\chi^2 = 1$, the dashed line the truth. Right: bias of $\hat{H}_0$ from fitting the pure Hubble law and the law with the q_0 term to supernovae up to $z_{\max}$, compared with the $\pm1\sigma$ statistical error of 100 supernovae (shaded); the circles are Monte Carlo checks.

How the measurement is done in research. Real analyses solve the same problem with the same structure, on a larger scale. The SH0ES measurement [24] writes the Cepheid magnitudes in the anchor galaxies (with geometric distances), the Cepheids and supernovae in the calibrator galaxies, and the Hubble-flow supernovae ($0.023 < z < 0.15$, with the q_0 and j_0 expansion terms) as one linear model $\boldsymbol{y} = A\boldsymbol{\theta}$ + noise with a full covariance matrix, and solves it by generalised least squares, $\hat{\boldsymbol{\theta}} = (A^{\mathsf{T}}\mathbf{C}^{-1}A)^{-1}A^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{y}$ (3.87). The parameter vector contains M and $5\log_{10} H_0$ together with the Cepheid period–luminosity parameters and the host distances, so the degeneracy of Section 3.17.4 is broken inside one fit and σ_M is propagated automatically. The Pantheon+ analysis [26] fits 1550 supernovae with a covariance matrix that includes calibration and light-curve systematics, and fits M, H_0 and Ω_{m0} jointly with the SH0ES calibrators, profiling or marginalising over the light-curve standardisation parameters. Two things our mock surveys ignored enter there: a magnitude-limited survey preferentially finds the brighter supernovae at large distance, a selection bias that no choice of estimator removes and that is calibrated by simulating the survey (step 5, Section 3.16.6); and correlated systematics, which are what the off-diagonal covariance is for (Section 3.10.5). The same problem can be posed as Bayesian inference, with priors on (M, H_0, Ω_{m0}) and a Markov chain exploring the posterior; Chapter 9 builds those chains, and Section 9.23.1 uses one to add the SH0ES value to a CMB chain by reweighting. How the error on H_0 is assembled from the errors of the rungs, before any data are taken, is the error budget of Section T2.12, and the Fisher matrix of Chapter 10 returns to H_0 as a parameter derived from the CMB in Section 10.6.3.

3.17.6 A second probe: cosmic chronometers, where H_0 enters linearly

Supernovae measure distances, and the expansion rate enters only through the integral that relates distance to $H(z)$. There is another method that measures $H(z)$ more directly. Consider galaxies that formed their stars early in cosmic history and then stopped forming new stars. Since no new stars are being added, their stellar populations simply grow older as the Universe evolves, and as the stars age, the spectrum of the galaxy changes in a predictable way. The observed spectrum can therefore be used to estimate the age of the stellar population at a given redshift. Comparing similar passively evolving galaxies at two nearby redshifts, the difference

in their ages measures the difference in cosmic time between those redshifts, which gives $\mathrm{d}t/\mathrm{d}z$. Since $1+z=a^{-1}$,

$$H(z) = -\frac{1}{1+z}\frac{\mathrm{d}z}{\mathrm{d}t} = -\frac{1}{1+z}\left(\frac{\mathrm{d}t}{\mathrm{d}z}\right)^{-1}. \tag{3.141}$$

These galaxies are therefore used as **cosmic chronometers**: clocks whose age differences give $H(z)$ without any distance. A chronometer survey delivers N triples (z_i, H_i, σ_i): a redshift, a measured expansion rate in km/s/Mpc, and its error bar.

The event, and what could have produced it. What we observed is the number H_i at redshift z_i. What made it come out that way? Two things: the true expansion rate at z_i, and the measurement error. In flat ΛCDM the true rate is $H_{\mathrm{th}}(z; H_0, \Omega_{m0}) = H_0\, E(z; \Omega_{m0})$, with the same E as in (3.139). The errors of the age measurements are independent from point to point and Gaussian with standard deviation σ_i. So for each candidate pair (H_0, Ω_{m0}) the noise needed to turn the prediction into the data is $H_i - H_0 E_i$, with $E_i \equiv E(z_i; \Omega_{m0})$, and how plausible that noise is gives the likelihood: $-2\ln\mathcal{L} = \chi^2 + \text{const}$ with

$$\chi^2(H_0, \Omega_{m0}) = \sum_{i=1}^{N} \frac{\left(H_i - H_0 E_i\right)^2}{\sigma_i^2}.$$

Here is the feature that makes the problem easy. The matter density enters E_i through a square root, nonlinearly, but H_0 only multiplies it: *the model is linear in* H_0. At fixed Ω_{m0} the χ^2 is therefore a quadratic function of H_0, a parabola, and its minimum can be written down. With the three sums $S_{HH} \equiv \sum_i H_i^2/\sigma_i^2$, $S_{HE} \equiv \sum_i H_i E_i/\sigma_i^2$ and $S_{EE} \equiv \sum_i E_i^2/\sigma_i^2$,

$$\begin{aligned}
\chi^2(H_0) &\overset{(1)}{=} S_{HH} - 2H_0\, S_{HE} + H_0^2\, S_{EE}, \\
\frac{\mathrm{d}\chi^2}{\mathrm{d}H_0} &\overset{(2)}{=} -2S_{HE} + 2H_0\, S_{EE} = 0 \quad\Longrightarrow\quad \hat{H}_0 = \frac{S_{HE}}{S_{EE}} = \frac{\sum_i H_i E_i/\sigma_i^2}{\sum_i E_i^2/\sigma_i^2}, \\
\chi^2(H_0) &\overset{(3)}{=} S_{EE}\left(H_0 - \hat{H}_0\right)^2 + \chi^2_{\min}, \qquad \chi^2_{\min} = S_{HH} - \frac{S_{HE}^2}{S_{EE}}.
\end{aligned} \tag{3.142}$$

(1) Expand $(H_i - H_0E_i)^2 = H_i^2 - 2H_0H_iE_i + H_0^2E_i^2$, divide by σ_i^2 and sum; H_0 does not depend on i.
(2) Differentiate the quadratic; $S_{EE} > 0$, so the stationary point is the minimum.
(3) Complete the square: $S_{EE}(H_0 - \hat{H}_0)^2 = S_{EE}H_0^2 - 2H_0S_{HE} + S_{HE}^2/S_{EE}$.

No minimiser is needed. The estimator is a *weighted ratio*, and it is also a weighted mean: each point gives its own estimate H_i/E_i of H_0, with variance σ_i^2/E_i^2, and (3.142) is $\sum_i w_i(H_i/E_i)/\sum_i w_i$ with the inverse variances $w_i = E_i^2/\sigma_i^2$ as weights, exactly the weighted mean of Section 3.7.

Bias, variance and the bound. The H_i are independent with $\mathbb{E}\,H_i = H_0E_i$ and $\mathrm{Var}\,H_i = \sigma_i^2$, and the E_i are fixed numbers at fixed Ω_{m0}. Then

$$\begin{aligned}
\mathbb{E}\,\hat{H}_0 &\overset{(1)}{=} \frac{\sum_i E_i\,\mathbb{E}\,H_i/\sigma_i^2}{S_{EE}} \overset{(2)}{=} \frac{H_0\sum_i E_i^2/\sigma_i^2}{S_{EE}} = H_0, \\
\mathrm{Var}\,\hat{H}_0 &\overset{(3)}{=} \frac{\sum_i E_i^2\,\mathrm{Var}\,H_i/\sigma_i^4}{S_{EE}^2} \overset{(4)}{=} \frac{S_{EE}}{S_{EE}^2} = \frac{1}{\sum_i E_i^2/\sigma_i^2}, \\
I(H_0) &\overset{(5)}{=} \mathbb{E}\left[\frac{\partial^2(-\ln\mathcal{L})}{\partial H_0^2}\right] \overset{(6)}{=} \frac{1}{2}\frac{\mathrm{d}^2\chi^2}{\mathrm{d}H_0^2} = S_{EE}.
\end{aligned} \tag{3.143}$$

(1) The expectation is linear and the denominator is not random.
(2) $\mathbb{E}\,H_i = H_0 E_i$.
(3) Variances of independent terms add, each multiplied by the square of its coefficient $E_i/(\sigma_i^2 S_{EE})$.
(4) $\operatorname{Var} H_i = \sigma_i^2$, so each term is E_i^2/σ_i^2.
(5) The Fisher information of Section 3.13.3.
(6) $-\ln\mathcal{L} = \chi^2/2 + \text{const}$, and the second derivative of the parabola, $2S_{EE}$, is not random, so the expectation does nothing.

The estimator is unbiased and its variance is $1/I(H_0)$: it sits on the Cramér–Rao bound (3.101), as it must, since the score $\partial \ln\mathcal{L}/\partial H_0 = -\frac{1}{2}\mathrm{d}\chi^2/\mathrm{d}H_0 = S_{EE}(\hat{H}_0 - H_0)$ is a linear function of $\hat{H}_0$. The completed square in line 3 of (3.142) says the same thing graphically: χ^2 rises by one when H_0 moves by $S_{EE}^{-1/2}$ from its minimum.

Compare with the supernova estimator (3.134). There the parameter $a = 5\log_{10} H_0$ entered the magnitudes *additively*, as a shift, and the estimator was a weighted mean of the y_i; here H_0 enters the rates *multiplicatively*, as a scale, and the estimator is a weighted ratio. Both are instances of one rule: *a parameter that enters the model linearly, with Gaussian noise, has a closed-form maximum-likelihood estimator that is unbiased and on the Cramér–Rao bound*. Unlike candidate (b) of Section 3.17.2, the single-point ratios H_i/E_i carry no bias, because the noise sits in the numerator only and enters linearly.

The matter density. The closed form holds at fixed Ω_{m0}. The matter density is not known exactly, and it must be profiled or marginalised. Profiling is cheap here for the same reason as for the distant supernovae of Section 3.17.5: at each Ω_{m0} the best H_0 is $\hat{H}_0(\Omega_{m0})$ of (3.142), and substituting it leaves the profile $\chi^2_p(\Omega_{m0}) = S_{HH} - S_{HE}^2/S_{EE}$, the $\chi^2_{\min}$ of line 3, a function of one variable for a one-dimensional minimiser. As in Section 3.14.2, letting Ω_{m0} float widens the error on H_0, because a higher matter density raises $E(z)$ at high redshift and is compensated by a lower H_0.

A numerical illustration. Script `code/ch03/18_chronometers.py` draws 30 points uniform in $0.07 < z < 2$ from $H_0 = 70$, $\Omega_{m0} = 0.3$, with 8% errors. For one survey (Figure 3.55) the closed form gives $\hat{H}_0 = 69.74 \pm 1.02$ km/s/Mpc at $\Omega_{m0} = 0.3$. Over 4000 surveys the estimates average 70.02 ± 0.02, unbiased, and scatter by 1.019, against the bound $S_{EE}^{-1/2} = 1.022$. With Ω_{m0} profiled, the first survey gives $\hat{H}_0 = 71.02$ and $\hat{\Omega}_{m0} = 0.282$, and over the surveys the H_0 estimates average 69.95 with a scatter of 3.08, three times wider than at fixed Ω_{m0}, while Ω_{m0} scatters by 0.041. The full chronometer analysis, with real data, the systematic errors of stellar-population models and the comparison with supernovae, is in Section 11.19. Chronometers are the one route to H_0 in the book that needs neither a calibrated candle, as the supernovae of Section T2.12 do, nor a sound horizon, as the CMB of Section 10.6.3 and the BAO of Section T2.15 do.

Where this leaves us. The likelihood is the probability of the data, the object of Chapter 2 and of Section 3.1, read backwards. For every trial value of the parameters it asks how plausible the fluctuation is that would turn the prediction into the data we have. Everything else follows from its shape. Its peak is the estimate. Its width is the error bar. Its peak value, compared with that of a saturated model, tests the fit. Maximising it over the parameters we do not care about removes the nuisances. When no formula is available, the bootstrap and the parametric simulation give the sampling distribution directly. To find the best estimator, sufficiency says what the data can tell us; candidates come from moments, maximum likelihood or a loss function; and the Cramér–Rao bound and the Rao–Blackwell theorem say when to stop

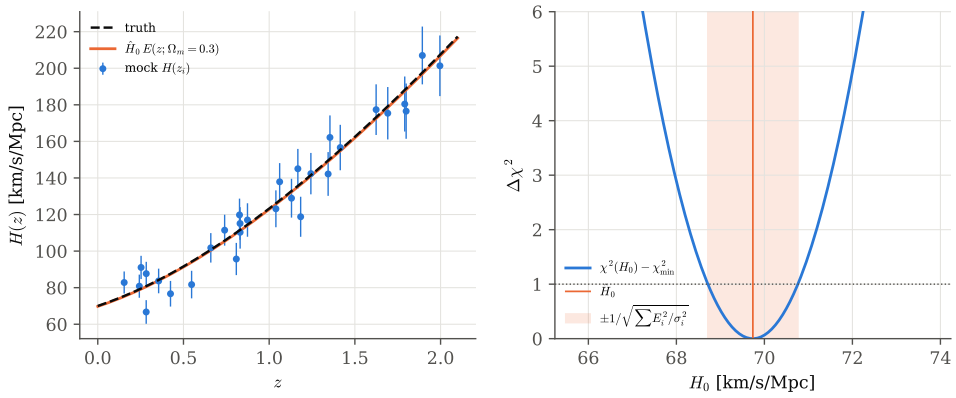


Figure 3.55: Cosmic chronometers. Left: one mock survey of $H(z_i)$ points with their error bars, the true expansion history (dashed) and the closed-form fit $\hat{H}_0 E(z)$ at $\Omega_{m0} = 0.3$ (orange). Right: the χ^2 of that survey as a function of H_0 is an exact parabola; its minimum is $\hat{H}_0$, and it rises by one (dotted) at $\hat{H}_0 \pm S_{EE}^{-1/2}$ (shaded).

improving. The full-sky power spectrum went through every one of these steps. Chapter 4 turns likelihood ratios and χ^2 values into tests and p-values; Chapter 5 multiplies the likelihood by a prior and normalises it; Chapter 9 explores it with Markov chains when it has too many dimensions to plot; Chapter 10 replaces it by its curvature, the Fisher matrix, to forecast experiments before they are built.

Recap: Part 3b in one page

- Data = model + noise. The likelihood $\mathcal{L}(\boldsymbol{\theta}) = p(\boldsymbol{d}_{\text{obs}} \mid \boldsymbol{\theta})$ is the probability (density) of the noise realisation needed to turn the prediction $d(\boldsymbol{\theta})$ into the data. It is a function of $\boldsymbol{\theta}$, not a distribution over $\boldsymbol{\theta}$: no normalisation, no Jacobian, only ratios matter.
- Independent data multiply: $\ell = \sum_i \ln f(x_i;\boldsymbol{\theta})$. The MLE solves $\partial\ell/\partial\boldsymbol{\theta} = 0$; Newton–Raphson $\boldsymbol{\theta} \to \boldsymbol{\theta} - H^{-1}\boldsymbol{g}$ when it cannot be solved by hand.
- Gaussian noise: $-2\ell = \chi^2 = \boldsymbol{r}^\mathsf{T}\mathbf{C}^{-1}\boldsymbol{r}$ (plus $\ln\det\mathbf{C}$ if $\mathbf{C}$ depends on $\boldsymbol{\theta}$). Linear models: $\hat{\boldsymbol{\theta}} = (A^\mathsf{T}\mathbf{C}^{-1}A)^{-1}A^\mathsf{T}\mathbf{C}^{-1}\boldsymbol{D}$ with covariance $(A^\mathsf{T}\mathbf{C}^{-1}A)^{-1}$, exactly. $\chi^2_{\min} \sim \chi^2_{N-m}$, mean $\nu = N - m$. Ignoring correlations keeps the fit unbiased but makes its error bar wrong.
- Counts: $-2\ell = 2\sum[\mu_i - n_i \ln \mu_i]$; the fit conserves the total. Baker–Cousins χ^2_λ fits and tests; Neyman and Pearson χ^2 bias the normalisation by $-\chi^2$ and $+\chi^2/2$.
- Error bar: $\hat{\sigma} = (-\ell''(\hat{\theta}))^{-1/2}$ or the $\Delta\ln\mathcal{L} = \frac{1}{2}$ crossings (better when ℓ is skewed). Why: $\operatorname{Var} S = -\mathbb{E}\,\ell'' = I_n$, and $\hat{\theta} \approx \mathcal{N}(\theta, 1/I_n)$.
- The MLE is consistent (Kullback–Leibler), equivariant ($\widehat{g(\theta)} = g(\hat{\theta})$), asymptotically efficient: Cramér–Rao $\operatorname{Var}\hat{\theta} \geq 1/I_n$ for unbiased estimators of regular models. Median on Gaussian data: efficiency $2/\pi$.
- Nuisance parameters: profile, $\mathcal{L}_p(\theta) = \max_\nu \mathcal{L}(\theta,\nu)$; its width is the full marginal error. Fixing ν underestimates the error by $\sqrt{1-\rho^2}$.
- Finding the estimator: write $\mathcal{L}$; find the sufficient statistic from the factorisation $p = g(U,\theta)h(\boldsymbol{d})$; build candidates; compare with $1/I_n$. In an exponential family the

MLE matches the model mean of U to its observed mean. Rao–Blackwell: $\mathbb{E}[W \mid U]$ never has larger variance than an unbiased W. Bayes estimators: posterior mean, median, mode for squared, absolute, zero–one loss. Full-sky $\widehat{C}_\ell = \sum_m |a_{\ell m}|^2/(2\ell+1)$ is sufficient, unbiased, and sits on the bound $2C_\ell^2/(2\ell+1)$.

- Bootstrap: resample the data with replacement (or simulate from the fit) and use the scatter of the replicas. It fails for statistics controlled by extremes, such as the maximum.

Chapter summary

A number computed from data is a random variable, with a sampling distribution over imagined repetitions of the experiment. The first idea is to ask how such an estimator scatters: its bias, its variance, the $1/\sqrt{N}$ law, the noise in variances and covariance matrices, the propagation of errors through formulas, and the combination of measurements. The second idea is to read the probability of the data backwards, as a function of the parameters. That gives the likelihood, which builds estimators and shows their error bars in its shape. On the pipeline of Chapter 0 these are the first tools for the last arrow, from a statistic's scatter to the physics: given the sampling distribution of a summary, they produce a number, an error bar and a check of the fit. The two tables list each concept, what it means operationally, and where it is used next.

Estimators and their errors (first half)

concept	operational meaning	used later in
statistical model	a family $f(x;\theta)$ of densities, one per parameter value; parametric if θ is finite-dimensional	every chapter
estimator, sampling distribution	a recipe $\hat{\theta} = g(x_1,\dots,x_N)$; its histogram over simulated repetitions of the experiment; standard error = its width	Chapters 4, 6 and 8
bias, variance, MSE	centre offset and width of the sampling distribution; MSE = bias^2 + variance; compare estimators at several true values	Chapters 5 and 10
consistency	$\hat{\theta} \to \theta$ in probability; check that bias and variance both vanish as $N \to \infty$	Chapter 9
$1/\sqrt{N}$, LLN, CLT	Var $\bar{x} = \sigma^2/N$; means become Gaussian with width $\sigma/\sqrt{N}$; fails for the Cauchy	Chapters 6, 7 and 9
sample variance, $N-1$	divide by the number of residual degrees of freedom; Var $S^2 = 2\sigma^4/(N-1)$ for Gaussian data	Chapters 4 and 7
cosmic variance	the variance of a variance: Var $\widehat{C}_\ell = 2C_\ell^2/(2\ell+1)$	Chapters 7 and 8
sample covariance, Hartlap factor	$\hat{\mathbf{C}}$ from N_s simulations is noisy; multiply $\hat{\mathbf{C}}^{-1}$ by $(N_s-p-2)/(N_s-1)$	Chapters 8 and 12
error propagation	$U = AVA^\mathsf{T}$ with the Jacobian A; exact for linear maps; replace by Monte Carlo near poles and for noisy denominators	Chapters 6 and 10
weighted mean	weights $1/\sigma_i^2$; $1/\sigma_{\hat{\mu}}^2 = \sum 1/\sigma_i^2$; check $\chi^2_{\min} \sim \chi^2_{N-1}$; correlated case $V^{-1}\mathbf{1}/(\mathbf{1}^\mathsf{T}V^{-1}\mathbf{1})$	Chapters 4 and 10

The likelihood (second half)

concept	operational meaning	used later in
likelihood	write the noise model, evaluate the probability of the observed data, vary $\boldsymbol{\theta}$; compare values only through ratios	every later chapter
not a density in θ	area arbitrary, no Jacobian; a prior is needed to make it a distribution	Chapter 5
sufficiency, likelihood principle	the likelihood depends on the data only through a sufficient statistic; proportional likelihoods carry the same evidence	Chapters 4, 5 and 7
factorisation theorem, exponential family	$p = g(U,\theta)h(\boldsymbol{d})$ identifies U; in $h\,e^{\eta\phi - A}$ the MLE solves $\mathbb{E}_{\hat\theta}\,\phi = \overline{\phi}$	Chapters 5 and 7
Bayes estimator	minimise the posterior expected loss: mean for squared, median for absolute, mode for zero–one loss	Chapters 5 and 9
Rao–Blackwell, UMVUE	average a crude unbiased estimator over a sufficient statistic; with a complete statistic the result is the best unbiased estimator	Chapter 6
MC-calibrated statistic	when $\mathcal{L}$ is intractable, simulate, calibrate the mean and covariance of a simple statistic	Chapters 6 and 8
full-sky $\widehat{C}_\ell$	sufficient, unbiased, attains Cramér–Rao: cosmic variance $2C_\ell^2/(2\ell+1)$	Chapters 7, 8 and 10
noise lives in magnitudes	write the likelihood in the variable where the noise is Gaussian (m, not d); distances from magnitudes are log-normal, $\mathbb{E}\,\tilde{d} = d\,e^{\kappa^2\sigma^2/2}$	Chapter 9
naive vs ML estimator of H_0	a line fitted to magnitude distances or a mean of ratios is biased and inconsistent; the weighted mean of $5\log_{10} H_0$ is sufficient and on the Cramér–Rao bound	Chapters 9 and 10 and Section T2.12
H_0–M degeneracy	SNe fix only $M - 5\log_{10} H_0$ (singular Fisher matrix); a calibration adds σ_M^2 to $\sigma^2_{5\log_{10} H_0}$; the Hubble tension is 0.18 mag in M	Sections T2.12.3, 9.23.1 and 11.20
maximum likelihood	maximise ℓ; solve the score equation or iterate Newton–Raphson from a moment estimate	Chapters 4, 8 and 10
method of moments	match sample moments to model moments; a quick, consistent starting point	Chapter 6
extended likelihood	multiply by Poisson$(n;\nu)$ when the number of events is random	Chapter 4
least squares	minimise $\chi^2 = \boldsymbol{r}^\mathsf{T}\mathbf{C}^{-1}\boldsymbol{r}$; normal equations; covariance $(A^\mathsf{T}\mathbf{C}^{-1}A)^{-1}$	Chapters 7, 8 and 10
goodness of fit	$p = \mathrm{P}(\chi^2_{N-m} \geq \chi^2_{\min})$; reduced $\chi^2 \approx 1 \pm \sqrt{2/\nu}$	Chapter 4
$\Delta\chi^2$ regions	$\chi^2 \leq \chi^2_{\min} + \Delta\chi^2$ with $\Delta\chi^2$ from χ^2_m (1, 2.30, 3.53 for 68%)	Chapters 9 and 10
correlated and parameter-dependent noise	keep $\ln\det\mathbf{C}(\boldsymbol{\theta})$; GW: $(D-h\|D-h)$ with $S_n(f)$; CMB: $\sum(2\ell+1)[\widehat{C}_\ell/C_\ell + \ln C_\ell]$	Chapters 7, 8 and 10
Poisson binned likelihood	$2\sum[\mu_i - n_i\ln\mu_i]$ (Cash); Baker–Cousins χ^2_λ to test; never $\sqrt{n_i}$ errors for few counts	Chapters 4 and 6
curvature error, $\Delta\ln\mathcal{L} = \frac{1}{2}$	$\hat\sigma = (-\ell'')^{-1/2}$, or the two crossings of $\ell_{\max} - \frac{1}{2}$; covariance $(-H)^{-1}$	Chapters 9 and 10
score, Fisher information	$\mathbb{E}\,S = 0$, $\mathrm{Var}\,S = -\mathbb{E}\,\ell'' = I_n = nI$; $\hat\theta \approx \mathcal{N}(\theta, 1/I_n)$	Chapters 5 and 10
consistency, invariance	the MLE converges to the minimiser of the Kullback–Leibler divergence; $\widehat{g(\theta)} = g(\hat\theta)$, but bias is not invariant	Chapters 5 and 9
Cramér–Rao bound, efficiency	$\mathrm{Var}\,\hat\theta \geq 1/I_n$ for unbiased estimators of regular models; attained by exponential families; fails for moving supports	Chapter 10
profile likelihood	maximise over nuisance parameters at each θ; read $\Delta\ln\mathcal{L}_p = \frac{1}{2}$; Wilks gives the coverage	Chapters 4, 5 and 10
bootstrap	resample the data with replacement (nonparametric) or simulate from the fit (parametric); Normal, pivotal and percentile intervals; fails for extremes	Chapters 6, 8 and 12

Some problems need estimators not built here: data with occasional wild readings, many parameters estimated at once, a parameter that lives in the covariance of the data rather than in their mean, data that arrive one reading at a time, and estimators that need the very value they estimate. Chapter T5 builds those, and sets out the questions that decide which estimator a given problem calls for.

Reading the literature

- **Main line.** [Wasserman Chs. 5–6] for convergence, the law of large numbers, the central limit theorem, estimators, bias, variance and consistency; [Wasserman Ch. 9] for the likelihood, maximum likelihood and its properties (consistency through the Kullback–Leibler divergence, equivariance, Fisher information, asymptotic normality, efficiency, the delta method, the parametric bootstrap), with sufficiency, the factorisation theorem, Rao–Blackwell and exponential families in its appendix; [Wasserman Ch. 8] for the bootstrap and its intervals.
- **The physicist's version.** [Cowan §§1.6–1.7, Ch. 5] for error propagation and estimators; [Cowan Ch. 6] for maximum likelihood with physics examples (lifetimes, the angular distribution, extended and binned likelihoods, goodness of fit); [Cowan Ch. 7] for least squares, binned χ^2 and the normalisation biases; [Cowan Ch. 8] for the method of moments.
- **Why least squares.** [Sivia §3.5]: maximum likelihood and least squares as what Bayes' theorem reduces to under a flat prior and Gaussian noise; [Sivia §2.3] for the weighted mean from a Gaussian posterior.
- **Careful derivations.** Casella and Berger give convergence and the delta method in [CasellaBerger §5.5], the likelihood principle and Birnbaum's argument in [CasellaBerger §6.3], maximum likelihood with restricted ranges and the invariance theorem in [CasellaBerger §7.2.2], in [CasellaBerger §6.2] sufficiency and the factorisation theorem, in [CasellaBerger §7.2.3] Bayes estimators, and in [CasellaBerger §7.3] mean squared error, the Cramér–Rao inequality with its attainment condition, and the Rao–Blackwell and Lehmann–Scheffé theorems.
- **Cosmology.** [HobsonBMC §1.3.1] for the four ingredients of Bayes' theorem and why the likelihood is the least controversial one; [HobsonBMC §2.4] for a χ^2 fit as a Gaussian likelihood and the marginalisation of an unknown noise level. [Padilla §§3.3–3.8] for the likelihood ratio, the Hessian, χ^2 goodness of fit, $\Delta\chi^2$ regions and a first look at the Fisher matrix.
- **Pictures.** [SeeingTheory pp. 16–18, 38–48] for interactive demonstrations of estimation, consistency and the central limit theorem.
- **Original papers.** The noise of inverted simulated covariance matrices [6]; the Poisson statistic for sparse spectra [20]; the likelihood-ratio χ^2 for histograms [21]; the asymptotic χ^2 distribution of likelihood ratios [22]; the bootstrap [23].

Chapter 4

Hypothesis Testing: Asking the Null

Where this chapter is going

A bump in a histogram may be the first sign of a new particle, but even a background-only experiment produces bins with unusually high counts. The more places we look, the more opportunities there are for adjacent bins to resemble a peak. To judge the bump we observed, we therefore need to specify what counted as a candidate in our search and ask how often background alone would produce a candidate at least as striking.

For this search the **null hypothesis** specifies a background-only experiment. Once we specify how to measure the size of a bump, the sampling distribution of that statistic under the null tells us how often background alone would give a value at least as extreme as the observed one (Section 4.1.1). That frequency is a tail probability, the **p-value**. Choosing the summary whose tail we compute (the *test statistic*) and specifying which departures count as extreme gives us a one- or two-sided test. We can then calculate how often the test rejects a true null hypothesis (its size) and how often it detects a specified alternative (its power). For a continuous test statistic we can explain why the p-value is uniform under the null. We then use the Neyman–Pearson lemma to examine the choice of statistic before turning to likelihood ratios, Wilks' theorem, and the χ^2 and Kolmogorov–Smirnov goodness-of-fit tests. We end by translating p-values into the "$Z\sigma$" language of particle physics and asking why a discovery needs 5σ.

Recall the pipeline of Chapter 0: a theory produces data, the data are compressed into a statistic S, the scatter of S is its sampling distribution $P(S \mid \theta)$, and from that scatter we infer the physics. Testing is the last arrow. It uses the sampling distribution *under one special* θ, the null, and asks whether the observed summary is typical of it. Every test can also be turned into a confidence interval, and a search that looks in many places needs its p-value corrected for the look-elsewhere effect; both follow from the same tail probabilities. Chapter 5 then asks the question a p-value cannot answer: how probable is the hypothesis itself, given the data?

4.1 A bump in a histogram: new particle or luck?

We have recorded events in which two photons were produced, and for each event we computed the invariant mass $m_{\gamma\gamma}$ of the photon pair. Figure 4.2 shows the histogram: thirty bins of 2 GeV between 100 and 160 GeV. Known physics produces a smooth, falling spectrum, and from simulation we expect 60 background events in the whole range, distributed as the dashed line. Near 126 GeV two neighbouring bins stick out. A new particle decaying to two photons would show up exactly like this: a narrow excess at its mass. But background alone also fluctuates. Every bin is a Poisson count, some bins come out high and some low, and a pair of high bins can happen by luck.

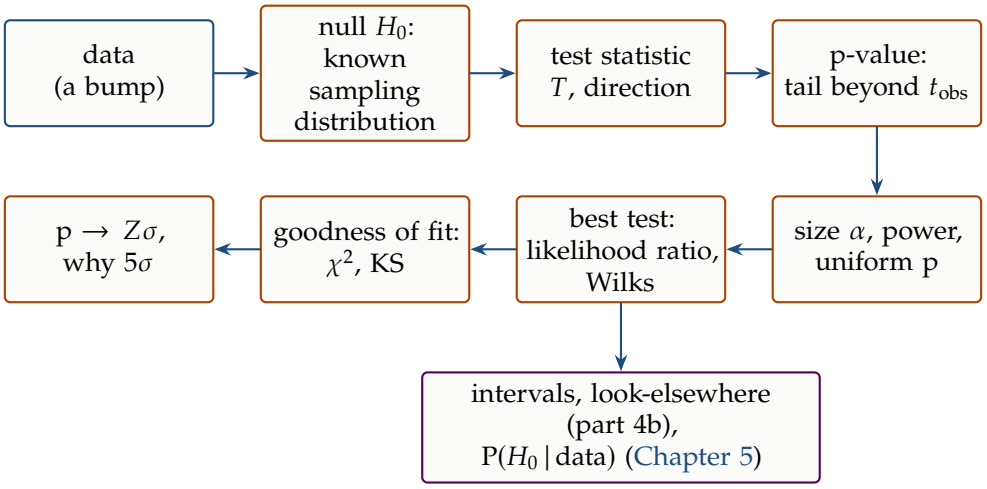


Figure 4.1: The route through this part. We start from data that look surprising and pick the hypothesis whose sampling distribution we know, the null. A summary of the data, the test statistic, and a direction of "more extreme" turn the question "could this be chance?" into a tail probability. The bottom row asks how good a test is, which test is best, how to test a whole model, and how to report the answer. The violet box points to what the p-value cannot tell us and to the next part.

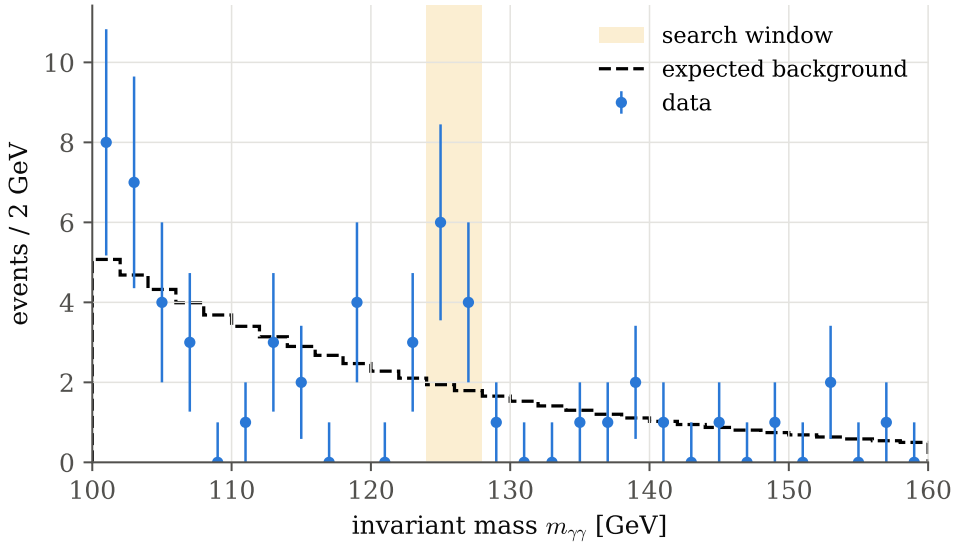


Figure 4.2: A simulated diphoton mass spectrum. The dots are the observed counts per 2 GeV bin, the dashed staircase is the background we expect with no new particle, and the shaded band is the two-bin window 124–128 GeV where the excess sits. The error bars are $\sqrt{n}$ and are drawn only to guide the eye; the pitfall at the end of Section 4.1 explains why they must not be used to judge the excess.

The first thing we can compute is not whether the particle exists. It is how surprising the bump would be *if nothing new were there*. That is weaker than what we want to know, but we can compute it exactly, and every discovery claim in physics starts with it.

The question in words. Before any formula, we can list the three things we need.

1. A hypothesis precise enough to *predict* the data, with no knob left to turn. "Only background" is such a hypothesis: the dashed line tells us the Poisson mean of every bin.

2. One number computed from the data that is large when there is a signal and small when

there is not. Here, the count in the window.

3. The answer to: if only background were present, how often would that number come out at least as large as the one we saw?

4.1.1 The null hypothesis is the one whose predictions we know

Every test starts by splitting the possible values of an unknown parameter into two sets, one for each hypothesis. This is the same for a particle search and for a medical trial. [Wasserman p. 149] opens with rats: expose one group to asbestos, leave another unexposed, and compare the disease rates. Either the two rates are the same, or they are not. In the bump, the unknown parameter is the expected number s of signal events in the window, and physics allows any $s \geq 0$. Writing θ for the parameter and Θ for the set of all its allowed values, the two hypotheses are two disjoint sets that together cover Θ:

$$H_0 : \theta \in \Theta_0 \qquad \text{versus} \qquad H_1 : \theta \in \Theta_1, \qquad \Theta_0 \cup \Theta_1 = \Theta, \quad \Theta_0 \cap \Theta_1 = \varnothing. \tag{4.1}$$

For the bump, $\theta = s$, $\Theta = [0, \infty)$, $\Theta_0 = \{0\}$ and $\Theta_1 = (0, \infty)$. For the rats, θ is the difference of the two disease rates, $\Theta_0 = \{0\}$, and Θ_1 is every nonzero difference.

The two hypotheses are not on an equal footing, and the asymmetry is the whole point. H_0 is the **null hypothesis**: the data came about by chance alone, from processes we already understand. We treat it this way *because we know its probability distribution*. With $s = 0$ the count in every bin is Poisson with a known mean, and we can compute, or simulate, the probability of any outcome. The alternative H_1 is a whole family: every $s > 0$, every possible mass and width of the new particle, each predicting something different. We cannot compute "the" probability of the data under H_1 without first choosing one of its members.

So we test "chance alone" not because we believe it, but because it is the one case whose probability density we actually know. A test therefore runs in the *prediction* direction of Chapter 1. We start from the one situation we can simulate, H_0, and ask what outcomes it produces and how often. Only then do we compare with what we saw.

The question a test asks. Suppose H_0 is true. Would a result like the one we got, or even more in this direction, then be a reasonable thing to expect?

If the answer is "hardly ever", the data give us evidence against H_0. If the answer is "quite often", they do not. Notice what we did *not* ask: how probable H_0 is. That question runs in the inference direction and needs Bayes' theorem, and we come back to it in Section 4.4.4 and Chapter 5.

4.1.2 Test statistic, rejection region, critical value

Data sets are long lists of numbers, and we cannot compare lists directly. We compress the data $\boldsymbol{x}$ into one number, the **test statistic** $T(\boldsymbol{x})$ [a function of the data alone, chosen to be large when H_1 is true]. For the bump, T is the count n in the window. Large n points to a signal, and n near the expected background points to nothing.

A test is then a rule for turning T into a decision. We choose a set of outcomes, the **rejection region** R, and agree in advance [Wasserman (10.2)], [Cowan §4.1]:

$$\boldsymbol{x} \in R \implies \text{reject } H_0, \qquad \boldsymbol{x} \notin R \implies \text{retain } H_0, \qquad R = \{\boldsymbol{x} : T(\boldsymbol{x}) > c\}. \tag{4.2}$$

The threshold c is the **critical value**. Cowan calls the rejection region the *critical region* and its complement the *acceptance region*. We say "retain" rather than "accept": not rejecting the

null only means the data did not contradict it, just as a jury that fails to convict has not proved innocence. The whole problem of testing is to choose a good T and a sensible c. Choosing c is the subject of Section 4.3; choosing T, of Section 4.6.

[Wasserman p. 150] adds a warning worth keeping in mind from the start: tests are for well-defined hypotheses. If the real question is "how big is the effect?", an estimate with a confidence interval answers it better than a test.

4.1.3 The p-value: the tail beyond what we saw

Instead of fixing c in advance and reporting only "reject" or "retain", we can report *where* the observed statistic t_{obs} falls in the null distribution. Figure 4.3 shows the picture that the whole chapter keeps returning to. The curve is $g(t \mid H_0)$, the sampling distribution of the test statistic when the null is true. The vertical line is the value we observed. The shaded area to the right is the probability, under H_0, of a result like ours or even more in the direction that signals H_1. That area is the p-value.

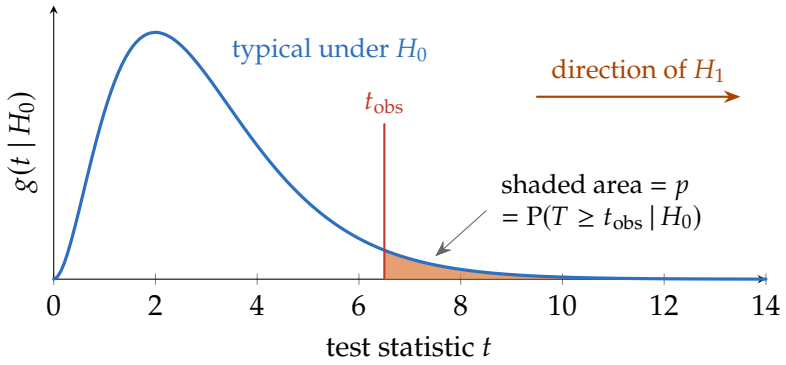


Figure 4.3: The p-value as a tail area. The blue curve is the distribution of the test statistic in a world where only chance operates. The observed value sits at the red line. The orange tail collects every outcome at least as extreme, in the direction the alternative predicts. A small tail means our result would be rare if the null were true.

Working definition 4.1 (p-value)

For a test that rejects H_0 when T is large, and a simple null $\Theta_0 = \{\theta_0\}$, the **p-value** of the observed data is

$$p = \mathrm{P}_{\theta_0}(T \geq t_{\text{obs}})$$

[the probability, computed as if H_0 were true, of a test statistic at least as large as the one observed]. It is computed from H_0 alone and is *not* the probability that H_0 is true. Operationally: simulate many data sets under H_0, compute T for each, and count the fraction with $T \geq t_{\text{obs}}$ [Wasserman Thm 10.12], [Cowan §4.5].

Three features of the definition matter.

- The p-value is a probability of *data*, not of hypotheses. The null value θ_0 sits in the subscript as a fixed assumption.
- It counts the observed value itself and everything beyond it. The words "or even more in this direction" are needed because, once the data are rich enough, the probability of getting exactly t_{obs} is tiny under any hypothesis, so it cannot measure surprise on its own.
- It needs a *direction*. We must say in advance which outcomes count as more extreme. Section 4.2 shows that this choice is part of the question.

4.1.4 The p-value of a Poisson excess

What we need. Under the null, the count N in the window adds up many rare, independent background events, so it is a Poisson variable (Chapter 2). Its mean b is the expected background in the window. So the first step is to find b, and the second is to add up the Poisson tail.

The expected background. Take the background to be exponential in mass, $f(m) \propto e^{-(m-m_0)/L}$ with $m_0 = 100\,\text{GeV}$, decay length $L = 25\,\text{GeV}$, and total expectation $B = 60$ over the range $[100, 160]\,\text{GeV}$. Its cumulative distribution on this range is $F(m) = \big(1 - e^{-(m-m_0)/L}\big)/\big(1 - e^{-60/L}\big)$, and the expected background in the window is

$$\begin{aligned} b &\overset{(1)}{=} B\,\big[F(128) - F(124)\big] \\ &\overset{(2)}{=} B\,\frac{e^{-24/25} - e^{-28/25}}{1 - e^{-60/25}} = 60 \times \frac{0.3829 - 0.3263}{1 - 0.0907} \approx 3.74. \end{aligned} \tag{4.3}$$

(1) The expected number of background events in a window is the total expectation times the fraction of the background probability that falls inside the window.
(2) Insert F. The denominators cancel between $F(128)$ and $F(124)$, and the ones cancel in the difference.

The tail. With b in hand, the p-value of observing n_{obs} events is the probability that a Poisson count with mean b comes out at n_{obs} or above [Cowan (4.38)]:

$$\begin{aligned} p = \mathrm{P}(N \geq n_{\text{obs}} \,|\, H_0) &\overset{(1)}{=} \sum_{n=n_{\text{obs}}}^{\infty} \mathrm{P}(N = n \,|\, s = 0) \\ &\overset{(2)}{=} \sum_{n=n_{\text{obs}}}^{\infty} \frac{b^n e^{-b}}{n!} \\ &\overset{(3)}{=} 1 - \sum_{n=0}^{n_{\text{obs}}-1} \frac{b^n e^{-b}}{n!}. \end{aligned} \tag{4.4}$$

(1) The events $\{N = n\}$ for different n are disjoint, so their probabilities add. The direction is "upwards", because a signal can only add events.
(2) Under H_0 the count is Poisson with mean b.
(3) The Poisson probabilities sum to one, because $\sum_{n\geq 0} b^n/n! = e^b$. So the infinite upper tail equals one minus a *finite* sum, which is what we actually evaluate.

Example: our bump. In the two-bin window there are $n_{\text{obs}} = 10$ events on an expected background of $b = 3.74$. The finite sum in (4.4) runs over $n = 0, \ldots, 9$, and

$$p = \mathrm{P}(N \geq 10 \,|\, b = 3.74) = 0.005172.$$

About one background-only experiment in 190 would give a window at least this full. Figure 4.4 shows the null distribution of the window count with this tail shaded, and the same distribution rebuilt from 10^6 simulated background-only experiments.

Example: Cowan's peak. [Cowan p. 60, Fig. 4.3] shows a histogram with a peak where two bins contain $n_{\text{obs}} = 11$ entries on an expected background of $b = 3.2$. The same formula gives

$$p = 1 - \sum_{n=0}^{10} \frac{3.2^n e^{-3.2}}{n!} = 4.972 \times 10^{-4},$$

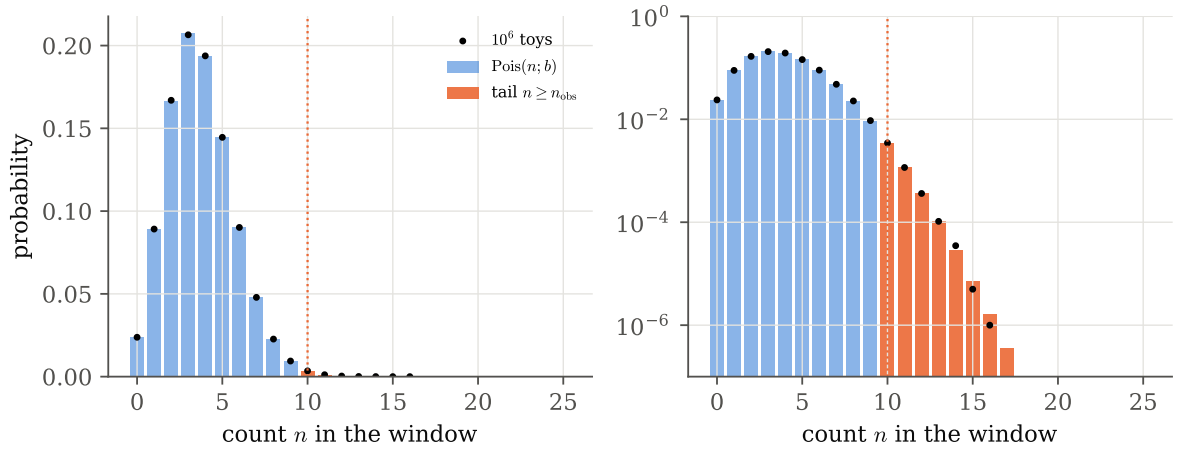


Figure 4.4: The null distribution of the window count, Pois$(n; b)$ with $b = 3.74$, on a linear scale (left) and a log scale (right). The orange bars are the outcomes at least as large as the observed $n_{\mathrm{obs}} = 10$, and their total height is the p-value. The black dots are the frequencies in a million simulated background-only experiments. They follow the bars even far out in the tail, where the log scale is needed to see them.

the value 5.0×10^{-4} quoted there. Fewer than one background-only histogram in two thousand would have such a full pair of bins *at that location*. The last three words matter: a peak could have appeared anywhere in the histogram, and [Cowan p. 61] points out that the fair question is how often a peak this improbable appears in *any* bin. That is the look-elsewhere effect, treated in the second part of the chapter.

Example: a few events on almost no background, and a background we do not quite know. The p-value is computed *entirely* from the null model, so an error in the expected background goes straight into it. Suppose we expect $b = 0.5$ background events and observe $n_{\mathrm{obs}} = 5$ [Cowan p. 59]. Then (4.4) gives $p = 1.721 \times 10^{-4}$. If the background estimate is off and the true value is $b = 0.8$, the same five events give $p = 0.001411$, almost ten times larger. A modest error in the background has turned a striking result into a much less striking one. This is why every real search spends most of its effort on the background estimate, and quotes p-values for a range of plausible b.

4.1.5 Python: exact tail and simulated tail

Listing 4.1 shows the lines of `code/ch04/01_bump_pvalue.py` that compute the p-value twice.

Listing 4.1: The window count, its exact Poisson p-value, and a Monte Carlo p-value (`code/ch04/01_bump_pvalue.py`)

```
win = (edges[:-1] >= 124.0) & (edges[1:] <= 128.0)   # two-bin window 124-128 GeV
b_win = b_bin[win].sum()
n_win = int(n_bin[win].sum())

# exact Poisson p-value:  P(N >= n_obs | b) = 1 - F(n_obs - 1)
p_exact = stats.poisson.sf(n_win - 1, b_win)
# Monte Carlo: background-only experiments in the window
M = 1_000_000
n_toys = rng.poisson(b_win, size=M)
p_mc = np.mean(n_toys >= n_win)
p_mc_err = np.sqrt(p_mc * (1 - p_mc) / M)
z_exact = stats.norm.isf(p_exact)
```

```
z_naive = (n_win - b_win) / np.sqrt(b_win)      # Gaussian approximation with sigma =
    sqrt(b)
z_wrong = (n_win - b_win) / np.sqrt(n_win)      # the "n +- sqrt(n)" error bar
```

The mask `win` picks the two bins between 124 and 128 GeV, and summing the expected background and the observed counts over it gives b and n_{obs}. The function `stats.poisson.sf(k, b)` is the *survival function* $\mathrm{P}(N > k)$, so asking for `sf(n_win - 1, b_win)` gives $\mathrm{P}(N \geq n_{\text{obs}})$, exactly the last line of (4.4). The off-by-one is the most common bug in this computation: $\mathrm{P}(N > n_{\text{obs}})$ leaves out the observed value itself.

The Monte Carlo version is the definition of the p-value turned into code, and it does not use the formula at all. It draws $M = 10^6$ background-only experiments, each a single Poisson count with mean b, and counts the fraction at least as large as the observation. It finds $p = 0.005133$, with a binomial standard error of $\sqrt{p(1-p)/M} = 7.146 \times 10^{-5}$, in agreement with the exact value 0.005172. Here the simulation is overkill, because the tail sum is exact. But when the test statistic is complicated and its null distribution has no formula, which is the usual situation in research, simulation is the only option.

The simulated data do contain a signal: 9 expected events from a narrow resonance at 126 GeV. The test does not know this. With these numbers it cannot tell a real signal of this size from a lucky fluctuation with much confidence. Section 4.3 asks how large a signal must be before a test catches it reliably.

Watch out: The $\sqrt{n}$ error bar answers the wrong question

It is tempting to write the observation as $n \pm \sqrt{n}$, subtract the background, and count standard deviations. For five events on $b = 0.5$ this gives 4.5 ± 2.2, “only two standard deviations” from zero, even though $p = 1.721 \times 10^{-4}$ says the excess is very unusual [Cowan p. 60]. The error bar $\sqrt{n}$ describes how a Poisson variable *with mean n* fluctuates down. The p-value asks how a Poisson variable *with mean b* fluctuates up, and only the second question is posed by the null. For our bump, $(n-b)/\sqrt{n} = 1.98$ and $(n-b)/\sqrt{b} = 3.24$, while the exact tail corresponds to 2.56 Gaussian standard deviations (the conversion is explained in Section 4.9). Neither shortcut is right for small counts. Compute the tail.

Two practical lessons from [Cowan p. 61] complete the recipe. The window must be fixed before we look at the data, with a width of a few times the mass resolution; widening or shifting it to make the excess look better destroys the meaning of the p-value, because the null distribution was computed for a fixed window. And the p-value of one window ignores the other bins, some of which are also a little high or low. A test of the whole histogram against the background shape is a goodness-of-fit test, the subject of Section 4.8.

4.2 Which way is “more extreme”? Sides, and the stopping rule

A coin is tossed $N = 20$ times and comes up heads $n_h = 17$ times [Cowan p. 58]. A fair coin gives 10 heads on average, so 17 looks suspicious. To turn the suspicion into a p-value we must say which other outcomes are “at least as extreme” as 17 heads. Which outcomes qualify depends on the departure from fairness we wanted to detect. The probability assigned to those outcomes also depends on the rule we chose in advance for when to stop tossing. The coin is small enough that we can do every sum by hand and see both effects directly.

The question in words.

1. Which alternative are we guarding against: a coin biased towards heads, or a coin biased either way?
2. Given that choice, which outcomes count as "this, or even more in this direction"?
3. Did the experimenter fix the number of tosses in advance, or stop for some other reason? The imaginary repetitions that the p-value averages over must follow the same plan.

4.2.1 Simple and composite hypotheses, one and two sides

A hypothesis that fixes the distribution of the data completely is **simple**. "The coin is fair", $p = \frac{1}{2}$, is simple: it predicts $n_h \sim \text{Binomial}(20, \frac{1}{2})$ [the number of successes in 20 independent tries, each with success probability $\frac{1}{2}$]. A hypothesis that leaves at least one parameter free is **composite**, for example $p \le \frac{1}{2}$ or $p \neq \frac{1}{2}$ [Wasserman p. 151], [Cowan §4.1].

The alternative decides the shape of the rejection region.

- A **two-sided test**, $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, guards against departures in both directions. For the coin, very many heads and very few heads are both evidence of bias, and a sensible statistic is the distance from the expected value, $T = |n_h - N/2|$.
- A **one-sided test**, $H_0 : \theta \le \theta_0$ versus $H_1 : \theta > \theta_0$ (or the mirror image), guards against a departure in one direction only. If the only suspicion is a coin loaded towards heads, then $T = n_h$ and only large values count.

In both cases the test statistic is built to be *large when H_1 is true*, so the rejection region is always "T large" and the p-value is always an upper tail of T. The direction lives in the choice of T. Figure 4.5 shows the two rejection regions for a statistic that is standard normal under the null, with the same total probability $\alpha = 0.05$ in each: all of it in one tail beyond $z = 1.645$, or half in each tail beyond ± 1.960.

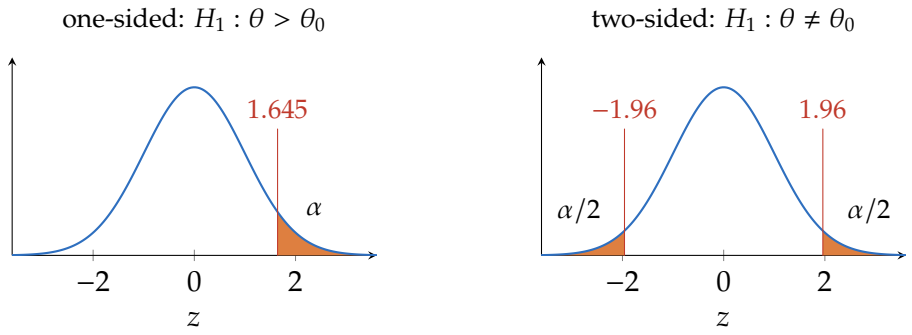


Figure 4.5: The same total probability $\alpha = 0.05$ placed in one tail (left) or split between two tails (right) of a standard normal null distribution. The one-sided test rejects sooner in the direction it cares about, at 1.645 instead of 1.96, and never rejects in the other direction.

Example: 17 heads in 20 tosses. The binomial probability of k heads for a fair coin is $\binom{20}{k}2^{-20}$, and $2^{20} = 1\,048\,576$. If we only feared a coin loaded towards heads, the p-value is

$$p_{\text{one}} = \mathrm{P}(n_h \ge 17 \mid p = \tfrac{1}{2}) \overset{(1)}{=} \frac{1}{2^{20}}\left[\binom{20}{17} + \binom{20}{18} + \binom{20}{19} + \binom{20}{20}\right] \overset{(2)}{=} \frac{1140 + 190 + 20 + 1}{1\,048\,576} = \frac{1351}{1\,048\,576} \approx 0.00129. \tag{4.5}$$

(1) Add the probabilities of the disjoint outcomes $17, 18, 19, 20$ heads.
(2) $\binom{20}{17} = \binom{20}{3} = \frac{20\cdot19\cdot18}{6} = 1140$, $\binom{20}{18} = \binom{20}{2} = 190$, $\binom{20}{19} = 20$, $\binom{20}{20} = 1$.

If any bias counts, the outcomes at least as far from 10 as 17 are $n_h \in \{0,1,2,3\} \cup \{17,\dots,20\}$, that is $|n_h - 10| \geq 7$. The binomial with $p = \frac{1}{2}$ is symmetric, $\binom{20}{k} = \binom{20}{20-k}$, so the lower tail equals the upper one and

$$p_{\text{two}} = \mathrm{P}\big(|n_h - 10| \geq 7 \,|\, p = \tfrac{1}{2}\big) = 2\,p_{\text{one}} = \frac{2702}{1\,048\,576} \approx 0.0026,$$

the value 0.0026 of [Cowan p. 58]. Figure 4.6 shows the eight outcomes that enter. The two answers differ by a factor of two, and both are correct answers to different questions.

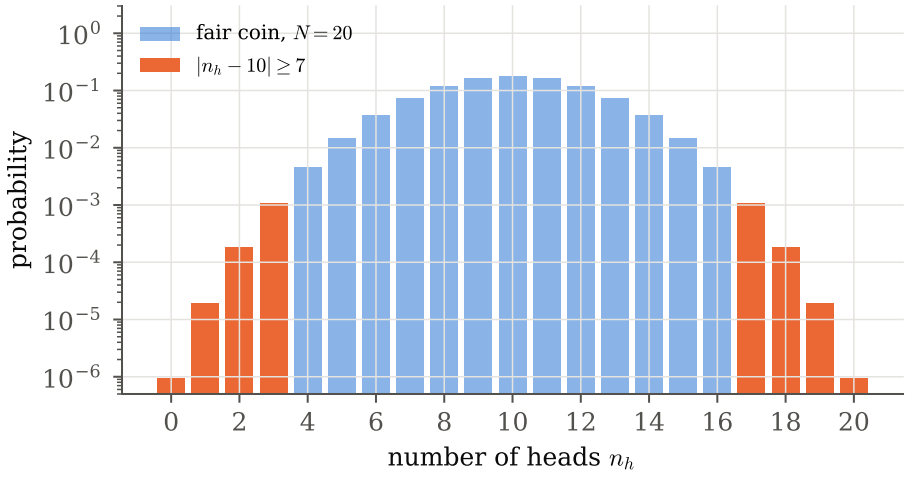


Figure 4.6: The binomial distribution of the number of heads in 20 tosses of a fair coin, on a logarithmic vertical scale so that the thin tails show. The orange bars are the outcomes at least as far from 10 as the observed 17, on either side. Their total is the two-sided p-value; the four bars on the right alone give the one-sided one.

A composite null. For a one-sided composite null, the p-value is the tail at the boundary of the null, so the composite test reduces to the simple test at θ_0. To see why, pose the one-sided coin question with the composite null $H_0 : p \leq \frac{1}{2}$, where p is now the probability of heads. The problem is that every p in Θ_0 gives a different tail probability $\mathrm{P}_p(n_h \geq 17)$. The rule [Wasserman Thm 10.12] is to report the largest of them, $\sup_{\theta\in\Theta_0} \mathrm{P}_\theta(T \geq t_{\text{obs}})$, so that the number is honest for every member of the null. The tail $\mathrm{P}_p(n_h \geq 17)$ grows with p, because a coin more inclined to heads gives more heads. So the largest tail is reached at the boundary $p = \frac{1}{2}$, and the answer is again 0.00129. The boundary value is the member of the null hardest to tell apart from the alternative.

When the null distribution is not symmetric. A skewed null needs a convention for its two-sided p-value. For the fair coin, “twice the one-sided tail” and “everything at least as far from the mean” agree. For a skewed null, such as a Poisson with small mean, they do not, and we must choose one of two rules: either double the smaller tail, or add the probabilities of all outcomes that are no more probable than the one observed. In searches for new physics the question is almost always one-sided anyway, because a signal can only add events.

Watch out: Choosing the side after seeing the data

The side is part of the question, so it must be chosen before looking at the data. Deciding on a one-sided test *because* the excess came out upwards halves the p-value without any new evidence. Under the null this procedure rejects whenever either tail holds α of the probability, so its true type I error rate is 2α, not α.

4.2.2 The same data, different p-values: the stopping rule

The same data can give different p-values, depending on how the experimenter decided when to stop. The reason is in what the p-value asks for: the probability of outcomes "like this or more extreme" in *repetitions of the experiment*. Those repetitions never happened. We imagine them, or, in the blunt words of [Lambert §2.8], to test a hypothesis we "invent fictitious samples". Lambert's example: if we hypothesise that heights are normal with mean 1.55 m and standard deviation 0.3 m and meet one person of 2.5 m, the p-value is the probability of meeting someone *at least* 2.5 m tall, although nobody taller was observed. To imagine the repetitions we must know the plan that produced the data, and the data alone do not tell us.

The data. [Kruschke §11.1] gives the example. A coin shows $z = 7$ heads in $N = 24$ tosses. Is it fair? We take two possible plans in turn.

Plan A: the number of tosses was fixed at $N = 24$. Then the random quantity is the number of heads z, it is Binomial$(24, \frac{1}{2})$ under the null, and few heads is the extreme direction:

$$p_A = \mathrm{P}(z \le 7 \,|\, N = 24) = \sum_{k=0}^{7} \binom{24}{k} 2^{-24} = 0.0320.$$

Plan B: the experimenter tossed until the seventh head. Now $z = 7$ is fixed by design and the random quantity is the number of tosses N it took. Needing *many* tosses to collect seven heads is the sign of a tail-heavy coin, so the extreme direction is $N \ge 24$. The event $\{N \ge 24\}$ is the same as "at most six heads in the first 23 tosses":

$$\begin{aligned} p_B = \mathrm{P}(N \ge 24 \,|\, z = 7) &\overset{(1)}{=} \mathrm{P}\big(\text{at most 6 heads in the first 23 tosses}\big) \\ &\overset{(2)}{=} \sum_{k=0}^{6} \binom{23}{k} 2^{-23} = 0.0173. \end{aligned} \tag{4.6}$$

(1) The seventh head arrives at toss 24 or later exactly when the first 23 tosses contain six heads or fewer.

(2) The first 23 tosses of a fair coin are Binomial$(23, \frac{1}{2})$, whatever happens afterwards.

The distribution of N itself is the negative binomial of Chapter 2: the n-th toss is the seventh head when it is a head and the previous $n - 1$ tosses hold exactly six heads, so $\mathrm{P}(N = n) = \binom{n-1}{6} 2^{-n}$. Figure 4.7 shows the two clouds of imaginary outcomes side by side.

What the two plans say. The very same sequence of tosses is "not significant" under one plan and "significant" under the other. With the conventional two-tailed threshold, which puts 2.5% in each tail, plan A gives $p_A = 0.0320$, above 0.025, and plan B gives $p_B = 0.0173$, below it. [Kruschke p. 309] adds two more plans. If the experimenter tossed for a fixed length of time, so that N itself is random (Poisson), the p-value is 0.024. If a second coin was tested in the same study and the question is whether *either* coin looks this extreme, the chance that at least one of two independent fixed-N experiments reaches the tail is $1 - (1 - p_A)^2 \approx 2p_A$, which is the 0.063 found there. The last case is our first taste of the look-elsewhere effect.

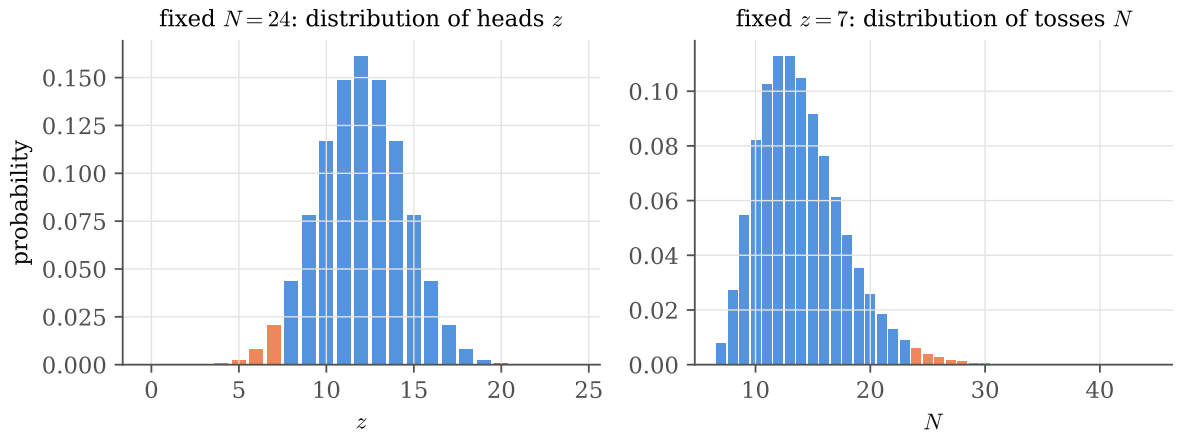


Figure 4.7: Two plans that can produce the same record of 7 heads in 24 tosses. Left: with $N = 24$ fixed, the random quantity is the number of heads, and its lower tail $z \le 7$ (orange) is the p-value. Right: when tossing stops at the seventh head, the random quantity is the number of tosses, and its upper tail $N \ge 24$ (orange) is the p-value. The observed data are identical; the imagined repetitions are not.

Example: Cowan's stopping rule. [Cowan p. 58] tells the same story with the 17 heads in 20 tosses. Suppose the plan was "toss until there are at least three heads and at least three tails", and the stopping happened at toss 20 (with 17 heads and 3 tails). Now the extreme outcomes are those that need 20 tosses or more:

$$\begin{aligned} \mathrm{P}(N_{\text{stop}} \ge 20) &\overset{(1)}{=} \mathrm{P}\big(\text{after 19 tosses, heads} \le 2 \text{ or tails} \le 2\big) \\ &\overset{(2)}{=} 2\,\mathrm{P}\big(\text{Binomial}(19, \tfrac{1}{2}) \le 2\big) \\ &\overset{(3)}{=} 2\,\frac{\binom{19}{0} + \binom{19}{1} + \binom{19}{2}}{2^{19}} = \frac{2\,(1 + 19 + 171)}{524\,288} \approx 7.286 \times 10^{-4}. \end{aligned} \tag{4.7}$$

(1) We are still tossing after toss 19 exactly when one of the two faces has appeared at most twice so far.
(2) The two events are disjoint (both at most 2 would mean at most 4 tosses, not 19), so their probabilities add, and by symmetry they are equal.
(3) Binomial sum with $\binom{19}{2} = 171$ and $2^{19} = 524\,288$.

That is 0.073% instead of the 0.26% of the fixed-N plan [Cowan p. 58]. Cowan's resolution is a convention: always interpret "similar experiments" as experiments with the same number of observations. It is arbitrary, but it makes a reported p-value unambiguous.

Python. The script `code/ch04/02_sides_and_stopping.py` evaluates every tail sum above with `scipy.stats.binom` and then checks each one by running the stated protocol many times.

Listing 4.2: Two stopping plans, and Cowan's rule, simulated (`code/ch04/02_sides_and_stopping.py`)

```
# ---------- (ii) 7 heads in 24 tosses, two stopping intentions ----------
N2, z2 = 24, 7
p_fixN = stats.binom.cdf(z2, N2, 0.5)                    # P(z <= 7 | N = 24)
# fixed z: stop at the 7th head.  "As extreme or more": needing N >= 24 tosses,
# i.e. at most 6 heads among the first 23 tosses.
p_fixz = stats.binom.cdf(z2 - 1, N2 - 1, 0.5)
# simulate the fixed-z protocol: number of tosses to get the 7th head = 7 + (tails before
    it)
```

```
Nstop = z2 + rng.negative_binomial(z2, 0.5, size=M)
p_fixz_mc = np.mean(Nstop >= N2)
p_fixN_mc = np.mean(rng.binomial(N2, 0.5, size=M) <= z2)

# ---------- (iii) Cowan: toss until >= 3 heads and >= 3 tails ----------
# needing >= 20 tosses <=> after 19 tosses, heads <= 2 or tails <= 2
p_cowan = 2 * stats.binom.cdf(2, 19, 0.5)
seqs = rng.integers(0, 2, size=(200_000, 60), dtype=np.int8)
h = np.cumsum(seqs, axis=1); t = np.arange(1, 61) - h
done = (h >= 3) & (t >= 3)
nstop = done.argmax(axis=1) + 1
p_cowan_mc = np.mean(nstop >= 20)
```

For plan B it uses the fact that the number of tails before the seventh head is a negative binomial variable, `rng.negative_binomial(7, 0.5)`, so the total number of tosses is seven plus that. For Cowan's rule it simply tosses: 200 000 sequences of 60 tosses, running totals of heads and tails with `np.cumsum`, and the first toss at which both totals reach three. The simulations give 0.0325, 0.0173 and 7.500×10^{-4} against the exact 0.0320, 0.0173 and 7.286×10^{-4}, and 0.0028 against 0.0026 for the two-sided coin test.

Key idea: A p-value depends on the outcomes that did not happen

A p-value is a sum over imaginary repetitions of the experiment. Which repetitions are imagined depends on the plan: the alternative we guard against (one or two sides), when the experimenter would have stopped, and how many other places or coins we would also have looked at. Change the plan and the p-value changes, even if the data do not.

A Bayesian analysis does not have this problem, because it uses only the probability of the data actually observed. In plan A that probability is $\binom{24}{7}p^7(1-p)^{17}$ and in plan B it is $\binom{23}{6}p^7(1-p)^{17}$. As functions of the coin's bias p, the two likelihoods differ only by a constant factor, which cancels in Bayes' theorem. Any inference built from the likelihood alone therefore reaches the same conclusion under both plans [Kruschke §11.1.7]. The p-value, which sums over outcomes that were not observed, does not. We take this up in Chapter 5. For now the practical rule is Cowan's: fix the plan in advance, state it, and compute the p-value for that plan.

4.3 Two ways to be wrong: size, power and how much data we need

A p-value tells us how surprising one data set is. Before the experiment runs, we face a different pair of questions. How often will our procedure cry "discovery" when there is nothing there? And if there *is* something there, of a given size, how often will we catch it? The first number protects the field from false claims. The second tells us whether the experiment is worth building. A detector that can only see a signal one time in ten is a bad investment, however careful the analysis.

[Wasserman p. 150] offers the legal trial as a model. The defendant is presumed innocent, the null, and is convicted only if the evidence is strong. Two mistakes are possible: convicting an innocent person, and acquitting a guilty one. A court that never convicts makes no mistakes of the first kind, and is useless. There is a trade-off, and it can be put into numbers.

The question in words.

1. If the null is true, what is the probability that the test rejects it?
2. If a particular alternative is true, what is the probability that the test rejects the null?

3. How do both depend on the critical value and on the amount of data, and how large must an experiment be to catch an effect of a given size?

4.3.1 Type I and type II errors

Every test ends in one of four situations [Wasserman Table 10.1]:

	retain H_0	reject H_0
H_0 true	correct	**type I error** (false alarm)
H_1 true	**type II error** (missed signal)	correct

Cowan calls them errors of the first and second kind [Cowan §4.1]. Picture the test statistic t with its two sampling distributions, $g(t \mid H_0)$ and $g(t \mid H_1)$, and a cut t_{cut} above which we reject (Figure 4.8). The probability of a type I error is the area of the null distribution above the cut, and the probability of a type II error is the area of the alternative below it [Cowan (4.1)–(4.2)]:

$$\alpha = \int_{t_{\text{cut}}}^{\infty} g(t \mid H_0)\, \mathrm{d}t, \qquad \beta_{\text{II}} = \int_{-\infty}^{t_{\text{cut}}} g(t \mid H_1)\, \mathrm{d}t. \tag{4.8}$$

Moving the cut to the right shrinks α and grows β_{II}; moving it left does the opposite. With a fixed statistic and a fixed amount of data we cannot shrink both. Only a better statistic (Section 4.6) or more data separates the two curves.

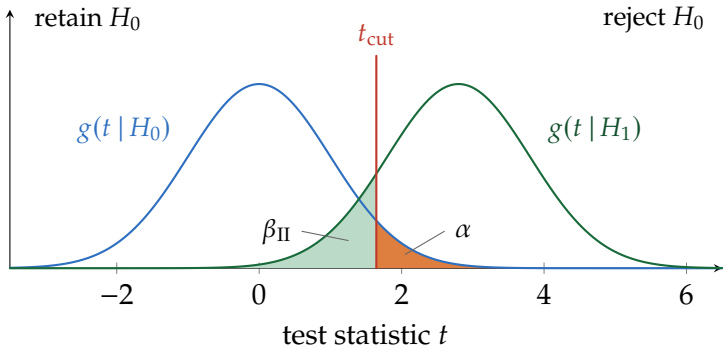


Figure 4.8: The statistic has one distribution under the null hypothesis (blue) and another under the specified alternative (green). We reject the null for values to the right of the red cut. The orange tail therefore gives the false-alarm probability α, while the pale green area to the left gives the probability of missing the signal, β_{II}. Sliding the cut trades one for the other.

4.3.2 The power function

The alternative is rarely a single value, so we want the rejection probability as a function of the true parameter.

Working definition 4.2 (power function, size, level)

The **power function** of a test with rejection region R is

$$\beta(\theta) = \mathrm{P}_\theta(\boldsymbol{X} \in R)$$

[the probability of rejecting H_0 when the true parameter is θ]. For $\theta \in \Theta_1$ it is the probability of a correct rejection; for $\theta \in \Theta_0$ it is a false-alarm rate. The **size** is the

worst false-alarm rate, $\alpha = \sup_{\theta\in\Theta_0}\beta(\theta)$, and a test has **level** α if its size is at most α [Wasserman Def. 10.1]. Operationally: simulate data at each θ and record the fraction of data sets that the test rejects.

A notational trap: Wasserman uses $\beta(\theta)$ for the power, while [Cowan (4.2)] uses β for the type II error probability, so that Cowan's power is $1-\beta$. Here we use Wasserman's convention and write the type II error probability at a point $\theta_1 \in \Theta_1$ as $\beta_{\text{II}} = 1 - \beta(\theta_1)$.

Example: the one-sided test for a Gaussian mean. We take n readings $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ with σ known and test $H_0 : \mu \le 0$ against $H_1 : \mu > 0$ by rejecting when the sample mean $\bar{X}$ exceeds a cut c [Wasserman Ex. 10.2]. Recall from Chapter 3 that $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$, so $Z = \sqrt{n}(\bar{X}-\mu)/\sigma$ is standard normal. The power function is

$$\begin{aligned}\beta(\mu) &\overset{(1)}{=} \mathrm{P}_\mu\big(\bar{X} > c\big)\\ &\overset{(2)}{=} \mathrm{P}_\mu\Big(\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} > \frac{\sqrt{n}(c-\mu)}{\sigma}\Big)\\ &\overset{(3)}{=} \mathrm{P}\Big(Z > \frac{\sqrt{n}(c-\mu)}{\sigma}\Big)\\ &\overset{(4)}{=} 1 - \Phi\Big(\frac{\sqrt{n}(c-\mu)}{\sigma}\Big).\end{aligned} \tag{4.9}$$

(1) Definition of the power function with $R = \{\bar{X} > c\}$.
(2) Subtract μ and multiply by $\sqrt{n}/\sigma > 0$ on both sides; this does not change the event.
(3) The left-hand side is standard normal whatever μ is.
(4) Φ is the standard normal CDF, $\Phi(z) = \mathrm{P}(Z \le z)$.

As μ increases the argument of Φ decreases, so $\beta(\mu)$ increases. The largest false-alarm rate over the null region $\mu \le 0$ is therefore at the boundary:

$$\alpha = \sup_{\mu\le 0}\beta(\mu) = \beta(0) = 1 - \Phi\Big(\frac{\sqrt{n}\,c}{\sigma}\Big) \quad\Longrightarrow\quad c = \frac{\sigma\, z_\alpha}{\sqrt{n}}, \qquad z_\alpha \equiv \Phi^{-1}(1-\alpha).$$

So the size-α test rejects when $\sqrt{n}\,\bar{X}/\sigma > z_\alpha$: the sample mean, measured in units of its own standard deviation, must exceed $z_{0.05} = 1.645$ for $\alpha = 0.05$.

Figure 4.9 (left) plots (4.9) for $n = 4$, 16 and 64 with $\sigma = 1$, together with simulated rejection rates. All three curves pass through $\alpha = 0.05$ at $\mu = 0$: the size does not depend on n, because the cut moves in as $1/\sqrt{n}$. What depends on n is the steepness. Quadrupling the data halves the width of $\bar{X}$'s distribution, so the curve rises over half the range of μ. In the simulation the rate of false alarms at $\mu = 0$ is 0.0508 for $n = 4$ and 0.0518 for $n = 64$, both equal to 0.05 within the simulation error $\sqrt{0.05\times 0.95/20\,000} \approx 0.0015$.

Example: how many readings do we need? Suppose we want power 0.8 at $\mu_1 = 0.5\sigma$, with $\alpha = 0.05$. Setting (4.9) equal to 0.8 at $\mu = \mu_1$, with $c = \sigma z_\alpha/\sqrt{n}$,

$$\begin{aligned}0.8 &\overset{(1)}{=} 1 - \Phi\Big(z_\alpha - \frac{\sqrt{n}\,\mu_1}{\sigma}\Big) \overset{(2)}{=} \Phi\Big(\frac{\sqrt{n}\,\mu_1}{\sigma} - z_\alpha\Big)\\ &\Longrightarrow \frac{\sqrt{n}\,\mu_1}{\sigma} - z_\alpha \overset{(3)}{=} z_{0.2} \Longrightarrow n \overset{(4)}{=} \Big(\frac{(z_\alpha + z_{0.2})\,\sigma}{\mu_1}\Big)^2 = \Big(\frac{1.645 + 0.842}{0.5}\Big)^2 \approx 24.7.\end{aligned} \tag{4.10}$$

(1) Insert $c = \sigma z_\alpha/\sqrt{n}$ into $\sqrt{n}(c-\mu_1)/\sigma$.

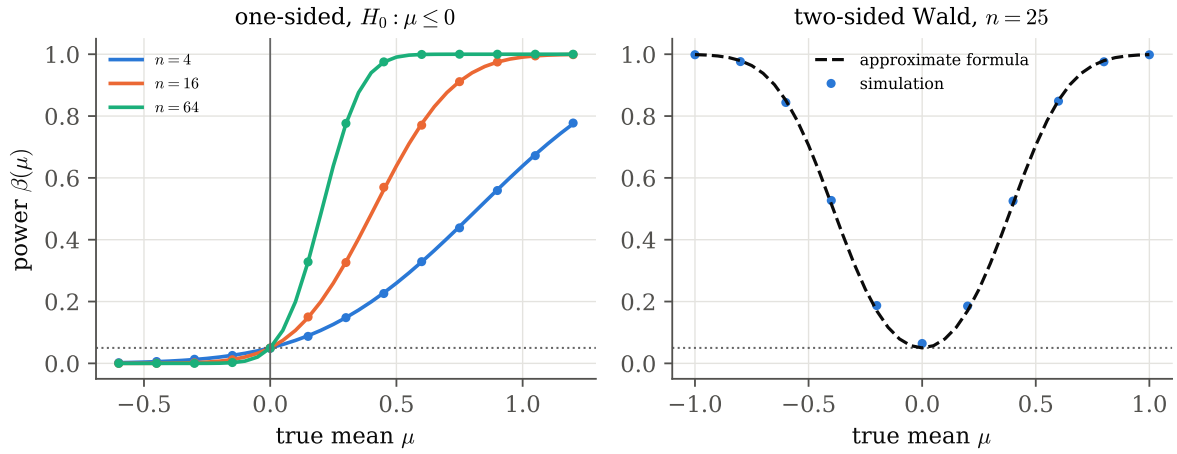


Figure 4.9: Power functions. Left: the one-sided test of $\mu \le 0$ for three sample sizes, with the formula as lines and simulated rejection rates as dots. Every curve crosses the dotted $\alpha = 0.05$ line at $\mu = 0$ and climbs faster the more data there are. Right: the two-sided Wald test with $n = 25$ and an estimated standard error; the dashed curve is the approximate formula and the dots are simulations. The curve is symmetric and bottoms out near α at $\mu = 0$.

(2) The normal CDF is symmetric: $1 - \Phi(x) = \Phi(-x)$.
(3) Apply Φ^{-1}; $\Phi^{-1}(0.8) = z_{0.2} = 0.842$ in the notation $z_a = \Phi^{-1}(1-a)$.
(4) Solve for n.

We need $n = 25$ readings, and the simulation with $n = 25$ rejects in a fraction 0.801 of experiments at $\mu = 0.5$. The formula says what experiment planning comes down to: the required number of readings grows as $1/\mu_1^2$, so detecting an effect half as large costs four times as much data. [Kruschke §11.5.1] makes the same point: planning an experiment is where sampling distributions are most useful.

Example: the power of the two-sided Wald test. Most tests are two-sided and use an estimate $\hat\theta$ with standard error $\widehat{\text{se}}$; the Wald test of Section 4.5 rejects $H_0 : \theta = \theta_0$ when $\left|\hat\theta - \theta_0\right| / \widehat{\text{se}} > z_{\alpha/2}$. Suppose the truth is $\theta_\star \neq \theta_0$ and $\hat\theta$ is approximately $\mathcal{N}(\theta_\star, \text{se}^2)$, so that $Z = (\hat\theta - \theta_\star)/\text{se}$ is approximately standard normal. Write $\delta = (\theta_0 - \theta_\star)/\text{se}$. Then

$$\begin{aligned}\beta(\theta_\star) &\overset{(1)}{=} \mathrm{P}_{\theta_\star}\Big(\frac{\hat\theta - \theta_0}{\text{se}} > z_{\alpha/2}\Big) + \mathrm{P}_{\theta_\star}\Big(\frac{\hat\theta - \theta_0}{\text{se}} < -z_{\alpha/2}\Big) \\ &\overset{(2)}{=} \mathrm{P}\big(Z > \delta + z_{\alpha/2}\big) + \mathrm{P}\big(Z < \delta - z_{\alpha/2}\big) \\ &\overset{(3)}{=} 1 - \Phi\Big(\frac{\theta_0 - \theta_\star}{\text{se}} + z_{\alpha/2}\Big) + \Phi\Big(\frac{\theta_0 - \theta_\star}{\text{se}} - z_{\alpha/2}\Big). \end{aligned} \tag{4.11}$$

(1) The two-sided rejection region is the union of two disjoint tails.
(2) $\hat\theta - \theta_0 = (\hat\theta - \theta_\star) + (\theta_\star - \theta_0)$, so dividing by se gives $Z - \delta$; move δ across.
(3) Write both tails with the normal CDF.

This is [Wasserman (10.6)]. Since se shrinks as $1/\sqrt{n}$, the argument δ grows, and one of the two terms tends to one: the power is large when $\theta_\star$ is far from θ_0 or when the sample is large [Wasserman p. 154]. In the right panel of Figure 4.9 the simulation follows the formula. At $\mu = 0$ it rejects in a fraction 0.0643 of experiments, a little above 0.05, because the standard error was estimated from only 25 readings. Section 4.5 explains that excess and removes it.

4.3.3 The power of a counting experiment

Return to a search like the bump: a window with known background $b = 3.2$ and an unknown signal $s \geq 0$, so that $N \sim \text{Pois}(s + b)$. Particle physics sets thresholds not as $\alpha = 0.05$ but as "3σ" and "5σ", meaning one-sided tail probabilities 0.00135 and 2.867×10^{-7} (Section 4.9 explains the translation). The test rejects when $N \geq n_{\text{crit}}$, where n_{crit} is the smallest count whose null tail is below the threshold:

$$n_{\text{crit}} = \min\{n : \mathrm{P}(N \geq n \mid s = 0) \leq p_{\text{thr}}\}, \qquad \beta(s) = \mathrm{P}(N \geq n_{\text{crit}} \mid s) = 1 - \sum_{k=0}^{n_{\text{crit}}-1} \frac{(s+b)^k e^{-(s+b)}}{k!}. \tag{4.12}$$

The first formula is the null tail (4.4) scanned in n. The second is the same tail sum evaluated with the signal switched on, because the power is the rejection probability under the alternative.

Example: 3σ and 5σ with $b = 3.2$. The script finds $n_{\text{crit}} = 11$ for 3σ and $n_{\text{crit}} = 17$ for 5σ. So Cowan's 11 events on 3.2 (Section 4.1.4) just reach 3σ. Figure 4.10 shows the power as a function of s. A signal of about 7.6 events is needed to cross 3σ half of the time, and about 13.6 to cross 5σ half of the time. With $s = 20$ the 5σ test fires with probability 0.924, and a direct simulation of 20 000 experiments gives 0.925.

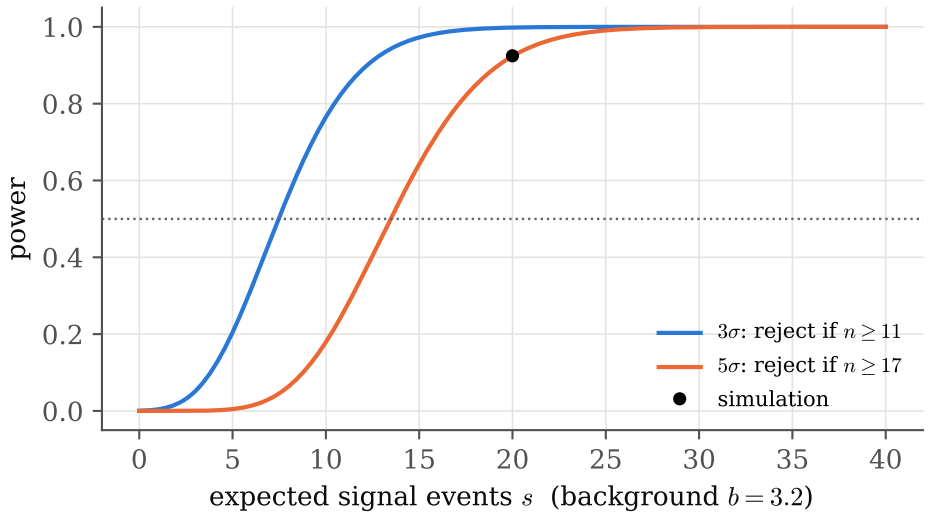


Figure 4.10: Power of a counting experiment with background $b = 3.2$, for the 3σ and the 5σ thresholds. Each curve rises from nearly zero to one as the expected signal grows. The dotted line marks power one half, and the black dot is a direct simulation at $s = 20$.

With discrete data a test usually has level α without having size exactly α: its false-alarm rate is below the target, and we call it *conservative*. The reason is that N takes only integer values, so the rejection probability under the null jumps as the cut moves from one integer to the next. For example, the size of the test "reject if $N \geq 11$" is $\mathrm{P}(N \geq 11 \mid b = 3.2) = 4.972 \times 10^{-4}$, smaller than the 3σ target 0.00135, and the next lower cut, $N \geq 10$, would overshoot it.

Listing 4.3: Critical count and power of a Poisson counting experiment (`code/ch04/03_power_curves.py`)

```
# ---------- (iii) counting experiment b = 3.2 ----------
b = 3.2
def n_crit(p_thr):
    """smallest n with P(N >= n | b) <= p_thr"""
    n = 0
```

```
    while stats.poisson.sf(n - 1, b) > p_thr:
        n += 1
    return n
p3, p5 = stats.norm.sf(3), stats.norm.sf(5)
n3, n5 = n_crit(p3), n_crit(p5)
s = np.linspace(0, 40, 201)
pow3 = stats.poisson.sf(n3 - 1, b + s)
pow5 = stats.poisson.sf(n5 - 1, b + s)
s50_3 = s[np.argmax(pow3 >= 0.5)]
s50_5 = s[np.argmax(pow5 >= 0.5)]
s_chk = 20.0
```

The helper `n_crit` steps n upward until the null tail `stats.poisson.sf(n - 1, b)` drops below the threshold, which is (4.12) read literally. The power curves are then the same survival function with mean `b + s`. Nothing here needs an approximation, which makes the counting experiment the best place to build intuition about power before we meet statistics whose distributions must be simulated.

Statistically significant is not the same as important. The power function of a consistent test climbs to one for *every* $\theta \in \Theta_1$ as $n \to \infty$, however close θ is to θ_0. With enough data, an effect too small to matter will be declared significant. [Wasserman p. 155] calls this the gap between statistical and scientific significance: a rejection says the effect is probably not zero, not that it is large. A confidence interval shows both at once, the distance from θ_0 and the precision, and is often the more informative report. Every test can be turned into a confidence interval, a construction taken up in the second part of the chapter.

4.4 The p-value is a random variable

A p-value changes from one repetition of an experiment to the next. Imagine a thousand groups, each running the same background-only experiment and each computing a p-value. They will not get the same number. The data fluctuate, so the test statistic fluctuates, so the p-value fluctuates [Cowan p. 58]. The p-value is therefore a random variable with its own distribution. That distribution is what gives "reject if $p < \alpha$" a known false-alarm rate. It also shows, more clearly than any argument in words, why a p-value is not the probability that the null is true.

The question in words.

1. How is the p-value related to the rejection regions of Section 4.3?
2. If the null is true, how are p-values distributed? If it is false?
3. In a collection of experiments where some nulls are true and some false, what fraction of the "significant" results are false alarms?

4.4.1 The smallest level at which we would reject

Suppose that for every $\alpha \in (0, 1)$ we have a size-α test, with rejection region $R_\alpha = \{T \geq c_\alpha\}$. A smaller α means a stricter test, a larger c_α and a smaller region, so the regions are nested. For our data, some of these tests reject and some do not, and there is a smallest level at which the test still rejects. [Wasserman Def. 10.11] takes that as the definition,

$$p = \inf\{\alpha : T(\boldsymbol{x}) \in R_\alpha\},$$

and [Wasserman Thm 10.12] shows that it is the same tail probability as before. For a simple null the argument is short. The observed value t_{obs} is on the boundary of the region exactly when $c_\alpha = t_{\text{obs}}$, and the size of that region is $\mathrm{P}_{\theta_0}(T \geq t_{\text{obs}})$. Any smaller α puts the cut above t_{obs} and retains; any larger one rejects. So

$$\text{reject at level } \alpha \quad \Longleftrightarrow \quad p \leq \alpha. \tag{4.13}$$

This is why a p-value is more useful than a bare "reject": every reader can apply their own α. For a composite null the tail probability depends on which $\theta \in \Theta_0$ is true, and the definition takes the worst case, $p = \sup_{\theta \in \Theta_0} \mathrm{P}_\theta(T \geq t_{\text{obs}})$, as we already did for the coin in Section 4.2.1.

A widely used vocabulary for p-values [Wasserman p. 157]:

p-value	evidence against H_0
< 0.01	very strong
0.01–0.05	strong
0.05–0.10	weak
> 0.10	little or none

Particle physicists would call $p = 0.01$ an unremarkable fluctuation, for reasons we come to in Section 4.9. The scale is a convention, and conventions differ between fields.

4.4.2 Under the null, p is uniform

Let T have a continuous distribution under H_0, with cumulative distribution function $F_0(t) = \mathrm{P}_{\theta_0}(T \leq t)$, increasing and continuous. For an observed value t, the p-value is the upper tail, $p = 1 - F_0(t)$. Now let the data be random, so that the p-value is the random variable $P = 1 - F_0(T)$. For any $u \in (0, 1)$,

$$\begin{aligned} \mathrm{P}_{\theta_0}(P \leq u) &\overset{(1)}{=} \mathrm{P}_{\theta_0}\big(1 - F_0(T) \leq u\big) \\ &\overset{(2)}{=} \mathrm{P}_{\theta_0}\big(F_0(T) \geq 1 - u\big) \\ &\overset{(3)}{=} \mathrm{P}_{\theta_0}\big(T \geq F_0^{-1}(1-u)\big) \\ &\overset{(4)}{=} 1 - F_0\big(F_0^{-1}(1-u)\big) \\ &\overset{(5)}{=} u. \end{aligned} \tag{4.14}$$

(1) Definition of P.
(2) Rearrange the inequality.
(3) F_0 is increasing, so applying F_0^{-1} to both sides keeps the direction of the inequality.
(4) $\mathrm{P}(T \geq a) = 1 - F_0(a)$ for a continuous variable, since $\mathrm{P}(T = a) = 0$.
(5) F_0 undoes F_0^{-1}.

A variable with $\mathrm{P}(P \leq u) = u$ for all u is uniform on $(0, 1)$ [Wasserman Thm 10.14]. The fact behind it is the **probability integral transform** [feeding a continuous random variable into its own CDF gives a uniform variable]. Run backwards, the same fact turns uniform random numbers into any distribution we like, by inversion (Chapter 6 uses it to simulate). Put together with (4.13), which says "reject at level α" is the same as "$p \leq \alpha$", it gives a test whose false-alarm probability is exactly α. So when nothing is there, a p-value is just a uniform random number. Values below 0.3 occur 30% of the time and values below 0.03 occur 3% of the time: both are things that happen, one ten times more often than the other.

When the alternative is true, the statistic tends to be larger, the tail beyond it smaller, and the p-values pile up near zero. How strongly they pile up is the power: by (4.13), $\mathrm{P}_\theta(P \leq \alpha) = \beta(\theta)$.

Example: p-values from Gaussian readings. Take $n = 10$ readings of $\mathcal{N}(\mu, 1)$ and the one-sided test of $\mu = 0$, with $p = 1 - \Phi(\sqrt{n}\,\bar{x})$. The script `code/ch04/04_pvalue_uniform.py` computes 200 000 such p-values for each of three true means. The left panel of Figure 4.11 shows the histograms. At $\mu = 0$ the histogram is flat, a fraction 0.0494 of the p-values fall below 0.05, and a Kolmogorov–Smirnov test of uniformity (Section 4.8.4) returns $p = 0.72$, no evidence against flatness. At $\mu = 0.3$ a fraction 0.242 fall below 0.05, and at $\mu = 0.8$ a fraction 0.813: these are the powers of the test at those means.

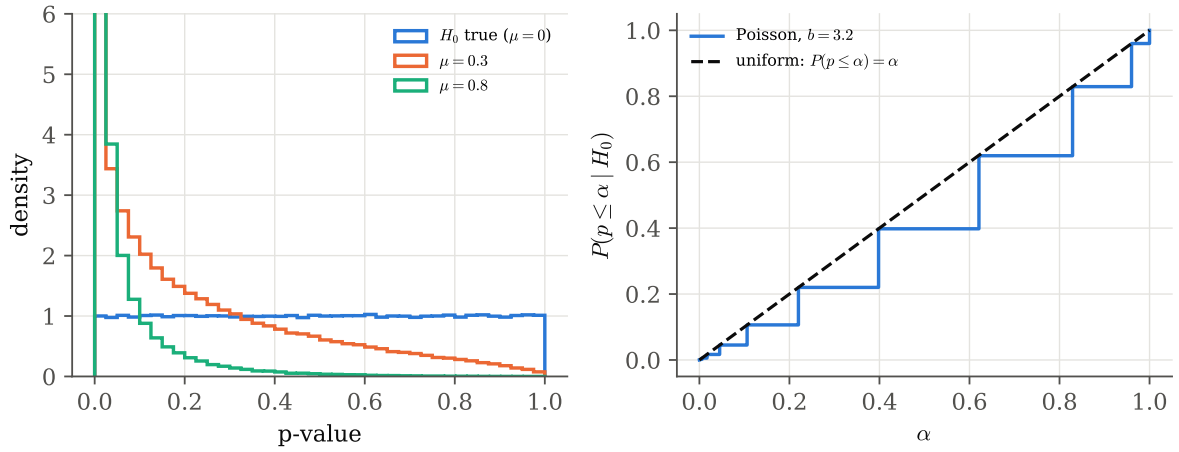


Figure 4.11: Left: histograms of p-values from many repetitions of the same experiment. With the null true (blue) the histogram is flat; with a real effect (orange, green) the p-values crowd towards zero, more so for the larger effect. Right: for Poisson counts with $b = 3.2$ the p-value takes only a few values, so its cumulative distribution is a staircase. The staircase touches the diagonal at its corners and otherwise stays below it, so "reject if $p \le \alpha$" never rejects more often than α.

4.4.3 Discrete data: valid, not uniform

For a count the argument breaks at step (4) of (4.14), because $\mathrm{P}(T = a)$ is no longer zero. The p-value can only take the values $\pi(n) = \mathrm{P}_0(N \ge n)$ for integer n, so it cannot be uniform. What survives is the inequality that matters for error control, the definition of a **valid p-value** [one that is at most α with probability at most α under the null, for every α] [CasellaBerger Def. 8.3.26]:

$$\mathrm{P}_\theta(P \le u) \le u \qquad \text{for every } \theta \in \Theta_0 \text{ and every } u \in [0, 1].$$

For a Poisson count the proof takes one line. The tail $\pi(n)$ decreases as n grows. Let n_u be the smallest integer with $\pi(n_u) \le u$. Then

$$\mathrm{P}_0(P \le u) \overset{(1)}{=} \mathrm{P}_0\big(\pi(N) \le u\big) \overset{(2)}{=} \mathrm{P}_0(N \ge n_u) \overset{(3)}{=} \pi(n_u) \overset{(4)}{\le} u. \tag{4.15}$$

(1) The p-value of an observed count N is $\pi(N)$.
(2) Because π is decreasing, $\pi(N) \le u$ holds exactly when $N \ge n_u$.
(3) Definition of π.
(4) Definition of n_u.

Equality holds only when u happens to be one of the attainable values $\pi(n)$, which are the corners of the staircase in the right panel of Figure 4.11. With $b = 3.2$ the simulated fraction of null experiments with $p \le 0.05$ is 0.0453, below 0.05. For a composite null, the supremum over Θ_0 in the definition of p keeps it valid for every member of the null [CasellaBerger Thm 8.3.27].

A large p-value is not evidence for the null. A large p can arise because H_0 is true, or because H_0 is false and the test has too little power to notice [Wasserman p. 157]. In the example above, at $\mu = 0.3$ three quarters of the p-values exceed 0.05 although the null is false. Failing to reject says "these data cannot tell". Testing is built to find evidence *against* a null, not to prove one [Wasserman p. 161].

4.4.4 What a p-value is not

The most common misreading of a p-value is "$p = 0.04$, so there is a 4% chance that the null is true". [Wasserman p. 157] and [Cowan p. 58] both warn against it. The two numbers answer questions in opposite directions:

$$p = \mathrm{P}\big(\text{data this extreme or more} \,|\, H_0\big) \qquad \text{versus} \qquad \mathrm{P}\big(H_0 \,|\, \text{data}\big).$$

The first is a prediction from the null; the second is an inference about it. Turning one into the other takes Bayes' theorem, and Bayes' theorem needs a number the p-value never uses: how often the null is true before we look. A simulation makes the difference concrete.

Example: a population of experiments. *The setting.* Picture a field in which 90% of the hypotheses people test are really null, $\pi_0 = 0.9$. The other 10% have an effect $\mu = 0.5$ in the same experiment as above: $n = 10$ Gaussian readings of unit variance, tested one-sided against $\mu = 0$. We simulate 10^6 experiments, each picking null or effect at random with these frequencies, and compute every p-value.

The question. Among the "significant" experiments, those with $p < 0.05$, what fraction had a true null?

The two routes to a significant result. A result can be significant because the null is true and chance gave a small p, which happens with probability 0.05 by uniformity. Or it can be significant because there is an effect and the test caught it, which happens with probability equal to the power at $\mu = 0.5$, 0.473.

Only then the numbers. Weighting each route by how often it is taken, Bayes' theorem gives

$$\begin{aligned} \mathrm{P}(H_0 \,|\, p < 0.05) &\overset{(1)}{=} \frac{\mathrm{P}(p < 0.05 \,|\, H_0)\,\pi_0}{\mathrm{P}(p < 0.05 \,|\, H_0)\,\pi_0 + \mathrm{P}(p < 0.05 \,|\, H_1)\,(1-\pi_0)} \\ &\overset{(2)}{=} \frac{0.05 \times 0.9}{0.05 \times 0.9 + 0.473 \times 0.1} \approx 0.487. \end{aligned} \tag{4.16}$$

(1) Bayes' theorem with the two hypotheses as the branches of a probability tree: first choose null or effect, then the p-value.

(2) Under the null P is uniform, so $\mathrm{P}(P < 0.05) = 0.05$; under the effect it is the power.

Almost half of the significant results are false alarms; the simulation counts a fraction 0.489. Now narrow the selection to results that only just made it, $0.04 < p < 0.05$. Under the null this band has probability 0.01 (uniformity again), and under the effect it has a probability we compute from (4.9). The same formula gives 0.682, and the simulation 0.681: two thirds of the barely significant results come from true nulls, while each of them reports "$p < 0.05$". Figure 4.12 shows why: the flat carpet of null p-values is ten times more populous than the effect experiments, and near $p = 0.05$ the carpet dominates.

Listing 4.4: A population of experiments: how many significant results are false alarms? (`code/ch04/04_pvalue_uniform.py`)

```
# ---------- (iii) a population of experiments ----------
```

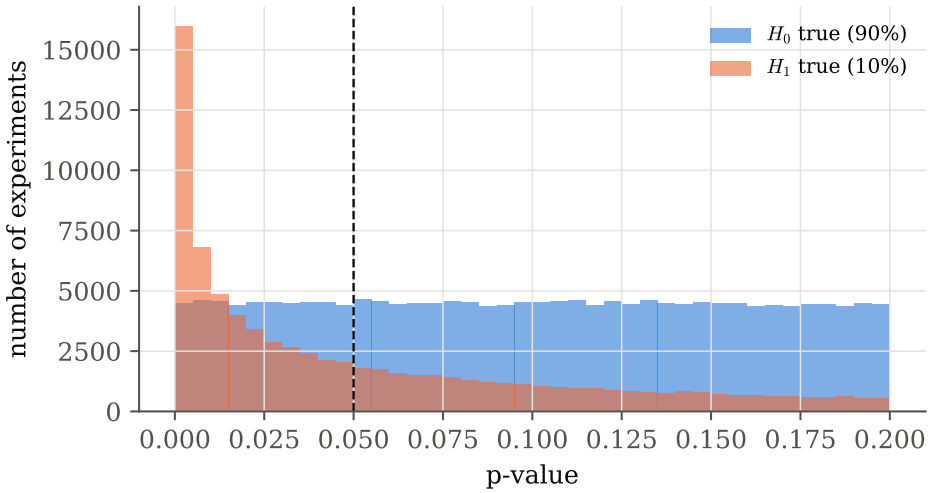


Figure 4.12: p-values from a population of experiments in which 90% of the nulls are true. The blue histogram (true nulls) is flat; the orange one (real effects) rises towards zero. Left of the dashed line at $p = 0.05$ both contribute, and just left of it the flat blue carpet is the larger.

```
pi0 = 0.9                                                    # fraction of experiments
    where H0 is true
mu_eff = 0.5                                                 # effect size when H0 is
    false
Mpop = 1_000_000
is_null = rng.random(Mpop) < pi0
mu = np.where(is_null, 0.0, mu_eff)
pp = stats.norm.sf(np.sqrt(n) * rng.normal(mu, sigma / np.sqrt(n)) / sigma)
power = np.mean(pp[~is_null] < 0.05)
sig = pp < 0.05
frac_null_sig = np.mean(is_null[sig])
band = (pp > 0.04) & (pp < 0.05)
frac_null_band = np.mean(is_null[band])
# the same, analytically: P(H0 | p in band) = pi0 P(band|H0) / [pi0 P(band|H0) + (1-pi0)
    P(band|H1)]
zlo, zhi = stats.norm.isf(0.05), stats.norm.isf(0.04)
d = np.sqrt(n) * mu_eff / sigma
PbandH1 = stats.norm.cdf(zhi - d) - stats.norm.cdf(zlo - d)
frac_null_band_th = pi0 * 0.01 / (pi0 * 0.01 + (1 - pi0) * PbandH1)
frac_null_sig_th = pi0 * 0.05 / (pi0 * 0.05 + (1 - pi0) * stats.norm.sf(zlo - d))
```

The code draws one Boolean per experiment, `is_null`, with probability π_0, and sets the true mean to 0 or 0.5 accordingly. It draws the sample mean directly from $\mathcal{N}(\mu, \sigma^2/n)$, which is its exact distribution, converts it to a p-value, and then simply counts: among the experiments with `pp < 0.05`, what fraction have `is_null`? The analytic lines below are (4.16) and its narrow-band version.

Watch out: p is not $\mathrm{P}(H_0 \mid \text{data})$

The p-value is computed assuming the null is true, so it cannot be the probability that the null is true. The fraction of significant results that are false alarms depends on the power and on how often nulls are true in the first place, neither of which enters p. In the example, $p < 0.05$ went with a 0.489 chance of a true null, and $0.04 < p < 0.05$ with a 0.681 chance. The inference direction is the subject of Chapter 5.

This does not make p-values useless. It makes them one ingredient. A small p-value says

the data are hard to produce by chance alone; whether that makes the alternative probable depends on how plausible it was to begin with. That is the logic behind the 5σ convention of Section 4.9, and it is exactly the prior-times-likelihood structure of Bayes' theorem.

4.5 Three workhorse tests: Wald, Student's t and permutation

A calorimeter is calibrated twice, a week apart, on the 661.7 keV gamma line of ^{137}Cs. The first run records 12 reconstructed peak positions, the second 15. The second run's mean is higher. Did the energy scale drift, or is this the scatter we expect from a detector with a resolution of a few keV? Situations like this, comparing two samples or comparing an estimate with a reference value, are the everyday business of testing. Most tests for them are one of three kinds.

1. The *Wald test*: divide the estimate's distance from the null value by its estimated standard error and compare with a normal distribution. It works for any estimator that is approximately normal.
2. The *t-test*: the same ratio, with the exact distribution that accounts for having estimated the standard error from few Gaussian readings.
3. The *permutation test*: no distributional assumption at all; the null distribution is built by shuffling the labels of the data.

Most powerful tests in the sense of Section 4.6 often do not exist for composite hypotheses, so these general-purpose tests carry most of the load [Wasserman p. 152].

4.5.1 The Wald test

Let $\hat\theta$ be an estimate of a scalar parameter θ with estimated standard error $\widehat{\mathrm{se}}$, and suppose $\hat\theta$ is approximately normal, as maximum-likelihood estimators are for large samples (Chapter 3). To test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ we measure the distance in standard errors,

$$W = \frac{\hat\theta - \theta_0}{\widehat{\mathrm{se}}}, \qquad \text{reject } H_0 \text{ when } |W| > z_{\alpha/2} \tag{4.17}$$

[Wasserman Def. 10.3]. Under the null, W is approximately standard normal, so

$$\mathrm{P}_{\theta_0}\big(|W| > z_{\alpha/2}\big) \longrightarrow \mathrm{P}\big(|Z| > z_{\alpha/2}\big) = \alpha \qquad (n \to \infty), \tag{4.18}$$

which is [Wasserman Thm 10.4]: the Wald test has size α asymptotically. The p-value of an observed w is the two-sided normal tail [Wasserman (10.7)]:

$$p = \mathrm{P}_{\theta_0}\big(|W| > |w|\big) \approx \mathrm{P}\big(|Z| > |w|\big) = 2\Phi(-|w|). \tag{4.19}$$

The standard error may also be computed at the null value, se_0, in place of the estimate; both versions are valid [Wasserman Remark 10.5]. For a binomial proportion $\hat p$ tested against p_0, for instance, one may use $\mathrm{se}_0 = \sqrt{p_0(1-p_0)/n}$, which needs no estimate at all.

Example: comparing two means. Two independent samples $X_1, \dots, X_m$ and $Y_1, \dots, Y_n$ have means μ_1, μ_2. The null is $\delta = \mu_1 - \mu_2 = 0$, the estimate is $\hat\delta = \bar X - \bar Y$, and because the samples are independent the variances add [Wasserman Ex. 10.8]:

$$\widehat{\mathrm{se}} = \sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}, \qquad W = \frac{\bar X - \bar Y}{\sqrt{s_1^2/m + s_2^2/n}}, \tag{4.20}$$

with s_1^2, s_2^2 the two sample variances. In Wasserman's cholesterol data [Wasserman Ex. 10.15] the two group means are 216.2 and 195.3 with standard errors 5 and 2.4, so $W = (216.2 - 195.3)/\sqrt{5^2 + 2.4^2} = 20.9/5.55 \approx 3.78$ and $p = 2\Phi(-3.78) = 0.00016$, the quoted 0.0002. The same recipe works for medians, with the standard error of the difference of sample medians taken from the bootstrap of Chapter 3 [Wasserman Ex. 10.9]; there $W = 2.4$ and $p = 0.02$.

Example: two classifiers, and paired data. [Wasserman Ex. 10.7] compares two prediction algorithms, and the same logic compares two event classifiers. If each is run on its own test set, of sizes m and n, the error counts are independent binomials and (4.20) applies with $s_i^2 = \hat{p}_i(1 - \hat{p}_i)$. If both run on the *same* test set, the two error rates are correlated and the variances no longer add. The fix is to test the per-event difference: let $X_i = 1$ if the first classifier is right on event i, Y_i likewise for the second, and $D_i = X_i - Y_i$. Then $\hat{\delta} = \bar{D}$, $\widehat{\text{se}} = S_D/\sqrt{n}$ with S_D the sample standard deviation of the D_i, and $W = \bar{D}/\widehat{\text{se}}$. This **paired comparison** removes the event-to-event variation that both classifiers share, which is usually most of it.

4.5.2 Small samples: Student's t

The Wald test's guarantee is asymptotic. With few readings, the estimated standard error is itself noisy, and a too-small $\widehat{\text{se}}$ inflates $|W|$. We can see how badly by simulation: draw n readings from $\mathcal{N}(0,1)$, so the null $\mu = 0$ is true, compute $T = \sqrt{n}\,\bar{X}/S$ with S the sample standard deviation, and count how often $|T| > 1.96$. Figure 4.13 shows the answer: the false-alarm rate is 0.190 at $n = 3$, 0.122 at $n = 5$, 0.081 at $n = 10$ and 0.059 at $n = 30$, instead of 0.05.

For Gaussian data the exact distribution of T is known. Write

$$T = \frac{\sqrt{n}(\bar{X} - \mu_0)}{S} \overset{(1)}{=} \frac{\sqrt{n}(\bar{X} - \mu_0)/\sigma}{S/\sigma} \overset{(2)}{=} \frac{Z}{\sqrt{V/(n-1)}}, \qquad Z \sim \mathcal{N}(0,1), \quad V = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}. \tag{4.21}$$

(1) Divide numerator and denominator by the unknown σ; the ratio is unchanged, so T does not depend on σ.

(2) Under the null the numerator is standard normal. In the denominator, $S/\sigma = \sqrt{V/(n-1)}$, and $(n-1)S^2/\sigma^2$ is χ^2_{n-1} and independent of $\bar{X}$ for Gaussian data (Chapter 3).

A standard normal divided by the square root of an independent χ^2_k/k is by definition a **Student t variable** with k degrees of freedom [a bell curve with heavier tails, because the denominator sometimes comes out small], with density [Wasserman p. 170]

$$f(t) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\,\Gamma\left(\frac{k}{2}\right)}\left(1 + \frac{t^2}{k}\right)^{-(k+1)/2}. \tag{4.22}$$

For $k = 1$ it is the Cauchy distribution of Chapter 2, and as $k \to \infty$ it becomes the standard normal, because $(1 + t^2/k)^{-(k+1)/2} \to e^{-t^2/2}$ and $V/k \to 1$ by the law of large numbers. So $T \sim t_{n-1}$ exactly, and the **t-test** rejects when $|T| > t_{n-1,\alpha/2}$, the upper $\alpha/2$ quantile of t_{n-1}. The same fact lets us compute, not just simulate, how often the Wald test rejects a true null. The Wald test uses the normal cut 1.96, so its size is $\mathrm{P}(|t_{n-1}| > 1.96)$, which for $n = 3$ is 0.189, the solid line in Figure 4.13. The t-test's simulated size is 0.050 at $n = 3$ and 0.050 at $n = 10$. For moderately large n the two tests are the same test.

For two samples with possibly different variances, the ratio (4.20) is not exactly t-distributed. **Welch's t-test** uses a t distribution with an effective number of degrees of freedom computed from s_1, s_2, m, n; it is what `scipy.stats.ttest_ind(..., equal_var=False)` does.

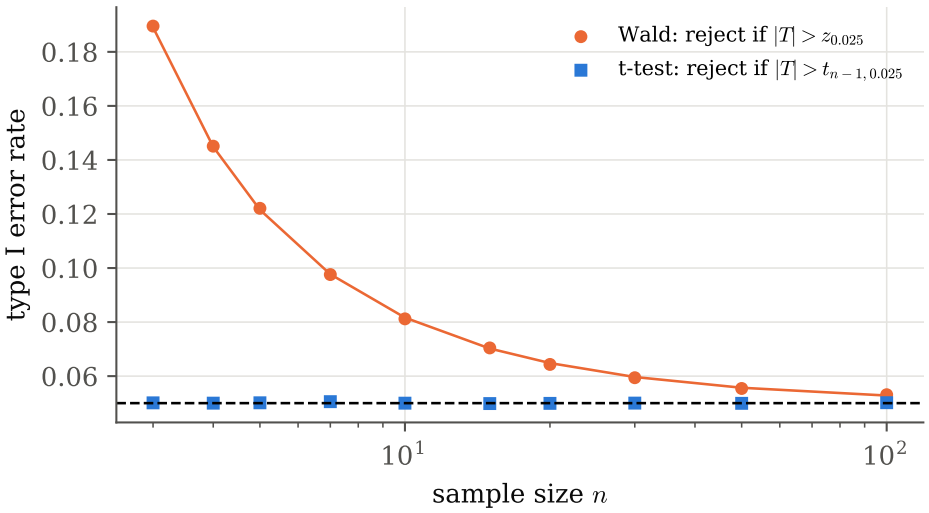


Figure 4.13: How often a test rejects a true null, as a function of the number of Gaussian readings. The Wald test, which uses the normal cut 1.96 with an estimated standard error, rejects too often for small samples (orange dots, with the exact curve as a line). The t-test, which uses the t quantile, stays on the dashed 0.05 line at every sample size (blue squares).

4.5.3 The permutation test

The t-test is exact only for Gaussian data. The **permutation test** is exact for any distribution, at the price of a computation [Wasserman §10.5]. The null it tests is that the two samples come from the same distribution, $H_0 : F_X = F_Y$. The key observation: if that is true, the labels "first sample" and "second sample" carry no information. Every way of splitting the pooled $N = m + n$ values into a group of m and a group of n was equally likely to have produced the data we saw. So we can build the null distribution of any statistic T by recomputing it for all $N!$ orderings of the pooled values, each with probability $1/N!$. That is the **permutation distribution**, and the p-value is the fraction of orderings with T larger than the observed value.

Example: the smallest possible permutation test. Take $(X_1, X_2, Y_1) = (1, 9, 3)$ and $T = \left|\bar{X} - \bar{Y}\right|$ [Wasserman Ex. 10.19]. The observed value is $|(1 + 9)/2 - 3| = 2$. The six orderings, with the first two entries playing the role of X:

ordering	T	probability
$(1,9,3), (9,1,3)$	$\lvert 5-3\rvert = 2$	$2/6$
$(1,3,9), (3,1,9)$	$\lvert 2-9\rvert = 7$	$2/6$
$(3,9,1), (9,3,1)$	$\lvert 6-1\rvert = 5$	$2/6$

Four of the six give $T > 2$, so $p = \mathrm{P}_0(T > 2) = 4/6 = 0.667$. The script enumerates the six orderings with `itertools.permutations` and gets the same number. (Some authors count $T \geq t_{\text{obs}}$, which includes the observed ordering itself; with continuous data and many orderings the difference is negligible.)

For real data $N!$ is astronomically large, so we sample orderings at random. The algorithm [Wasserman p. 163] is:

1. compute t_{obs} from the data as labelled;
2. shuffle the pooled data, split into groups of m and n, and recompute T;

3. repeat step 2 B times, giving $T_1, \dots, T_B$;
4. the p-value is approximately $\frac{1}{B}\sum_{j=1}^{B} I(T_j > t_{\text{obs}})$, with I the indicator (one if true, zero if false).

Why is this legitimate? Under the null the N values are exchangeable: their joint distribution does not change when we reorder them. Given the set of values that occurred, every assignment of labels is equally likely, so the shuffled statistics are draws from the exact conditional null distribution of T. No normality, no large-sample limit, no estimate of a standard error is involved. The permutation test is most useful when there are few samples and the statistic has no convenient formula for its distribution. Wasserman's gene-expression example is of this kind ([Wasserman Ex. 10.20]: ten patients, two tumour types, T the absolute difference of medians = 710, $p = 0.045$).

Example: the calorimeter calibration. Back to the two calibration runs. We simulate run 1 as 12 readings from $\mathcal{N}(661.7, 4^2)$ keV and run 2 as 15 readings with a true shift of 3.3 keV (script `05_wald_t_permutation.py`). The observed means are 662.07 and 665.67 keV with sample standard deviations 4.30 and 3.55 keV. By (4.20),

$$\begin{aligned} \hat{\delta} &= 665.67 - 662.07 = 3.60 \text{ keV}, \\ \widehat{\text{se}} &= \sqrt{4.30^2/12 + 3.55^2/15} = 1.54 \text{ keV}, \\ W &= 3.60/1.54 = 2.33, \end{aligned}$$

so the Wald p-value is $2\Phi(-2.33) = 0.0198$. Welch's t-test gives 0.0298, larger because it accounts for the uncertainty in the two standard deviations. The permutation test, with $B = 100\,000$ shuffles of the 27 values, gives 0.0251. Figure 4.14 shows the permutation distribution of $\bar{x}_2 - \bar{x}_1$ with the observed difference marked. The three p-values are close to one another: they ask the same question in three ways. None of them would convince a particle physicist of a new effect. Nor could they be expected to. With twelve and fifteen readings of 4 keV resolution, the true shift of 3.3 keV is only $3.3/\sqrt{4^2/12 + 4^2/15} \approx 2.1$ standard errors on average, and a single pair of runs can easily show much less or much more. This experiment is too small to find the shift reliably.

Listing 4.5: Wald, Welch and permutation tests for two calibration runs (`code/ch04/05_wald_t_permutation.py`)

```
m, n = 12, 15
run1 = rng.normal(661.7, 4.0, size=m)                    # Cs-137 line, keV, resolution 4 keV
run2 = rng.normal(665.0, 4.0, size=n)                    # a 3.3 keV scale shift in run 2
d_hat = run2.mean() - run1.mean()
se_hat = np.sqrt(run1.var(ddof=1) / m + run2.var(ddof=1) / n)
W = d_hat / se_hat
p_wald = 2 * stats.norm.sf(abs(W))
p_welch = stats.ttest_ind(run2, run1, equal_var=False).pvalue
pooled = np.concatenate([run1, run2])
B = 100_000
Tperm = np.empty(B)
for j in range(B):
    perm = rng.permutation(pooled)
    Tperm[j] = perm[m:].mean() - perm[:m].mean()
p_perm = np.mean(np.abs(Tperm) >= abs(d_hat))
p_perm_scipy = stats.permutation_test((run2, run1), lambda a, b: a.mean() - b.mean(),
                                      n_resamples=20_000, alternative="two-sided",
                                      random_state=rng).pvalue
```

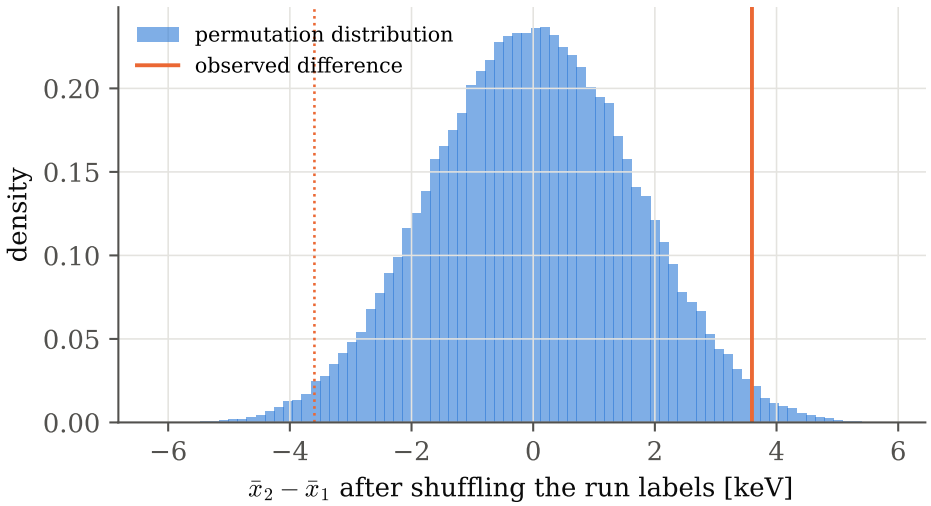


Figure 4.14: The permutation distribution for the calibration example: the difference of the two run means after randomly reassigning the 27 readings to the two runs, 100 000 times. The solid orange line is the observed difference and the dotted one its mirror image; the fraction of shuffles beyond either line is the two-sided p-value.

The loop is the algorithm above, line for line: `rng.permutation` shuffles the pooled readings, the first m play run 1 and the rest run 2, and the difference of means is stored. The p-value counts shuffles with $|T_j| \geq |t_{\text{obs}}|$ (two-sided). As a cross-check the script calls `scipy.stats.permutation_test` with 20 000 resamples, which returns 0.0273. Both numbers are Monte Carlo estimates, so they carry a random error. For the shorter run of 20 000 resamples that error is $\sqrt{p(1-p)/20\,000}$, a few thousandths at most. The two estimates differ by 1.8 times their combined error; more shuffles would bring them together. In large samples the permutation test usually agrees with the asymptotic tests, and it is most valuable for small samples [Wasserman p. 164].

Tests and intervals are two views of one object. There is a simple link between the Wald test and the familiar interval $\hat{\theta} \pm z_{\alpha/2}\widehat{\text{se}}$. By (4.17), $|\hat{\theta} - \theta_0| > z_{\alpha/2}\widehat{\text{se}}$ says precisely that θ_0 lies outside that interval. So the size-α Wald test rejects θ_0 if and only if θ_0 is outside the $1-\alpha$ interval [Wasserman Thm 10.10]. The second part of the chapter turns that observation into a general machine for building confidence intervals from tests.

4.6 The best test: the Neyman–Pearson lemma

A trigger has to decide, event by event, whether a collision looks like signal or like background. Each event comes with several measured quantities: energies, angles, isolation variables. Any selection keeps some signal and some background. At a fixed signal efficiency, say 90%, one selection keeps less background than another, and we would like the best one. Put as a question about tests: among all tests with the same size α, which has the largest power? For two simple hypotheses there is a complete answer, and it is one of the few genuinely optimal results in statistics.

The question in words.

1. We have a fixed budget α of false-alarm probability. Which outcomes should we spend it

on?

2. Does the answer depend on the particular alternative, or is one test best against a whole family of alternatives?

3. What do we do when the densities are only available as simulated events in many dimensions?

4.6.1 Spending a false-alarm budget

Take a simple null and a simple alternative, $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, with densities $f_0(\boldsymbol{x}) = f(\boldsymbol{x} \mid \theta_0)$ and $f_1(\boldsymbol{x}) = f(\boldsymbol{x} \mid \theta_1)$ for the whole data set. A test is a choice of rejection region R. Its size and power are

$$\alpha = \int_R f_0(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, \qquad \beta(\theta_1) = \int_R f_1(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.$$

Think of each small cell $\mathrm{d}\boldsymbol{x}$ of data space as an item we can put into R. It costs $f_0\,\mathrm{d}\boldsymbol{x}$ of the false-alarm budget and buys $f_1\,\mathrm{d}\boldsymbol{x}$ of power. The budget is fixed. To buy as much power as possible we should take the items with the best value for money first, the cells where the ratio f_1/f_0 is largest, and keep taking them in order of decreasing ratio until the budget is spent. The region we end up with is $R = \{\boldsymbol{x} : \ f_1(\boldsymbol{x})/f_0(\boldsymbol{x}) > k\}$, with the threshold k fixed by the budget.

The same answer comes out of the Lagrange-multiplier method familiar from mechanics. We maximise $\int_R f_1$ subject to $\int_R f_0 = \alpha$, so we maximise $\int_R (f_1 - k f_0)\,\mathrm{d}\boldsymbol{x}$ for a multiplier k. The integrand can be positive or negative, and the integral over R is largest when R contains every point where $f_1 - k f_0 > 0$ and no point where it is negative. The multiplier is the exchange rate between power and size.

Key idea: Neyman–Pearson lemma

For testing a simple $H_0 : \theta = \theta_0$ against a simple $H_1 : \theta = \theta_1$, the test that rejects when the **likelihood ratio**

$$\Lambda(\boldsymbol{x}) = \frac{f(\boldsymbol{x} \mid \theta_1)}{f(\boldsymbol{x} \mid \theta_0)} = \frac{\mathcal{L}(\theta_1)}{\mathcal{L}(\theta_0)}$$

exceeds a constant k, with k chosen so that $\mathrm{P}_{\theta_0}(\Lambda > k) = \alpha$, has the largest power of all tests of level α [Wasserman Thm 10.30], [CasellaBerger Thm 8.3.12], [Cowan (4.12)]. The ratio collapses any number of measured variables into one number, and no other function of the data does better.

Why it is true. The budget argument is the proof in disguise. To make it exact, describe a test by its **test function** $\phi(\boldsymbol{x})$ [one where the test rejects, zero where it retains], so that size and power are $\int \phi f_0$ and $\int \phi f_1$. Let ϕ be the likelihood-ratio test and ϕ' any other test of level α, with power functions β and β'. At every point, $[\phi(\boldsymbol{x}) - \phi'(\boldsymbol{x})][f_1(\boldsymbol{x}) - k f_0(\boldsymbol{x})] \geq 0$: where $f_1 > k f_0$ we have $\phi = 1 \geq \phi'$, and where $f_1 < k f_0$ we have $\phi = 0 \leq \phi'$, so the two factors never have opposite signs. Integrating,

$$\begin{aligned} 0 &\overset{(1)}{\leq} \int \big[\phi(\boldsymbol{x}) - \phi'(\boldsymbol{x})\big]\big[f_1(\boldsymbol{x}) - k f_0(\boldsymbol{x})\big]\mathrm{d}\boldsymbol{x} \\ &\overset{(2)}{=} \big[\beta(\theta_1) - \beta'(\theta_1)\big] - k\big[\beta(\theta_0) - \beta'(\theta_0)\big] \\ &\overset{(3)}{\leq} \beta(\theta_1) - \beta'(\theta_1). \end{aligned} \tag{4.23}$$

(1) The integrand is non-negative at every point, as just argued.
(2) Multiply out: $\int \phi f_1 = \beta(\theta_1)$, $\int \phi f_0 = \beta(\theta_0)$, and likewise for ϕ'.
(3) $\beta(\theta_0) = \alpha$ for the likelihood-ratio test, and $\beta'(\theta_0) \le \alpha$ because ϕ' has level α; so the bracket multiplying $k \ge 0$ is non-negative, and dropping it can only increase the right-hand side.

So $\beta(\theta_1) \ge \beta'(\theta_1)$: no level-$\alpha$ test beats the likelihood ratio [CasellaBerger p. 315]. For discrete data replace integrals by sums. Notice what the proof used: nothing about the shape of the densities, the dimension of $\boldsymbol{x}$, or large samples. The lemma is exact.

Example: the Gaussian mean, and a uniformly best test. Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, σ known, and test $\mu = 0$ against a single $\mu = \mu_1 > 0$. The log of the likelihood ratio is

$$\begin{aligned} \ln \Lambda &\overset{(1)}{=} \sum_{i=1}^{n} \Big[-\frac{(x_i - \mu_1)^2}{2\sigma^2} + \frac{x_i^2}{2\sigma^2} \Big] \\ &\overset{(2)}{=} \sum_{i=1}^{n} \frac{2\mu_1 x_i - \mu_1^2}{2\sigma^2} \overset{(3)}{=} \frac{n\mu_1}{\sigma^2}\,\bar{x} - \frac{n\mu_1^2}{2\sigma^2}. \end{aligned} \tag{4.24}$$

(1) The likelihood is a product of Gaussians; its log is a sum. The normalisations $1/\sqrt{2\pi\sigma^2}$ are the same under both hypotheses and cancel.
(2) Expand $(x_i - \mu_1)^2 = x_i^2 - 2\mu_1 x_i + \mu_1^2$; the x_i^2 terms cancel.
(3) $\sum_i x_i = n\bar{x}$.

Because $\mu_1 > 0$, $\ln \Lambda$ increases with $\bar{x}$, so "$\Lambda > k$" is the same event as "$\bar{x} > c$" for some c. The most powerful test is the one-sided test of Section 4.3.2, with $c = \sigma z_\alpha / \sqrt{n}$ to give size α [CasellaBerger Ex. 8.3.15].

The value of μ_1 has disappeared from the rejection region. Neither the event $\bar{x} > c$ nor the cut $c = \sigma z_\alpha / \sqrt{n}$ mentions it. So the same region is most powerful against *every* $\mu_1 > 0$. A test with this property is **uniformly most powerful**. Such tests exist when the likelihood ratio is monotone in one statistic, as for one-sided problems in exponential families. For two-sided alternatives they usually do not: the best test against $\mu_1 > 0$ rejects for large $\bar{x}$, the best against $\mu_1 < 0$ rejects for small $\bar{x}$, and no single region is best against both [Wasserman p. 152].

Example: discreteness restricts the available sizes. Let $X \sim \text{Binomial}(2, \theta)$, testing $\theta = \frac{1}{2}$ against $\theta = \frac{3}{4}$ [CasellaBerger Ex. 8.3.14]. The pmfs are $\binom{2}{x}\theta^x(1-\theta)^{2-x}$, giving $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ and $(\frac{1}{16}, \frac{6}{16}, \frac{9}{16})$ for $x = 0, 1, 2$, so the ratios are

$$\Lambda(0) = \frac{1/16}{1/4} = \frac{1}{4}, \qquad \Lambda(1) = \frac{6/16}{1/2} = \frac{3}{4}, \qquad \Lambda(2) = \frac{9/16}{1/4} = \frac{9}{4}.$$

Ranking the outcomes by Λ and filling the budget: rejecting on $\{2\}$ is the most powerful test of size $\mathrm{P}_{1/2}(X = 2) = \frac{1}{4}$; rejecting on $\{1, 2\}$ is the most powerful test of size $\frac{3}{4}$. No non-randomised test has size 0.05 except the one that never rejects. With discrete data the lemma still says what to do, but only certain sizes can be achieved, as we saw for the Poisson counts in Section 4.3.3.

4.6.2 Selecting particles: efficiency, purity and the ratio

Selecting events is a test, and the lemma tells us how to do it best. [Cowan §4.2] takes a charged particle of known momentum whose ionisation t is measured. Electrons and pions deposit different amounts on average, so t has density $g(t \mid e)$ for electrons and $g(t \mid \pi)$ for pions.

Selecting electron candidates with a cut $t < t_{\text{cut}}$ defines two efficiencies [Cowan (4.3)–(4.4)],

$$\varepsilon_e = \int_{-\infty}^{t_{\text{cut}}} g(t \mid e)\,\mathrm{d}t = 1 - \alpha, \qquad \varepsilon_\pi = \int_{-\infty}^{t_{\text{cut}}} g(t \mid \pi)\,\mathrm{d}t = \beta_{\text{II}}.$$

Here the *signal* plays the role of the null: rejecting "it is an electron" when it is one loses signal, with probability α, and keeping a pion is a type II error. The purity of the selected sample follows from Bayes' theorem with the particle fractions a_e and $a_\pi = 1 - a_e$ as prior probabilities [Cowan (4.8)–(4.11)]:

$$P_e = \frac{\text{electrons kept}}{\text{particles kept}} = \frac{a_e\,\varepsilon_e}{a_e\,\varepsilon_e + a_\pi\,\varepsilon_\pi}. \tag{4.25}$$

This is the same structure as (4.16): the purity of a selection, like the fraction of true discoveries, needs the prior fractions. At fixed efficiency ε_e the purity is largest when ε_π is smallest, which is the Neyman–Pearson problem. With one variable and a monotone ratio a single cut is automatically optimal. With a vector of variables $\boldsymbol{x}$ the lemma says to keep the events with $g(\boldsymbol{x} \mid e)/g(\boldsymbol{x} \mid \pi) > c$ [Cowan (4.12)–(4.13)].

In practice $g(\boldsymbol{x} \mid H)$ is known only through simulated events. With M histogram bins per variable and n variables there are M^n bins to fill, which becomes impossible beyond a few variables [Cowan p. 51]. One retreat is a linear statistic $t = \boldsymbol{a}^{\mathsf{T}}\boldsymbol{x}$. Fisher chose the direction $\boldsymbol{a}$ that maximises the separation of the two class means relative to the spread within the classes, $(\tau_0 - \tau_1)^2/(\Sigma_0^2 + \Sigma_1^2)$, and found [Cowan (4.23)]

$$\boldsymbol{a} \propto \boldsymbol{W}^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1), \qquad \boldsymbol{W} = \boldsymbol{V}_0 + \boldsymbol{V}_1, \tag{4.26}$$

with $\boldsymbol{\mu}_k$ and $\boldsymbol{V}_k$ the mean vector and covariance matrix of $\boldsymbol{x}$ under hypothesis k. When both classes are Gaussian with the *same* covariance $\boldsymbol{V}$, Fisher's discriminant is exactly the log likelihood ratio. Writing $Q_k = (\boldsymbol{x} - \boldsymbol{\mu}_k)^{\mathsf{T}}\boldsymbol{V}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)$ for the quadratic form in each Gaussian exponent,

$$\begin{aligned} \ln \frac{\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_0, \boldsymbol{V})}{\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_1, \boldsymbol{V})} &\overset{(1)}{=} -\tfrac{1}{2}Q_0 + \tfrac{1}{2}Q_1 \\ &\overset{(2)}{=} (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^{\mathsf{T}}\boldsymbol{V}^{-1}\boldsymbol{x} - \tfrac{1}{2}\big(\boldsymbol{\mu}_0^{\mathsf{T}}\boldsymbol{V}^{-1}\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1^{\mathsf{T}}\boldsymbol{V}^{-1}\boldsymbol{\mu}_1\big). \end{aligned} \tag{4.27}$$

(1) The normalisations $(2\pi)^{-n/2}\det \boldsymbol{V}^{-1/2}$ are equal and cancel.

(2) Expand $Q_k = \boldsymbol{x}^{\mathsf{T}}\boldsymbol{V}^{-1}\boldsymbol{x} - 2\boldsymbol{\mu}_k^{\mathsf{T}}\boldsymbol{V}^{-1}\boldsymbol{x} + \boldsymbol{\mu}_k^{\mathsf{T}}\boldsymbol{V}^{-1}\boldsymbol{\mu}_k$ (using the symmetry of $\boldsymbol{V}^{-1}$). The quadratic term in $\boldsymbol{x}$ is the same for both and cancels, leaving a linear function of $\boldsymbol{x}$.

The coefficient vector is $\boldsymbol{V}^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$, proportional to Fisher's with $\boldsymbol{W} = 2\boldsymbol{V}$ [Cowan (4.27)–(4.28)]. When the covariances differ, the $\boldsymbol{x}^{\mathsf{T}}\boldsymbol{V}_k^{-1}\boldsymbol{x}$ terms no longer cancel, the optimal boundary is curved, and a linear statistic loses power. Nonlinear methods, neural networks and boosted trees, are in this light approximations to the likelihood ratio learned from simulated events [Cowan §4.4.2]: a classifier trained to separate two classes outputs, at best, a monotone function of $g(\boldsymbol{x} \mid H_0)/g(\boldsymbol{x} \mid H_1)$.

Example: three selections, and three hundred random ones. We make signal and background Gaussian in two variables (script `06_neyman_pearson.py`), with different means and different covariance matrices, and compare three selections at a signal efficiency of 90%: a cut on x_1 alone, Fisher's linear discriminant (4.26) trained on 20 000 events of each class, and the exact log likelihood ratio. Figure 4.15 shows the events and the three decision boundaries. The fractions of background kept are 0.397 for the x_1 cut, 0.256 for Fisher and 0.177 for the likelihood ratio.

The lemma claims that no selection does better than the likelihood ratio. To check this, the script also draws 300 random quadratic selections $g(\boldsymbol{x}) = \boldsymbol{x}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{x} + \boldsymbol{v}^{\mathsf{T}}\boldsymbol{x} > c$, each tuned to keep

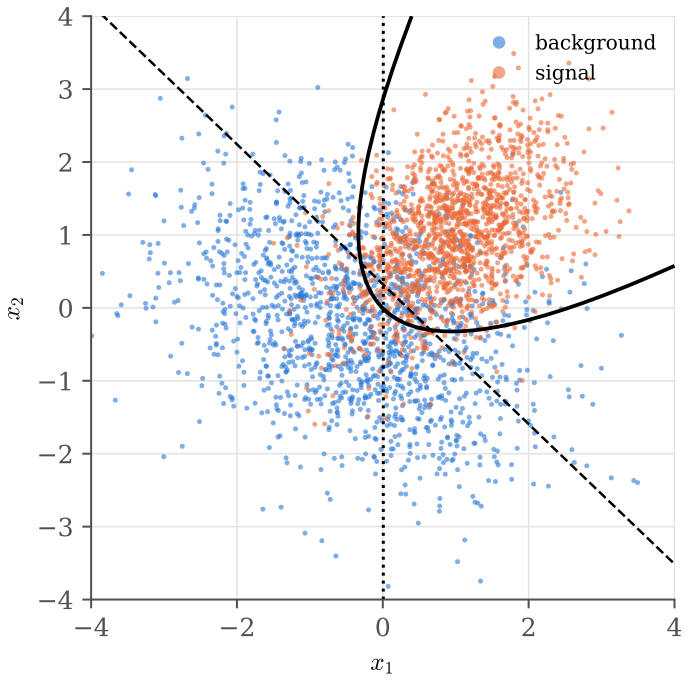


Figure 4.15: Signal (orange) and background (blue) events in two measured variables, with three boundaries that each keep 90% of the signal: a cut on x_1 alone (dotted), Fisher's straight line (dashed), and the curved contour of constant likelihood ratio (solid). The curved boundary follows the shape of the two clouds, which have different orientations, and so admits the least background.

90% of the signal. The best of them keeps a background fraction 0.199, above the likelihood ratio's 0.177. (Because the true ratio here is itself quadratic, a random quadratic can come close, but it cannot win.) Figure 4.16 shows the whole trade-off as a **ROC curve** [background efficiency against signal efficiency, as the cut is varied]. The likelihood-ratio curve lies below the others at every signal efficiency.

Listing 4.6: Three selections compared at equal signal efficiency (code/ch04/06_neyman_pearson.py)

```
def lnr(x):                          # log likelihood ratio  ln[ f(x|signal) / f(x|background) ]
    return S.logpdf(x) - Bk.logpdf(x)

# Fisher linear discriminant from training samples: a proportional to (W_s + W_b)^{-1}
    (mu_s - mu_b)
tr_s, tr_b = xs[:20000], xb[:20000]
Wsum = np.cov(tr_s.T) + np.cov(tr_b.T)
a_fish = np.linalg.solve(Wsum, tr_s.mean(0) - tr_b.mean(0))

stats_s = {"cut on $x_1$ alone": xs[:, 0], "Fisher linear": xs @ a_fish, "likelihood
    ratio": lnr(xs)}
stats_b = {"cut on $x_1$ alone": xb[:, 0], "Fisher linear": xb @ a_fish, "likelihood
    ratio": lnr(xb)}

eff_s = np.linspace(0.01, 0.99, 99)
roc = {}
for name in stats_s:
    thr = np.quantile(stats_s[name], 1 - eff_s)                # keep events with
    statistic > thr
    sb = np.sort(stats_b[name])
    roc[name] = 1 - np.searchsorted(sb, thr) / M              # background efficiency
i90 = np.argmin(np.abs(eff_s - 0.90))
bkg90 = {k: v[i90] for k, v in roc.items()}
```

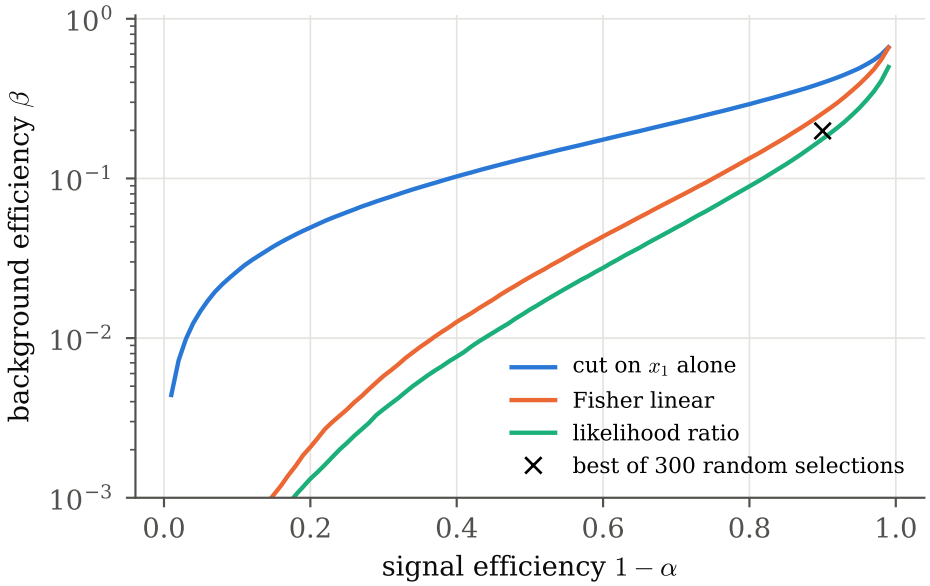


Figure 4.16: Background kept versus signal kept, as the threshold of each selection is varied. Lower is better. The likelihood ratio (green) is lowest everywhere, Fisher's linear discriminant (orange) is next, a cut on one variable (blue) is worst, and the cross marks the best of 300 random quadratic selections at 90% signal efficiency, which still lies above the green curve.

The function `lnr` is the log likelihood ratio, evaluated with the exact densities because we generated the data; in a real analysis they would come from simulation or be learned. Fisher's coefficients come from one line of linear algebra, `np.linalg.solve(Wsum, tr_s.mean(0) - tr_b.mean(0))`, which is (4.26) without forming an inverse. For each statistic the threshold for a given signal efficiency is a quantile of its signal distribution (`np.quantile`), and the background efficiency is the fraction of background above it, found by sorting.

Watch out: Most powerful for one alternative only

The Neyman–Pearson lemma is about *two* completely specified hypotheses. Real alternatives have free parameters: an unknown mass, an unknown coupling. A test tuned to one mass hypothesis can be poor for another, and a uniformly most powerful test rarely exists. The generalisation that handles free parameters, by fitting them, is the likelihood-ratio test of the next section; it is not guaranteed optimal, but it inherits the lemma's logic and is the standard tool.

4.7 Likelihood-ratio tests and Wilks' theorem

The Neyman–Pearson lemma needs two completely specified hypotheses. A real search never has them. In the diphoton bump the background normalisation is known only roughly, the signal strength is unknown, and even the null "no signal" leaves the background level free. Such unknown parameters, which we must allow for but do not want to report, are **nuisance parameters** [parameters of the model that are not the question, like a background level or a calibration constant]. We need a test that works with composite hypotheses, and a way to get its null distribution without simulating every possible value of the nuisance parameters.

The recipe is the natural generalisation of the lemma: in the ratio of likelihoods, give each

hypothesis its best shot by maximising over its free parameters. The surprise, due to Wilks [22], is that the resulting statistic has, for large samples, a null distribution that does not depend on the model at all.

The question in words.

1. How much better does the data fit if we free the parameters that the null fixes?
2. How large must that improvement be before it is surprising, if the null is true?
3. When does the universal answer fail, and what do we do then?

4.7.1 The likelihood-ratio statistic

Let $\mathcal{L}(\boldsymbol{\theta})$ be the likelihood of the data as a function of all parameters $\boldsymbol{\theta} \in \Theta$ (Chapter 3), and let the null restrict them to $\Theta_0 \subset \Theta$. The **likelihood ratio** is

$$\lambda(\boldsymbol{x}) = \frac{\sup_{\boldsymbol{\theta}\in\Theta_0} \mathcal{L}(\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta}\in\Theta} \mathcal{L}(\boldsymbol{\theta})} = \frac{\mathcal{L}(\hat{\boldsymbol{\theta}}_0)}{\mathcal{L}(\hat{\boldsymbol{\theta}})}, \qquad -2\ln\lambda = 2\big[\ln\mathcal{L}(\hat{\boldsymbol{\theta}}) - \ln\mathcal{L}(\hat{\boldsymbol{\theta}}_0)\big], \tag{4.28}$$

where $\hat{\boldsymbol{\theta}}$ is the maximum-likelihood estimate over all of Θ and $\hat{\boldsymbol{\theta}}_0$ the one restricted to Θ_0. Because Θ_0 is part of Θ, the restricted maximum can never be higher, so $0 < \lambda \le 1$ and $-2\ln\lambda \ge 0$. It is zero when the null fits as well as anything, and grows as the unrestricted fit pulls away. We reject for large $-2\ln\lambda$.

Two conventions need a word. First, [Wasserman Def. 10.21] calls the quantity $2\ln[\mathcal{L}(\hat{\boldsymbol{\theta}})/\mathcal{L}(\hat{\boldsymbol{\theta}}_0)]$ itself "λ", so Wasserman's λ is our $-2\ln\lambda$. Second, the denominator maximises over all of Θ rather than over the alternative alone, $\Theta \setminus \Theta_0$. This changes the statistic very little and makes its theory much simpler [Wasserman p. 164].

Key idea: Wilks' theorem

Let $\boldsymbol{\theta}$ have r free parameters and the null fix $r - q$ of them, leaving q free. If the null is true, the model is regular and the true value lies in the interior of Θ_0, then for large samples

$$-2\ln\lambda \;\sim\; \chi^2_{r-q}$$

[the degrees of freedom are the number of parameters the null fixes], whatever the model [Wasserman Thm 10.22], [CasellaBerger Thm 10.3.3]. The p-value is $\mathrm{P}(\chi^2_{r-q} > -2\ln\lambda_{\mathrm{obs}})$.

For example, with five parameters and the null $\theta_4 = \theta_5 = 0$, the statistic is asymptotically $\chi^2_{5-3} = \chi^2_2$ [Wasserman p. 164]. The nuisance parameters, the ones left free under both hypotheses, cost nothing in degrees of freedom: they are refitted in numerator and denominator.

4.7.2 Why the answer is χ^2

Take one parameter, θ, and a simple null $\theta = \theta_0$, so $r - q = 1$. Write $\ell(\theta) = \ln\mathcal{L}(\theta)$ and expand it around its maximum $\hat{\theta}$ [CasellaBerger p. 395]:

$$\begin{aligned} -2\ln\lambda &\overset{(1)}{=} 2\big[\ell(\hat{\theta}) - \ell(\theta_0)\big] \\ &\overset{(2)}{=} 2\Big[\ell(\hat{\theta}) - \Big(\ell(\hat{\theta}) + \ell'(\hat{\theta})(\theta_0 - \hat{\theta}) + \tfrac{1}{2}\ell''(\hat{\theta})(\theta_0 - \hat{\theta})^2 + \dots\Big)\Big] \\ &\overset{(3)}{=} -\ell''(\hat{\theta})\,(\hat{\theta} - \theta_0)^2 + \dots \end{aligned}$$

$$\overset{(4)}{=} \left(\frac{\hat\theta - \theta_0}{\widehat{\mathrm{se}}}\right)^2 + \dots . \tag{4.29}$$

(1) Definition (4.28) with $\hat\theta_0 = \theta_0$.
(2) Taylor expansion of $\ell(\theta_0)$ about $\hat\theta$.
(3) $\ell'(\hat\theta) = 0$ at the maximum, and the $\ell(\hat\theta)$ terms cancel.
(4) The curvature of the log likelihood at its peak is the observed information, and its inverse is the variance of the estimate: $-\ell''(\hat\theta) = 1/\widehat{\mathrm{se}}^{\,2}$ (Chapter 3).

For a large sample, $\hat\theta$ is approximately normal about the true value with standard deviation se. Under the null the true value is θ_0, so the bracket in the last line is approximately a standard normal Z, and $-2\ln\lambda \approx Z^2 \sim \chi^2_1$. The dropped terms are of higher order in $\hat\theta - \theta_0$, which shrinks as $1/\sqrt{n}$. Near its peak, every smooth log likelihood is a parabola, the likelihood-ratio statistic is the drop of that parabola between $\hat\theta$ and θ_0, and the drop is the square of the Wald statistic of Section 4.5. With k fixed parameters the same expansion gives a quadratic form in a k-dimensional normal vector weighted by its inverse covariance, which is a sum of k squared independent standard normals: χ^2_k.

The linear-Gaussian case, where Wilks is exact. If the data are $y_i = \mu_i(\boldsymbol\theta) + \epsilon_i$ with known Gaussian errors σ_i, then $-2\ln\mathcal{L}(\boldsymbol\theta) = \chi^2(\boldsymbol\theta) + \text{const}$ with $\chi^2(\boldsymbol\theta) = \sum_i (y_i - \mu_i(\boldsymbol\theta))^2/\sigma_i^2$ (Chapter 3), and

$$-2\ln\lambda \overset{(1)}{=} -2\ln\mathcal{L}(\hat{\boldsymbol\theta}_0) + 2\ln\mathcal{L}(\hat{\boldsymbol\theta}) \overset{(2)}{=} \chi^2_{\min}(H_0) - \chi^2_{\min}(H_1) \equiv \Delta\chi^2. \tag{4.30}$$

(1) Definition (4.28).
(2) Maximising the likelihood is minimising χ^2, and the constants cancel in the difference.

When the model is linear in the parameters, $\Delta\chi^2$ is exactly χ^2_{r-q} for any sample size. Here is the geometric reason. Divide each residual by its σ_i; the noise vector $\boldsymbol\epsilon/\boldsymbol\sigma$ is then an isotropic Gaussian in N dimensions. A least-squares fit with r linear parameters is the orthogonal projection of the data onto an r-dimensional subspace; the null's fit projects onto a q-dimensional subspace inside it. Under the null, the difference of the two squared residual lengths is the squared length of the noise projected onto the $(r-q)$-dimensional part of the big subspace orthogonal to the small one. An isotropic Gaussian projected onto any k-dimensional subspace is an isotropic Gaussian in k dimensions, and its squared length is χ^2_k.

Example: the Poisson rate. With $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Pois}(\lambda)$ and $H_0 : \lambda = \lambda_0$ [CasellaBerger Ex. 10.3.2], the log likelihood is $\ell(\lambda) = -n\lambda + S\ln\lambda + \text{const}$ with $S = \sum_i X_i$, and its maximum is at $\hat\lambda = S/n$. Then

$$\begin{aligned} -2\ln\lambda &\overset{(1)}{=} 2\big[\ell(\hat\lambda) - \ell(\lambda_0)\big] \overset{(2)}{=} 2\big[-n\hat\lambda + S\ln\hat\lambda + n\lambda_0 - S\ln\lambda_0\big] \\ &\overset{(3)}{=} 2n\left[(\lambda_0 - \hat\lambda) - \hat\lambda\ln\frac{\lambda_0}{\hat\lambda}\right]. \end{aligned} \tag{4.31}$$

(1) Definition, with a simple null.
(2) Insert ℓ; the $\ln X_i!$ terms in the constant cancel.
(3) Use $S = n\hat\lambda$ and $\ln\hat\lambda - \ln\lambda_0 = -\ln(\lambda_0/\hat\lambda)$.

Casella and Berger simulate it for $\lambda_0 = 5$ and $n = 25$. Our 100 000 simulated experiments give the percentiles below; they sit close to those of χ^2_1 already at this modest sample size (their own table has 1.630, 2.726, 3.744 and 6.304 from 10 000 experiments).

percentile	0.80	0.90	0.95	0.99
simulated $-2\ln\lambda$	1.630	2.726	3.757	6.685
χ^2_1	1.642	2.706	3.841	6.635

Example: is there a quadratic term? Twenty points y_i at $x_i \in [-1, 1]$ carry Gaussian noise of known $\sigma = 0.5$, and the truth is a constant. We fit a constant, a straight line and a parabola $a + bx + cx^2$, and form $\Delta\chi^2$ for two nulls: $c = 0$ (one constraint) and $b = c = 0$ (two constraints). Over 100 000 simulated data sets the first has mean 1.002 and variance 2.011, the second mean 2.003 and variance 3.995, matching χ^2_1 (mean 1, variance 2) and χ^2_2 (mean 2, variance 4). Figure 4.17 shows both examples.

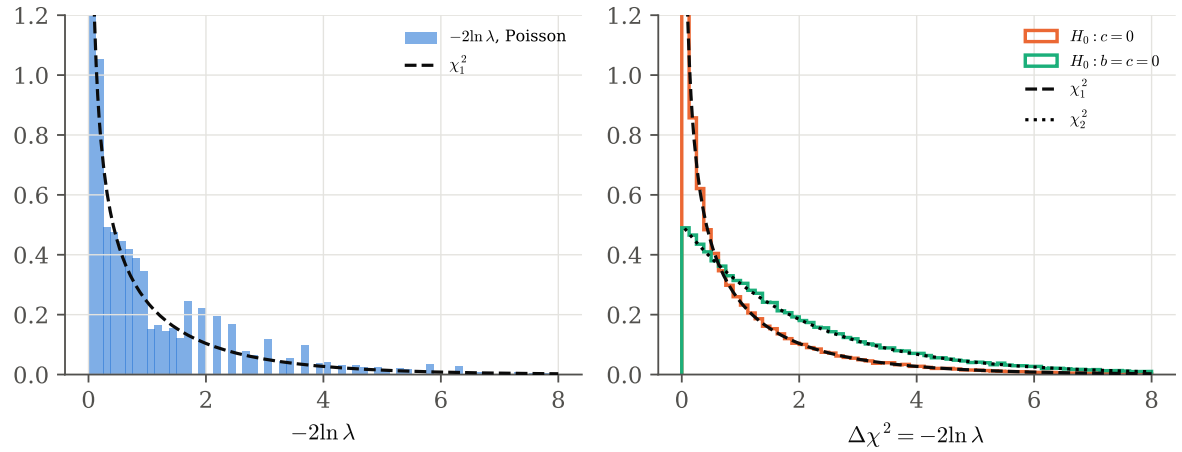


Figure 4.17: Wilks' theorem checked twice. Left: the likelihood-ratio statistic for a Poisson rate (25 counts, $\lambda_0 = 5$) follows the χ^2_1 density (dashed) closely. Right: $\Delta\chi^2$ between nested polynomial fits with Gaussian noise follows χ^2_1 when one coefficient is tested (orange) and χ^2_2 when two are (green, with the dotted density). For linear models the agreement is exact, not only asymptotic.

Example: Mendel's peas by likelihood ratio. Mendel counted $X = (315, 101, 108, 32)$ peas in four classes, $n = 556$, where his theory predicts probabilities $p_0 = (\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16})$ [Wasserman Ex. 10.18]. The multinomial likelihood is $\prod_j p_j^{X_j}$, maximised at $\hat{p}_j = X_j/n$, so

$$-2\ln\lambda = 2\sum_{j=1}^{4} X_j \ln\frac{X_j/n}{p_{0j}}$$
$$= 2\left[315\ln\frac{315/556}{9/16} + 101\ln\frac{101/556}{3/16} + 108\ln\frac{108/556}{3/16} + 32\ln\frac{32/556}{1/16}\right] = 0.48$$

[Wasserman Ex. 10.23]. The unrestricted model has three free probabilities (four that sum to one), the null none, so the reference is χ^2_3 and the p-value is about 0.92. The Pearson statistic of Section 4.8 gives 0.47 on the same data: for large counts the two statistics agree to second order.

4.7.3 When the null sits on a boundary: $Z = \sqrt{q_0}$

When the null value sits on the edge of the allowed region, Wilks' theorem needs a correction, and the corrected answer is very simple. The case that matters is the counting experiment of Section 4.3.3: a count $N \sim \text{Pois}(s + b)$, with known expected background b, unknown expected

signal s, and the physical constraint $s \geq 0$. The null $s = 0$ sits on the *edge* of the allowed region, which breaks the interior condition of Wilks' theorem.

The fit. The log likelihood is $\ell(s) = N \ln(s+b) - (s+b) + \text{const}$. Its maximum is at $s = N - b$ if that is allowed, and at the boundary $s = 0$ otherwise, so $\hat{s} = \max(N - b, 0)$.

The statistic. The discovery statistic is the likelihood-ratio statistic (4.28) for the null $s = 0$:

$$q_0 = -2 \ln \frac{\mathcal{L}(0)}{\mathcal{L}(\hat{s})} \overset{(1)}{=} 2\big[N \ln(\hat{s}+b) - (\hat{s}+b) - N \ln b + b\big] \overset{(2)}{=} \begin{cases} 2\big[N \ln(N/b) - (N-b)\big], & N > b, \\ 0, & N \leq b. \end{cases} \tag{4.32}$$

(1) Insert ℓ at $\hat{s}$ and at 0.
(2) For $N > b$, $\hat{s} + b = N$. For $N \leq b$ the best allowed fit is $\hat{s} = 0$, the same as the null, and the ratio is one. A deficit is never evidence for a signal.

Its null distribution. For a large background, N is nearly Gaussian about b and falls below it half of the time; then $q_0 = 0$. In the other half, the unconstrained Wilks argument applies and q_0 is χ^2_1. So the null distribution is the mixture $\frac{1}{2}\delta(q_0) + \frac{1}{2}\chi^2_1$, a result that goes back to Chernoff. Its tail is

$$\mathrm{P}(q_0 \geq Z^2 \,|\, H_0) \overset{(1)}{=} \tfrac{1}{2}\,\mathrm{P}\big(\chi^2_1 \geq Z^2\big) \overset{(2)}{=} \tfrac{1}{2}\,\mathrm{P}\big(|Z'| \geq Z\big) \overset{(3)}{=} \Phi(-Z), \qquad Z' \sim \mathcal{N}(0,1), \tag{4.33}$$

(1) For $Z > 0$ only the χ^2_1 half contributes, with weight $\frac{1}{2}$.
(2) A χ^2_1 variable is the square of a standard normal.
(3) The two tails of Z' are equal, $\mathrm{P}(|Z'| \geq Z) = 2\Phi(-Z)$.

so the significance of an observation is simply $Z = \sqrt{q_0}$ standard deviations [27]. The same half-χ^2, and $Z = \sqrt{q_0}$, survive when the background and calibrations are nuisance parameters profiled in q_0, as the second part of the chapter shows in its section on discovery and exclusion with nuisance parameters. Figure 4.18 tests this with two million background-only toys at $b = 3.2$ and at $b = 100$. The fraction with $q_0 = 0$ is 0.603 and 0.527, above one half because the Poisson puts extra probability at and just below its mean. The fraction with $q_0 \geq 9$, that is $Z \geq 3$, is 0.001769 and 0.001262, against the asymptotic 0.00135: the formula is good to about 7% at $b = 100$ and to about 30% at $b = 3.2$. For Cowan's peak, 11 events on 3.2, $Z = \sqrt{q_0} = 3.40$, while the exact Poisson tail corresponds to 3.29.

Listing 4.7: The discovery statistic for a counting experiment (`code/ch04/07_wilks.py`)

```
# ---------- (iii) counting experiment with s >= 0 ----------
def q0(nobs, b):
    nobs = np.asarray(nobs, float)
    with np.errstate(divide="ignore", invalid="ignore"):
        q = 2 * (nobs * np.log(nobs / b) - (nobs - b))
    return np.where(nobs > b, q, 0.0)
```

The function is (4.32) verbatim: the `np.where` enforces $\hat{s} \geq 0$, and the `errstate` guard silences the harmless warnings from $\ln 0$ in the branch that is discarded.

Example: the expected significance, and when $s/\sqrt{b}$ is enough. Before an experiment we want its *median* significance if a signal s is present. We can get it without simulation [27]:

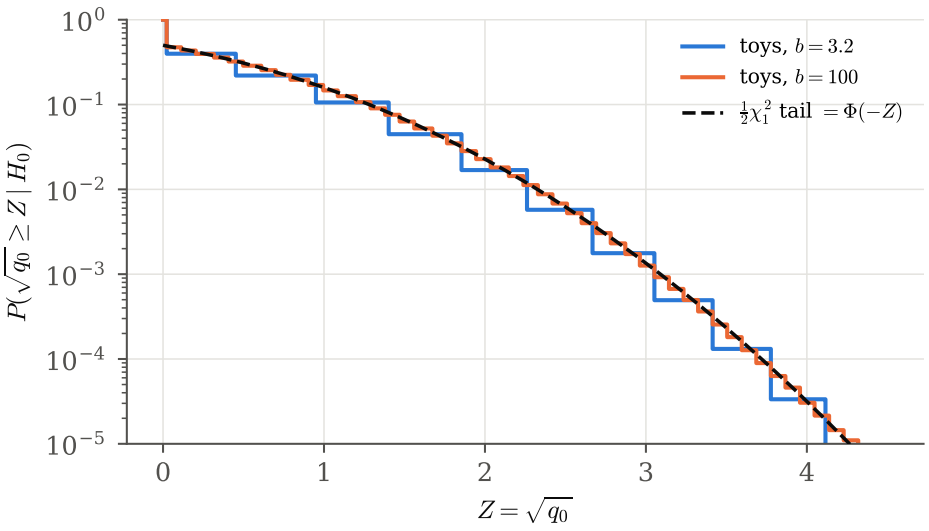


Figure 4.18: The probability that background alone reaches a significance $Z = \sqrt{q_0}$ or more, from toys with $b = 3.2$ (blue) and $b = 100$ (orange), compared with the asymptotic half-χ^2_1 prediction $\Phi(-Z)$ (dashed). The large-background staircase follows the prediction closely; the small-background one jumps in coarse steps because only integer counts are possible, and wanders above and below the line.

evaluate q_0 on the "Asimov" data set, in which N is replaced by its expectation $s + b$. Then

$$\begin{aligned} Z_A^2 = q_0\big(N = s+b\big) &\overset{(1)}{=} 2\left[(s+b)\ln\frac{s+b}{b} - s\right] \overset{(2)}{=} 2\left[(s+b)\ln\left(1+\frac{s}{b}\right) - s\right] \\ &\overset{(3)}{=} 2\left[(s+b)\left(\frac{s}{b} - \frac{s^2}{2b^2} + \frac{s^3}{3b^3} - \dots\right) - s\right] \\ &\overset{(4)}{=} \frac{s^2}{b}\left[1 - \frac{s}{3b} + \dots\right]. \end{aligned} \tag{4.34}$$

(1) (4.32) with $N = s + b > b$; $N - b = s$.

(2) $(s + b)/b = 1 + s/b$.

(3) $\ln(1 + u) = u - u^2/2 + u^3/3 - \dots$ with $u = s/b$.

(4) Multiply out: $(s + b)(s/b) = s^2/b + s$, whose s cancels the $-s$; $(s + b)(-s^2/2b^2) = -s^2/2b - s^3/2b^2$; $(s + b)(s^3/3b^3) = s^3/3b^2 + \dots$. Collecting, $s^2/b - s^2/2b = s^2/2b$ and $-s^3/2b^2 + s^3/3b^2 = -s^3/6b^2$; the factor 2 gives $s^2/b - s^3/3b^2$.

So for $s \ll b$ the familiar $Z \approx s/\sqrt{b}$ is recovered, and the first correction shows it *overestimates* the significance. With $s = 10$ on $b = 100$, $Z_A = 0.984$, $s/\sqrt{b} = 1.000$, and the median of toy significances is 0.984. With $s = 5$ on $b = 3.2$ the shortcut fails badly: $Z_A = 2.33$ against $s/\sqrt{b} = 2.80$.

How this is used in real research: Profile likelihood ratios at the LHC

The discovery of the Higgs boson and most searches since use exactly this statistic, with many nuisance parameters for backgrounds, luminosity and energy scales. Each is "profiled": maximised separately in numerator and denominator of λ. The asymptotic formulae, $Z = \sqrt{q_0}$ and the Asimov median, are derived for this general case in [27]. Our toys *verify* them against an analytic answer. In research the same toys are the *instrument*: where counts are small, parameters sit on boundaries or the signal mass is scanned, the null distribution of q_0 is generated by simulation, and the asymptotic formula serves only as a cross-check.

Where Wilks does not apply. Three conditions carried the derivation, and each fails somewhere in physics. The parameter must be in the interior (the boundary gives the half-χ^2 above). The hypotheses must be nested, one a special case of the other. And every parameter of the alternative must mean something under the null. The last condition fails in a bump hunt: the signal mass is meaningless when the signal strength is zero, and maximising over it inflates q_0 beyond any χ^2. That is the look-elsewhere effect in its likelihood form [15], and it leads to the trials factor treated in the second part of the chapter.

4.8 Goodness of fit: Pearson's χ^2 and Kolmogorov–Smirnov

So far the null has always been tested against a definite direction: a signal in a window, a bias of a coin, a shift of a calibration. Often we want something vaguer. Here is the background model for the whole mass spectrum, thirty bins of it; does it describe the data *at all*? No particular alternative is in mind, just "something else". This is a **goodness-of-fit** test [Cowan §4.5]: the test statistic itself measures the disagreement between data and model, and its large values are the extreme direction.

The question in words.

1. How do we add up the disagreement of many bins into one number with a known null distribution?
2. What changes when the total is fixed, when parameters were fitted to the same data, and when counts are small?
3. Can we avoid binning altogether?

4.8.1 The χ^2 distribution, recalled

If $Z_1, \dots, Z_k$ are independent standard normals, $V = \sum_{i=1}^{k} Z_i^2$ has the χ^2 distribution with k degrees of freedom (Chapter 2), with density, mean and variance [Wasserman p. 159]

$$f(v) = \frac{v^{k/2-1} e^{-v/2}}{2^{k/2}\Gamma(k/2)} \quad (v > 0), \qquad \mathbb{E}\, V = k, \qquad \operatorname{Var} V = 2k. \tag{4.35}$$

The mean is immediate, $\mathbb{E}\, Z_i^2 = 1$ for each term, and the variance follows from $\mathbb{E}\, Z^4 = 3$: $\operatorname{Var} Z_i^2 = 3 - 1 = 2$ per term. Its upper quantile $\chi^2_{k,\alpha}$ is defined by $\mathrm{P}(\chi^2_k > \chi^2_{k,\alpha}) = \alpha$, and a p-value is the upper tail beyond the observed value. For large k the distribution is close to $\mathcal{N}(k, 2k)$.

4.8.2 Pearson's statistic

A histogram has N bins with observed counts n_i and expected counts ν_i under the model. If the counts are independent Poisson variables and none of the ν_i is small, each n_i is approximately Gaussian with mean ν_i and variance ν_i, so $(n_i - \nu_i)/\sqrt{\nu_i}$ is approximately a standard normal. Adding the squares [Cowan (4.39)],

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i} \;\sim\; \chi^2_N \qquad \text{(approximately, for } \nu_i \gtrsim 5\text{)}. \tag{4.36}$$

The approximation does not care what physical variable was histogrammed or what shape the model has; the test is **distribution free**. Each term is the squared deviation in units of its own standard deviation, so χ^2 is the total disagreement counted in "sigmas squared".

If instead the total $n = \sum_i n_i$ is fixed, the counts are multinomial with cell probabilities p_{0i} and expected counts $E_i = np_{0i}$, and Pearson's statistic is [Wasserman Def. 10.16], [Cowan (4.41)]

$$T = \sum_{j=1}^{k} \frac{(X_j - E_j)^2}{E_j} \;\sim\; \chi^2_{k-1} \tag{4.37}$$

under the null, asymptotically [Wasserman Thm 10.17]. The one lost degree of freedom is the constraint $\sum_j (X_j - E_j) = 0$: the deviations are not independent, because an excess in one cell must be paid for by deficits elsewhere. The case $k = 2$ shows the mechanism exactly. With cell probabilities p and $1-p$, $X_2 - n(1-p) = -(X_1 - np)$, so

$$\begin{aligned} T &\overset{(1)}{=} \frac{(X_1 - np)^2}{np} + \frac{(X_1 - np)^2}{n(1-p)} \\ &\overset{(2)}{=} (X_1 - np)^2 \, \frac{(1-p) + p}{np(1-p)} \overset{(3)}{=} \left(\frac{X_1 - np}{\sqrt{np(1-p)}} \right)^2 . \end{aligned} \tag{4.38}$$

(1) The second deviation is minus the first, and its square is the same.
(2) Common denominator.
(3) $(1-p) + p = 1$.

The bracket is the binomial count standardised by its own mean np and standard deviation $\sqrt{np(1-p)}$, which by the de Moivre–Laplace theorem of Chapter 2 is approximately standard normal. So $T \approx Z^2 \sim \chi^2_1$: two cells, one degree of freedom. For k cells the same happens in $k-1$ dimensions: the constraint confines the standardised deviation vector to a hyperplane, and its squared length is a sum of $k-1$ independent squared normals.

Example: Mendel's peas. Mendel crossed peas and counted $X = (315, 101, 108, 32)$ progeny of four types, $n = 556$; his theory predicts $p_0 = (\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16})$ [Wasserman Ex. 10.18]. The expected counts are $E = (312.75, 104.25, 104.25, 34.75)$ and

$$\begin{aligned} T &= \frac{(315-312.75)^2}{312.75} + \frac{(101-104.25)^2}{104.25} + \frac{(108-104.25)^2}{104.25} + \frac{(32-34.75)^2}{34.75} \\ &= 0.016 + 0.101 + 0.135 + 0.218 = 0.470. \end{aligned}$$

With $k - 1 = 3$ degrees of freedom the 5% critical value is 7.815, far above, and the p-value is $\mathrm{P}(\chi^2_3 > 0.470) = 0.925$. So the data do not contradict Mendel's theory. They do not prove it either. A test like this can show evidence to reject the null, but it cannot prove the null [Wasserman p. 161], because a failure to reject may come from a true null or from low power. (A footnote there recalls that Mendel's results have been suspected of fitting *too* well. A p-value very close to one is also improbable under the null.)

Watch out: χ^2/ν near one is not enough

It is common to quote χ^2 divided by the number of degrees of freedom and call anything near one a good fit. The ratio hides the scale: the spread of χ^2_ν/ν is $\sqrt{2/\nu}$, which shrinks with ν. For $\chi^2 = 15$ on 10 degrees of freedom the p-value is 0.13, perfectly fine; for $\chi^2 = 150$ on 100, the same ratio, it is 9.0×10^{-4} [Cowan p. 62]. Always quote χ^2 and ν separately, or the p-value.

4.8.3 Fitted parameters and small counts

Usually the model has parameters, such as a background slope, that are estimated from the same histogram. The fit pulls the expectation towards the data and makes χ^2 smaller than it would be at the true parameters. If s parameters are estimated by maximising the multinomial likelihood of the binned counts, $Q(\boldsymbol{\theta}) = \prod_j p_j(\boldsymbol{\theta})^{N_j}$, then [Wasserman Thm 10.29]

$$Q = \sum_{j=1}^{k} \frac{\big(N_j - n p_j(\hat{\boldsymbol{\theta}})\big)^2}{n p_j(\hat{\boldsymbol{\theta}})} \;\sim\; \chi^2_{k-1-s} : \tag{4.39}$$

each fitted parameter removes one degree of freedom, for the same geometric reason as in Wilks' theorem (Section 4.7.2): the fit projects the deviations onto a smaller subspace. This count holds only when the parameters are fitted to the binned counts. If they are estimated instead from the *unbinned* data, which is often more natural, the distribution of Q is no longer χ^2_{k-1-s}; a theorem of Chernoff and Lehmann puts it between χ^2_{k-1-s} and χ^2_{k-1} [Wasserman p. 169]. The reason is that the unbinned estimate is not tuned to the bins, so it removes less of the disagreement.

Example: fitting a decay length. The script `code/ch04/08_goodness_of_fit.py` draws 400 decay times with mean $\tau = 2$, histograms them in $k = 8$ bins, $[0,1), [1,2), \ldots, [7,\infty)$, and fits τ two ways: by maximising the binned multinomial likelihood, and by the unbinned maximum-likelihood estimate $\hat{\tau} = \bar{t}$. Over 4000 repetitions the mean of Q is 6.01 with the binned fit, as $\chi^2_{k-2} = \chi^2_6$ predicts, and 6.06 with the unbinned one; each mean carries a random error of about 0.05. Both fits see the same histograms, so their difference is measured much better: the unbinned fit leaves Q larger by 0.051 ± 0.002 on average, which puts it between χ^2_6 and χ^2_7, as Chernoff–Lehmann says (Figure 4.19, right). Here the difference is small; it matters more when the bins are few and the parameters many.

Example: when the bins are sparse. The χ^2_N approximation needs every ν_i to be reasonably large. With a few counts per bin, $(n_i - \nu_i)^2/\nu_i$ is far from a squared Gaussian: a Poisson with small mean is skewed, and a single count of 5 where 1 was expected contributes 16. [Cowan p. 63] finds $\chi^2 = 29.8$ for 20 bins of his bump histogram, where most bins have fewer than five entries, and shows that the χ^2_{20} tail (0.073) understates the true p-value (0.11, from Monte Carlo). Our version uses 20 Poisson bins with expected counts between about 1 and 4. Simulating 200 000 histograms under the null, the 95% point of Pearson's statistic is 33.5, against 31.4 for χ^2_{20}. For an observed value of 30 the simulated p-value is 0.091 and the χ^2 formula would give 0.070. With 25 times more data in the same shape the simulated 95% point becomes 31.4, on the χ^2 value (Figure 4.19, left). The remedy is Cowan's: when counts are small, simulate the null distribution of the statistic you computed.

Binning is a choice, and different choices give different p-values [Cowan p. 63]. Bins must be wide enough for the approximation, yet wide bins throw away where in the bin each event fell. Simulating the null distribution relaxes the first constraint. Avoiding bins altogether removes the choice.

4.8.4 The Kolmogorov–Smirnov test

For n individual measurements $x_1, \ldots, x_n$ we can compare the data with a hypothesised continuous CDF F directly. The **empirical CDF** [the fraction of the data at or below x]

$$F_n(x) = \frac{1}{n} \#\{i : \; x_i \le x\}$$

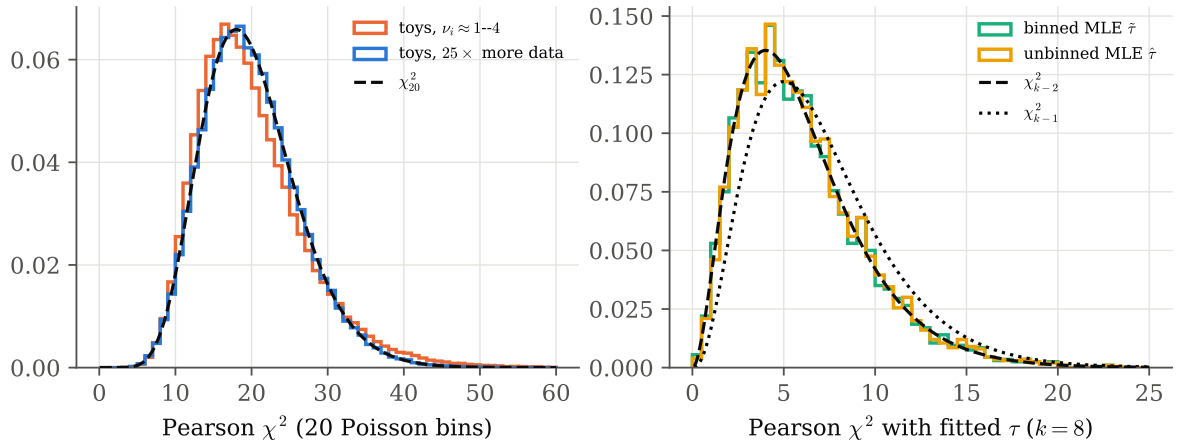


Figure 4.19: Left: the null distribution of Pearson's χ^2 for 20 Poisson bins. With one to four expected counts per bin (orange) it has a longer tail than χ^2_{20} (dashed); with 25 times more data (blue) it lies on the curve. Right: after fitting a decay length, Pearson's statistic follows χ^2_{k-2} when the fit uses the binned counts (green) and sits slightly towards χ^2_{k-1} (dotted) when the fit uses the unbinned decay times (yellow).

is a staircase that climbs by $1/n$ at each data point. The **Kolmogorov–Smirnov statistic** is its largest vertical distance from the model,

$$D_n = \sup_x |F_n(x) - F(x)| \,. \tag{4.40}$$

The supremum over all x reduces to a check at the data points. Sort them, $x_{(1)} \le \cdots \le x_{(n)}$. Between two data points F_n is flat and F increases, so the gap $F_n - F$ is largest just after a jump and $F - F_n$ is largest just before one. At $x_{(i)}$ the staircase jumps from $(i-1)/n$ to i/n, hence

$$D_n = \max\bigl(D_n^+, D_n^-\bigr), \qquad D_n^+ = \max_i\Bigl(\frac{i}{n} - F(x_{(i)})\Bigr), \qquad D_n^- = \max_i\Bigl(F(x_{(i)}) - \frac{i-1}{n}\Bigr). \tag{4.41}$$

Why is the null distribution of D_n the same for every continuous F? Set $u_i = F(x_i)$. Under the null the u_i are uniform on $(0,1)$ by the probability integral transform of Section 4.4.2, and because F is increasing, "$x_i \le x$" is the same as "$u_i \le F(x)$". Writing $u = F(x)$,

$$|F_n(x) - F(x)| \overset{(1)}{=} \Bigl|\frac{1}{n}\#\{i : u_i \le u\} - u\Bigr| \overset{(2)}{=} |G_n(u) - u| \,, \tag{4.42}$$

(1) Replace each event $x_i \le x$ by the equivalent $u_i \le u$.
(2) G_n is the empirical CDF of the uniform sample $u_1, \ldots, u_n$.

so D_n is the Kolmogorov–Smirnov distance of a uniform sample from the uniform CDF, whatever F was. Its distribution can be tabulated once; for large n, Kolmogorov showed that $\sqrt{n}D_n$ converges to a limiting distribution with tail $\mathrm{P}(K > x) = 2\sum_{j\ge1}(-1)^{j-1}e^{-2j^2x^2}$.

Example: a decay length that is not what we thought. Fifty decay times are drawn with mean 1.3, and we test the null that the mean is 1, $F(x) = 1 - e^{-x}$. Figure 4.20 (left) shows the empirical staircase below the model curve over most of the range, since longer lifetimes make small times rarer. Our own implementation of (4.41) finds $D_{50} = 0.1790$, so $\sqrt{n}D_n = 1.266$. The exact p-value from `scipy.stats.kstest` (which also returns $D = 0.1790$) is 0.0714; the Kolmogorov limit gives 0.0812; and 40 000 simulated null samples of size 50 give 0.0726. With only fifty events, a 30% change of lifetime gives modest evidence against the null at best, and another fifty decays could easily give a much larger p-value.

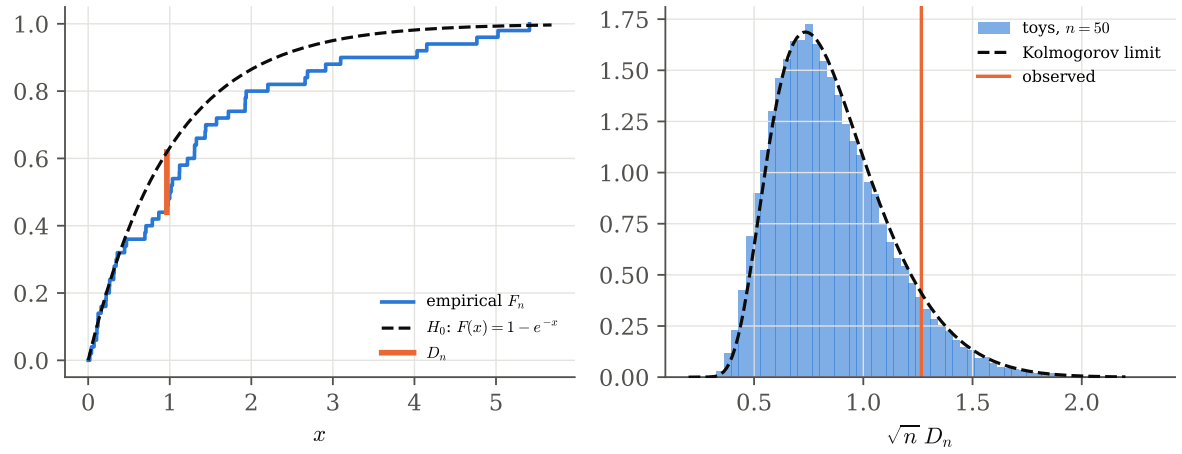


Figure 4.20: Left: the empirical CDF of fifty decay times (blue staircase) and the null model's CDF (dashed). The orange bar marks the largest vertical gap, D_n. Right: the null distribution of $\sqrt{n}\, D_n$ for samples of fifty, from simulation (histogram) and Kolmogorov's large-n limit (dashed), with the observed value marked.

Listing 4.8: The Kolmogorov–Smirnov statistic from scratch (`code/ch04/08_goodness_of_fit.py`)

```
# ---------- (iv) Kolmogorov-Smirnov from scratch ----------
def ks_stat(sample, cdf):
    """D_n = sup_x |F_n(x) - F(x)|; the sup is attained just before or at a data point."""
    xs = np.sort(sample); n = xs.size
    F = cdf(xs)
    d_plus = np.max(np.arange(1, n + 1) / n - F)   # F_n jumps up to i/n at x_(i)
    d_minus = np.max(F - np.arange(0, n) / n)      # just before the jump F_n = (i-1)/n
```

The function sorts the sample, evaluates the model CDF at each point, and takes the two maxima of (4.41) with the vectors `np.arange(1, n+1)/n` (the tops of the steps) and `np.arange(0, n)/n` (the bottoms). It agrees with scipy to every digit printed. The Monte Carlo p-value calls the same function on samples drawn from the null.

Two cautions. The distribution-free property holds only when F is fully specified in advance. If its parameters are estimated from the same data, the fitted CDF hugs the data and D_n is typically smaller, so the tabulated p-values are too large; the null distribution must then be simulated, fitting the parameters afresh on each simulated sample. And like every goodness-of-fit test, a large p-value does not certify the model: it may only mean the test has little power against the departure that is actually present [Wasserman p. 169]. Relatives of KS that weight the whole curve rather than its largest gap, such as the Cramér–von Mises statistic, are often more powerful [Cowan p. 63].

4.9 From p to $Z\sigma$, and why a discovery needs 5σ

Particle physicists rarely quote p-values. They say that an excess is "a 2.6σ fluctuation", that 3σ counts as evidence, and that a discovery needs 5σ. The language comes from Gaussian error bars, but it is used for every kind of test statistic, including Poisson counts and likelihood ratios that are nothing like Gaussian. So "Z sigma" must be a relabelling of the p-value, and it is: a precise one-to-one translation. Once we have it, we can ask why particle physics sets the bar so high, when other fields are content with $p < 0.05$.

The question in words.

1. What exactly does "Z sigma" mean for a test statistic that is not Gaussian?
2. How small is the p-value of a 5σ result, and how does it compare with 3σ?
3. What goes wrong with a lower threshold, given many searches, imperfect backgrounds and new physics that is rare?

4.9.1 The translation

The **significance** Z of a p-value is the number of standard deviations above the mean at which a standard normal variable has the same upper-tail probability [27]:

$$p = \int_{Z}^{\infty} \frac{e^{-t^2/2}}{\sqrt{2\pi}}\,\mathrm{d}t = 1 - \Phi(Z) = \Phi(-Z), \qquad Z = \Phi^{-1}(1-p). \tag{4.43}$$

The definition says nothing about the distribution of the actual statistic. A Poisson count, a χ^2, a permutation statistic: whatever gave the p-value, Z is only a relabelling of it on a scale that physicists find intuitive. For the bump of Section 4.1, $p = 0.005172$ becomes $Z = 2.56$. Some authors use the two-sided convention $p = 2\Phi(-Z)$, which corresponds to $|\text{Gaussian}| \geq Z$; discovery claims use the one-sided version, because only an excess counts. The table from `code/ch04/09_p_to_z.py`:

Z	one-sided $\Phi(-Z)$	two-sided $2\Phi(-Z)$
1	0.1587	0.3173
2	0.02275	0.0455
3	0.00135	0.0027
4	3.167×10^{-5}	6.334×10^{-5}
5	2.867×10^{-7}	5.733×10^{-7}
6	9.866×10^{-10}	1.973×10^{-9}

In the other direction, $p = 0.05$ corresponds to $Z = 1.645$ one-sided and 1.960 two-sided. The usual 5% threshold of the life sciences is a 1.6 to 2σ effect, and a 5σ result has a p-value about 170 000 times smaller.

Example: reading the scale in both directions. A search reports a one-sided p-value of 2.5×10^{-4}. Then $Z = \Phi^{-1}(1 - 2.5 \times 10^{-4}) = 3.48$, a little above the 3σ line of "evidence" and well short of 5σ. Had the same number been a two-sided p-value, the one-sided tail would be half of it, 1.25×10^{-4}, and $Z = 3.66$; quoting Z without saying which convention is in use can move the claim by a fifth of a sigma. In the other direction, an analysis planned to reach $Z = 4.2$ must see a background-only tail no bigger than $\Phi(-4.2) = 1.3 \times 10^{-5}$, about one chance in 75 000.

How fast the tail falls. For large Z there is a simple estimate of the tail, useful for checking orders of magnitude without a table. Write $\varphi(t) = e^{-t^2/2}/\sqrt{2\pi}$ and note that $\varphi'(t) = -t\varphi(t)$. Integrating by parts,

$$\begin{aligned} \Phi(-Z) = \int_{Z}^{\infty} \varphi(t)\,\mathrm{d}t &\overset{(1)}{=} \int_{Z}^{\infty} \frac{1}{t}\,\big(t\varphi(t)\big)\,\mathrm{d}t \overset{(2)}{=} \Big[-\frac{\varphi(t)}{t}\Big]_{Z}^{\infty} - \int_{Z}^{\infty} \frac{\varphi(t)}{t^2}\,\mathrm{d}t \\ &\overset{(3)}{=} \frac{\varphi(Z)}{Z} - \int_{Z}^{\infty} \frac{\varphi(t)}{t^2}\,\mathrm{d}t \overset{(4)}{\approx} \frac{\varphi(Z)}{Z}\Big(1 - \frac{1}{Z^2}\Big). \end{aligned} \tag{4.44}$$

(1) Multiply and divide by t.

(2) Parts, with $u = 1/t$ and $\mathrm{d}v = t\varphi(t)\,\mathrm{d}t$, so $v = -\varphi(t)$ and $\mathrm{d}u = -\mathrm{d}t/t^2$.
(3) The boundary term vanishes at infinity.
(4) The remaining integral is smaller by a factor of about $1/Z^2$; the same trick applied to it gives the correction shown.

At $Z = 5$, $\varphi(5) = e^{-12.5}/\sqrt{2\pi} = 1.49 \times 10^{-6}$, so the estimate is $1.49 \times 10^{-6} \times 0.96/5 = 2.85 \times 10^{-7}$, against the exact 2.867×10^{-7}. The factor $e^{-Z^2/2}$ dominates: each extra sigma near $Z = 5$ costs a factor of $e^{-(Z+1)^2/2+Z^2/2} = e^{-Z-1/2} \approx 250$ in probability.

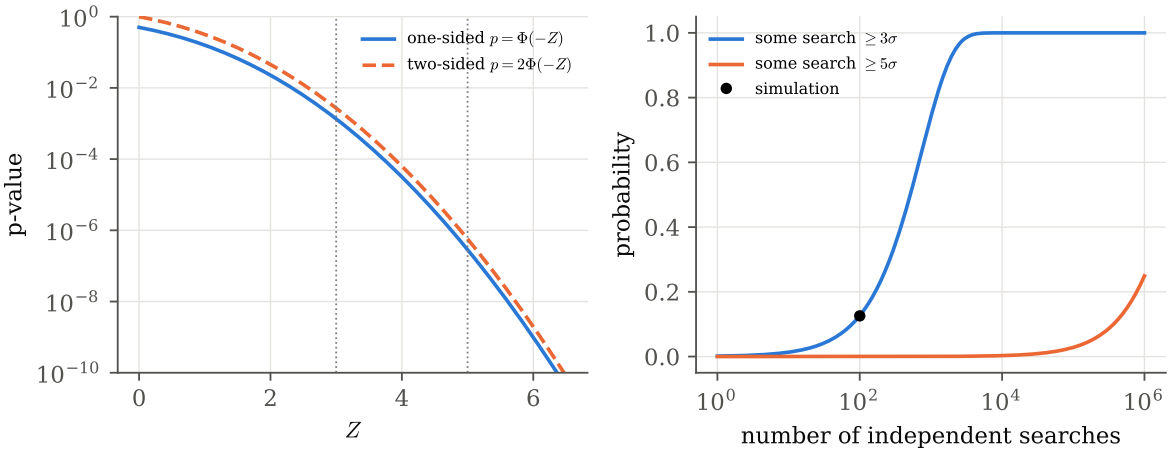


Figure 4.21: Left: the p-value that corresponds to a significance Z, one-sided (solid) and two-sided (dashed), on a log scale; the dotted lines mark 3σ and 5σ. Right: the probability that at least one of many independent background-only searches reaches 3σ (blue) or 5σ (orange). With a hundred searches the chance of a 3σ fluctuation somewhere is already about one in eight; the black dot is a direct simulation.

4.9.2 Why five

Three separate arguments each call for a threshold far stricter than $p < 0.05$: there are many searches, the backgrounds are imperfect, and new physics is rare. We take them one at a time.

Many searches. A laboratory does not run one test. It scans masses, channels, final states, and every analysis group does the same. If N independent searches each have probability p_1 of a false alarm, the probability that at least one of them fires is

$$\mathrm{P}(\text{at least one}) \overset{(1)}{=} 1 - \mathrm{P}(\text{none}) \overset{(2)}{=} 1 - (1 - p_1)^N \overset{(3)}{\approx} N p_1 \quad (N p_1 \ll 1). \tag{4.45}$$

(1) Complement.
(2) Independence: the probability that all N stay quiet is the product.
(3) $(1 - p_1)^N \approx 1 - N p_1$ to first order.

With $N = 100$ and a 3σ threshold, $p_1 = 0.00135$ gives 0.126, and a simulation of the maximum of 100 standard normals over 200 000 trials gives 0.126. One lab in eight would see a 3σ effect from nothing. At 5σ the same number is 2.866×10^{-5}. This is Kruschke's two coins of Section 4.2.2 on a larger scale, and it is the look-elsewhere effect that the second part of the chapter quantifies properly for scans over a mass.

Imperfect backgrounds. The p-value is computed from the null model alone, so an error in the null model goes straight into it. In Section 4.1.4 moving the background from 0.5 to 0.8

events changed the p-value of five events by a factor of about eight. Real detectors also have tails in their resolution functions that no Gaussian captures. A 5σ threshold leaves a margin: an effect that survives it is unlikely to be an artefact of a background that is off by a modest amount. A 3σ effect is not so robust.

New physics is rare. The third argument is the population argument of Section 4.4.4, and it needs Bayes' theorem. *The event* is "the search crossed the threshold". *The question* is what produced it: a real effect, or a false alarm on a null. Suppose that of all the hypotheses that searches test, only a fraction $\pi_1 = 10^{-3}$ are real. *Each route in turn:* when an effect is real, the search reaches the threshold half the time, so the power is 0.5; when the null is true, it reaches it with the false-alarm probability p_{thr} of the threshold. Figure 4.22 draws the probability tree. *Only then the numbers.* The fraction of threshold crossings that are real, the *purity* of the discoveries, is

$$\mathrm{P}(\text{real} \,|\, \text{crossed}) \overset{(1)}{=} \frac{\pi_1 \times 0.5}{\pi_1 \times 0.5 + (1-\pi_1)\, p_{\text{thr}}}, \tag{4.46}$$

(1) Bayes' theorem on the tree: the two branches that end in "crossed" are "real, then detected" and "null, then false alarm".

exactly the purity formula (4.25) of event selection, with hypotheses in place of particles. With $p_{\text{thr}} = 0.00135$ (3σ) the purity is 0.270: nearly three quarters of the "discoveries" would be false. With $p_{\text{thr}} = 2.867 \times 10^{-7}$ (5σ) it is 0.9994. The numbers π_1 and the power are illustrative, and a frequentist cannot even write π_1 into the p-value. But the lesson is robust: when true effects are rare, the threshold must be far smaller than π_1 for a crossing to mean what we want it to mean. Extraordinary claims need extraordinary evidence, and 5σ is how particle physics encodes that without writing down a prior. Chapter 5 writes it down.

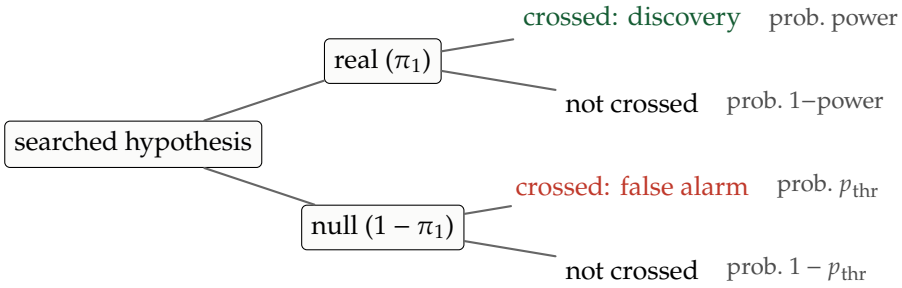


Figure 4.22: The probability tree behind the 5σ convention. First nature decides whether the hypothesis being searched for is real, with a small probability π_1; then the experiment either crosses the threshold or not. Among all crossings, the false alarms (red) come from the heavily populated null branch, the discoveries (green) from the thin real branch. Only a tiny false-alarm probability keeps the red leaf smaller than the green one.

What 5σ does not say. A 5σ result has $p = 2.867 \times 10^{-7}$. It does not follow that the probability of the null is 2.867×10^{-7}, nor that the new particle exists with probability $1 - 2.867 \times 10^{-7}$. As in Section 4.4.4, those are inference-direction statements and need a prior. The convention in particle physics is to call 3σ "evidence" and 5σ an "observation", to report the p-value (or Z) together with the expected significance of the experiment, and to leave the probability of the hypothesis to a Bayesian analysis.

Recap

- The null hypothesis is the one whose sampling distribution we know: "it happened by chance". A test runs in the prediction direction, from H_0 to the outcomes it produces.
- The p-value is the probability, under H_0, of a test statistic at least as extreme as observed, in the direction the alternative predicts. The direction, and the experimental plan (stopping rule, number of places looked), are part of the question.
- A test has size α (false-alarm rate) and a power function $\beta(\theta)$ (rejection rate under the truth). Power grows with effect size and with $\sqrt{n}$; experiment planning is a power calculation.
- Under a continuous null the p-value is uniform, so "reject if $p \le \alpha$" has size α. For discrete data it is valid: $\mathrm{P}(p \le \alpha) \le \alpha$.
- A p-value is not $\mathrm{P}(H_0 \mid \text{data})$: that needs the prior frequency of true nulls and the power (Bayes).
- Wald, t and permutation tests cover most everyday comparisons; the t-test is exact for small Gaussian samples, the permutation test for any distribution.
- Neyman–Pearson: for simple hypotheses the likelihood ratio is the most powerful statistic. With free parameters, $-2\ln\lambda$ is asymptotically χ^2 with as many degrees of freedom as the null fixes (Wilks); on a boundary it is half χ^2_1 and $Z = \sqrt{q_0}$.
- Pearson's χ^2 and Kolmogorov–Smirnov test a whole model; fitted parameters, small counts and estimated CDFs change their null distributions, which can always be simulated.
- $Z = \Phi^{-1}(1-p)$. Discovery needs 5σ because of many searches, imperfect backgrounds and the rarity of real effects.

4.10 From a test to an interval: the values we cannot reject

A new muon trigger is being commissioned. *What happened:* in a test run, 24 muons that we know passed through the detector are sent in, and the trigger fires on 7 of them. (The numbers are those of Kruschke's coin, $z = 7$ heads in $N = 24$ flips [Kruschke §11.3.1]; a trigger efficiency θ, the probability that the trigger fires on a muon, is a heads probability with a different name.) The design goal was an efficiency of one half. A test of $H_0 : \theta = 0.5$ answers one narrow question: do the data contradict the design value? They do not, quite: the two-sided p-value is 0.064, just above 0.05. But the engineer who must decide whether to rebuild the trigger has a different question. *Which efficiencies are compatible with* 7 *firings out of* 24*?* Is 0.4 still possible? Is 0.2? What about 0.6?

The tools of the first part of the chapter already answer this. Each candidate value θ_0 defines its own null hypothesis, $H_0 : \theta = \theta_0$, and we can test each one at level α. Some candidates are rejected; the rest survive. The set of survivors is a confidence interval. So a p-value asks how surprising the data are under one hypothesis, and a confidence interval asks the same question of every hypothesis at once and lists the ones that pass.

The question in words.

1. What property should a reported range $[a, b]$ have, so that quoting it is an honest summary of an experiment, whatever the true θ is?
2. How do we build a range with that property from tests we already know?
3. What does the range say, and what does it not say?

4.10.1 What a confidence interval promises

An interval estimate is a pair of numbers $a(\boldsymbol{X})$ and $b(\boldsymbol{X})$ computed from the data $\boldsymbol{X}$, with $a \le b$, together with the claim "$a \le \theta \le b$" [CasellaBerger Def. 9.1.1]. The data are random, so a and b are random too: rerun the experiment and the interval moves. The parameter does not move. It has one fixed value, which we do not know. So the one thing we can ask of a recipe for intervals is this: over repeated runs of the experiment, how often does the moving interval land on the fixed parameter?

Working definition 4.3 (confidence interval, coverage)

A $1-\alpha$ **confidence interval** for θ is a pair of statistics $a(\boldsymbol{X}) \le b(\boldsymbol{X})$ [functions of the data only, no unknown parameter inside] such that [Wasserman (6.9)]

$$\mathrm{P}_\theta\big(a(\boldsymbol{X}) \le \theta \le b(\boldsymbol{X})\big) \;\ge\; 1-\alpha \qquad \text{for every } \theta \in \Theta.$$

The left side, read as a function of θ, is the **coverage probability** [the fraction of repeated experiments, generated with this θ, whose interval contains it] [CasellaBerger Def. 9.1.4]. Its smallest value over θ is the **confidence coefficient** [CasellaBerger Def. 9.1.5]. *Operational test:* simulate many experiments at a fixed θ, build the interval each time, count the hits; repeat for other θ.

Three words in the definition carry the weight. "Every" θ: we do not know the truth, so the guarantee must hold whatever it is. "At least": for discrete data an exact $1-\alpha$ is impossible, and we then accept intervals that cover more often. "Repeated experiments": the probability is about the recipe, not about the one interval we computed. For a parameter vector the same definition applies to a region (an ellipse, say) instead of an interval, and we speak of a **confidence set**.

We do not need to imagine one experiment repeated forever for this to be useful [Wasserman p. 92]. On day one we measure one quantity and quote a 95% interval. On day two we measure a completely different quantity, on day three another. If every one of those intervals comes from a recipe with 95% coverage, then over a career 95% of the quoted intervals contain their true values, even though no two of them are about the same parameter. The pollster who writes "83% in favour, accurate to within 4 points 95 percent of the time" makes exactly this promise about a lifetime of different questions [Wasserman Ex. 6.13].

4.10.2 Inverting a family of tests

A level-α test of $H_0 : \theta = \theta_0$ has an **acceptance region** $A(\theta_0)$ [the set of data sets for which that null is retained], with

$$\mathrm{P}_{\theta_0}\big(\boldsymbol{X} \in A(\theta_0)\big) \;\ge\; 1-\alpha. \tag{4.47}$$

The acceptance region lives in the space of *data*. We now define a set in the space of *parameters*, one for each observed data set $\boldsymbol{x}$:

$$C(\boldsymbol{x}) = \big\{\theta_0 :\; \boldsymbol{x} \in A(\theta_0)\big\}, \tag{4.48}$$

the set of null values that the data $\boldsymbol{x}$ fail to reject. The two sets are tied together by a statement that is true by definition [CasellaBerger p. 340]:

$$\boldsymbol{x} \in A(\theta_0) \quad \Longleftrightarrow \quad \theta_0 \in C(\boldsymbol{x}). \tag{4.49}$$

Key idea: Test inversion

If $A(\theta_0)$ is the acceptance region of a level-α test of $\theta = \theta_0$, for every θ_0, then the set of values not rejected, $C(\boldsymbol{X}) = \{\theta_0 : \boldsymbol{X} \in A(\theta_0)\}$, is a $1-\alpha$ confidence set. Conversely, from any $1-\alpha$ confidence set, $A(\theta_0) = \{\boldsymbol{x} : \theta_0 \in C(\boldsymbol{x})\}$ is the acceptance region of a level-α test [CasellaBerger Thm. 9.2.2]. Tests and intervals are the same object read in two directions.

The proof is short, and it shows where "for every θ" comes from. Let θ be the true value, whatever it is. We ask: how often does the set of survivors $C(\boldsymbol{X})$ contain it? Then

$$\mathrm{P}_\theta\big(\theta \in C(\boldsymbol{X})\big) \overset{(1)}{=} \mathrm{P}_\theta\big(\boldsymbol{X} \in A(\theta)\big) \overset{(2)}{\geq} 1-\alpha. \tag{4.50}$$

(1) The event "the random set $C(\boldsymbol{X})$ contains θ" is, by (4.49) with $\theta_0 = \theta$, the same event as "the random data fall in the acceptance region of the test of the true value". Same event, same probability.

(2) The test of $\theta_0 = \theta$ has level α, (4.47), and this holds for the true value like for any other.

We never needed to know which θ is true. We built a test for *every* candidate. Whichever candidate is true, its own test decides whether the interval covers it. That test wrongly rejects with probability at most α, so the interval misses with probability at most α. The converse runs the same chain backwards: $\mathrm{P}_{\theta_0}(\boldsymbol{X} \notin A(\theta_0)) = \mathrm{P}_{\theta_0}(\theta_0 \notin C(\boldsymbol{X})) \leq \alpha$.

In practice we use the first direction. Building a level-α test is easy: choose a statistic and a tail, as in the first part of the chapter. Guessing the right shape of an interval is hard. Inversion solves the hard problem with the easy one, and every property of the test carries over to the interval: one-sided tests give one-sided intervals (limits), two-sided tests give two-sided intervals, a test based on a sufficient statistic gives an interval based on it, and a strangely shaped acceptance region gives a strangely shaped set, which need not even be an interval [CasellaBerger p. 341].

Example: the Gaussian mean. Take $X_1, \ldots, X_n$ independent, each $\mathcal{N}(\mu, \sigma^2)$ with σ known, and the two-sided test of $\mu = \mu_0$ that retains when the sample mean $\bar{x}$ is within $z_{\alpha/2}\sigma/\sqrt{n}$ of μ_0, where $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$ is the standard normal quantile (1.96 for $\alpha = 0.05$) [CasellaBerger Ex. 9.2.1]. Then

$$\begin{aligned} \boldsymbol{x} \in A(\mu_0) &\overset{(1)}{\Longleftrightarrow} |\bar{x} - \mu_0| \leq z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \\ &\overset{(2)}{\Longleftrightarrow} -z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \bar{x} - \mu_0 \leq z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \\ &\overset{(3)}{\Longleftrightarrow} \bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}. \end{aligned} \tag{4.51}$$

(1) Definition of the acceptance region of this test.

(2) An absolute value is at most c exactly when the quantity lies between $-c$ and c.

(3) Subtract $\bar{x}$ throughout and multiply by -1, which reverses both inequalities; then read the result as a condition on μ_0.

The last line is $C(\boldsymbol{x})$: the familiar $\bar{x} \pm z_{\alpha/2}\sigma/\sqrt{n}$. Figure 4.23 draws it. In the plane of (data, parameter), the test keeps the band between two parallel lines. Reading the band horizontally,

at fixed μ_0, gives the acceptance region, a range of $\bar{x}$. Reading it vertically, at the observed $\bar{x}$, gives the confidence interval, a range of μ. The test fixes the parameter and asks which data are consistent with it; the interval fixes the data and asks which parameters are consistent with them.

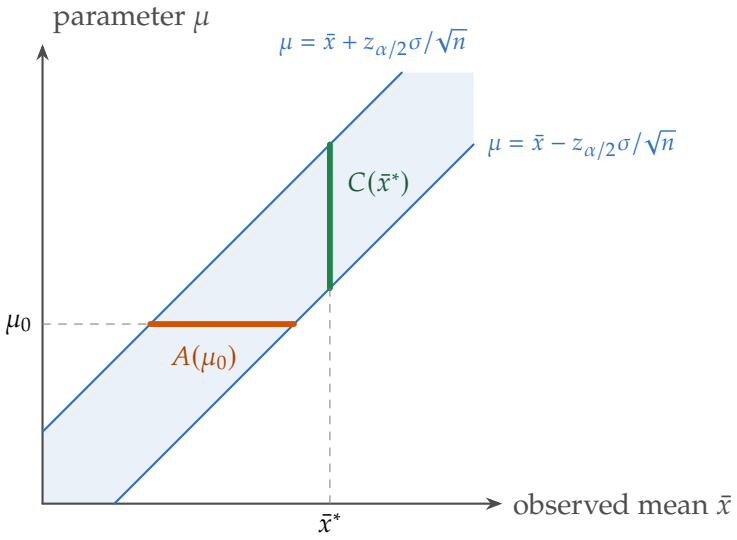


Figure 4.23: The Gaussian test and its interval are one band. The shaded band holds the pairs (observed mean, parameter) that the level-α test does not reject. A horizontal cut at a hypothesised μ_0 gives the range of data the test accepts (orange). A vertical cut at the mean actually observed gives the range of parameters the data do not reject (green): the confidence interval.

Example: the trigger efficiency, exactly. Back to the trigger: $N = 24$ muons, of which $z = 7$ fired it. The number of firings Z is binomial, Binomial(N, θ). *The question:* which θ_0 does a two-sided test at level α fail to reject? The test splits α equally between the two tails, so a candidate can fail in two ways.

- It is too *small* if, under θ_0, seeing z or more firings has probability at most $\alpha/2$.
- It is too *large* if, under θ_0, seeing z or fewer firings has probability at most $\alpha/2$.

The survivors are the θ_0 that fail neither way:

$$\mathrm{P}(Z \geq z \mid \theta_0) > \frac{\alpha}{2} \qquad \text{and} \qquad \mathrm{P}(Z \leq z \mid \theta_0) > \frac{\alpha}{2}. \tag{4.52}$$

As θ_0 grows, each muon is more likely to fire the trigger, so the first tail grows and the second shrinks. The first condition therefore cuts off small θ_0, the second cuts off large θ_0, and the survivors form an interval $[\theta_L, \theta_U]$, whose ends solve $\mathrm{P}(Z \geq z \mid \theta_L) = \alpha/2$ and $\mathrm{P}(Z \leq z \mid \theta_U) = \alpha/2$. This is the **Clopper–Pearson** interval [CasellaBerger Ex. 9.2.5]. For our trigger it is

$$\theta \in [0.126,\ 0.511] \qquad (95\%),$$

the numbers Kruschke finds for the same data [Kruschke p. 321]. The design value 0.5 is inside, which is the interval version of $p = 0.064 > 0.05$. So is 0.2, and so is every efficiency from about 0.13 to about 0.51. That answers the engineer: 0.2 and 0.4 are still possible, 0.6 is not. Seven firings in twenty-four is weak information.

Figure 4.24 shows the whole family of tests at once. For every θ_0 on the horizontal axis we plot the two-sided p-value of $H_0 : \theta = \theta_0$. This **p-value function** is largest near the observed

fraction $\hat{\theta} = 0.292$ and falls off on both sides; the confidence interval is the set where it stays above α. The curve is not a probability density for θ: its height at θ_0 is a tail probability of the data, and the area under it has no meaning (it need not be one) [Kruschke p. 323].

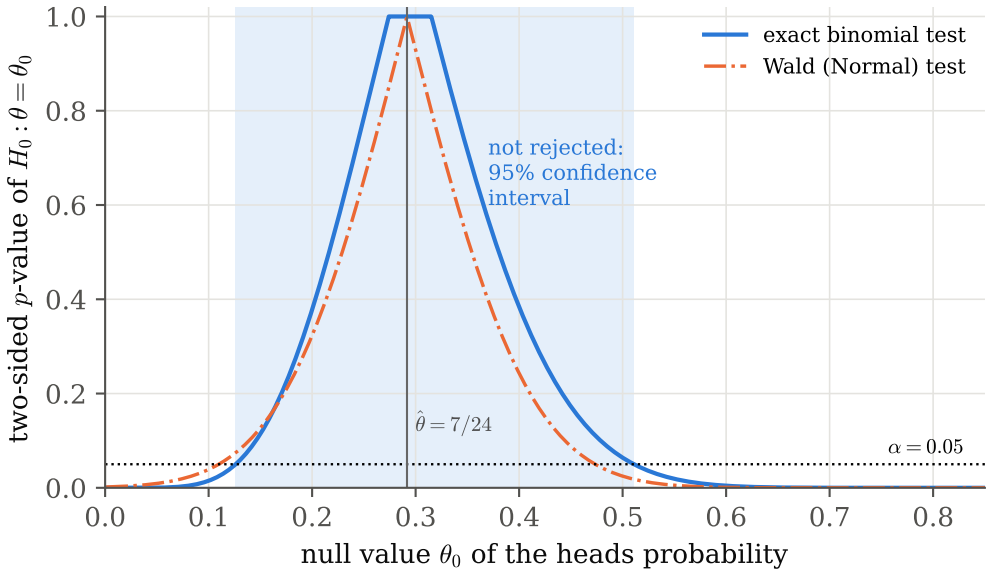


Figure 4.24: The two-sided p-value of $\theta = \theta_0$ for 7 successes in 24 trials, as a function of the hypothesised θ_0. The solid curve uses the exact binomial tails, the dash-dotted one the Normal (Wald) approximation. Where a curve stays above the dotted line at $\alpha = 0.05$, the null value survives; the shaded strip is the exact 95% confidence interval. The Wald curve is symmetric about $\hat{\theta}$ and shifted to the left of the exact one.

Two cheaper intervals for the same data. The exact interval needs a root finder. Two shortcuts avoid it. The first, the Normal (Wald) interval, treats the observed fraction $\hat{\theta} = z/N$ as Gaussian with mean θ and estimated standard error $\widehat{\mathrm{se}} = \sqrt{\hat{\theta}(1-\hat{\theta})/N}$. Its coverage is right only in the large-sample limit [Wasserman Thm. 6.16]. To see why, write $\hat{\theta}_n$ for the estimate from n trials and suppose $(\hat{\theta}_n - \theta)/\widehat{\mathrm{se}}$ tends in distribution to $Z \sim \mathcal{N}(0,1)$ (here Z is a standard normal variable, not the count). Then

$$\mathrm{P}_\theta\big(\hat{\theta}_n - z_{\alpha/2}\widehat{\mathrm{se}} < \theta < \hat{\theta}_n + z_{\alpha/2}\widehat{\mathrm{se}}\big) \overset{(1)}{=} \mathrm{P}_\theta\Big(-z_{\alpha/2} < \frac{\hat{\theta}_n - \theta}{\widehat{\mathrm{se}}} < z_{\alpha/2}\Big) \xrightarrow{(2)} \mathrm{P}\big(-z_{\alpha/2} < Z < z_{\alpha/2}\big) \overset{(3)}{=} 1 - \alpha. \tag{4.53}$$

(1) The same rearrangement of inequalities as in (4.51), now with $\hat{\theta}_n$ and $\widehat{\mathrm{se}}$.
(2) Convergence in distribution: probabilities of intervals converge to those of the limit.
(3) The definition of $z_{\alpha/2}$ puts $\alpha/2$ in each tail.

For the trigger, $\widehat{\mathrm{se}} = 0.0928$ and the interval is $[0.110, 0.474]$. It is symmetric about $\hat{\theta}$, while the exact one reaches further up, because a binomial with θ below one half has a longer right tail. The second shortcut, from Hoeffding's inequality, gives a third interval, $\hat{\theta} \pm \epsilon_N$ with $\epsilon_N^2 = \ln(2/\alpha)/(2N)$, which covers with probability at least $1 - \alpha$ for *every* N, not just in the limit [Wasserman Ex. 6.15]. Here $\epsilon_N = 0.277$ and the interval is $[0.014, 0.569]$, almost twice as long. That is the price of a guarantee that uses nothing about the binomial except that each trial is between 0 and 1. So the trade is: the Normal interval is shorter, but its coverage is only approximately right [Wasserman Ex. 6.17]. How far off it is, we measure in Section 4.12.

Example: inverting a likelihood-ratio test. Test inversion works with any test, including the likelihood-ratio test. Casella and Berger take n lifetimes from an exponential distribution with mean λ [CasellaBerger Ex. 9.2.3]. The likelihood is $\lambda^{-n}e^{-\sum x_i/\lambda}$, it is largest at $\hat{\lambda} = \sum x_i/n$, and the ratio of the likelihood at λ_0 to its maximum is

$$\frac{\lambda_0^{-n}e^{-\sum x_i/\lambda_0}}{\hat{\lambda}^{-n}e^{-n}} \overset{(1)}{=} \left(\frac{\sum x_i}{n\lambda_0}\right)^n e^{\,n}\, e^{-\sum x_i/\lambda_0}. \tag{4.54}$$

(1) Insert $\hat{\lambda} = \sum x_i/n$ in the denominator, so that $e^{-\sum x_i/\hat{\lambda}} = e^{-n}$, and collect powers.

The test retains λ_0 when this ratio is large, above some threshold k. Write $T = \sum x_i/\lambda_0$. Then the ratio is $(T/n)^n e^n e^{-T}$, so the condition is $T^n e^{-T} \geq k^*$, where the constant n^n/e^n has been moved into the new threshold $k^* = kn^n/e^n$. The function $T^n e^{-T}$ rises and then falls, so it is above a level exactly on a stretch $b \leq T \leq a$ whose ends have equal height, $a^n e^{-a} = b^n e^{-b}$. Solving $b \leq \sum x_i/\lambda_0 \leq a$ for λ_0 gives the interval $\sum x_i/a \leq \lambda \leq \sum x_i/b$.

Its coverage. The interval covers the true λ exactly when $T = \sum X_i/\lambda$ lies between b and a. Each X_i/λ is exponential with mean one, whatever λ is, so T is a sum of n exponentials of mean one, a Gamma$(n, 1)$ variable, with the same distribution for every λ. For $n = 2$ its density is te^{-t}, and

$$\mathrm{P}(b \leq T \leq a) \overset{(1)}{=} \int_b^a te^{-t}\,\mathrm{d}t \overset{(2)}{=} \left[-te^{-t}\right]_b^a + \int_b^a e^{-t}\,\mathrm{d}t \overset{(3)}{=} e^{-b}(b+1) - e^{-a}(a+1). \tag{4.55}$$

(1) The Gamma$(2, 1)$ density.
(2) Integration by parts with $u = t$, $\mathrm{d}v = e^{-t}\mathrm{d}t$.
(3) Evaluate both terms and collect.

Requiring 0.90 and $a^2e^{-a} = b^2e^{-b}$ at the same time gives $a = 5.480$, $b = 0.441$ and a confidence coefficient of 0.90006. As a check: $e^{-0.441} \times 1.441 = 0.9272$ and $e^{-5.480} \times 6.480 = 0.0270$, whose difference is 0.900; and $5.480^2e^{-5.480} = 0.125 = 0.441^2e^{-0.441}$. The interval for two lifetimes is $[\sum x_i/5.48,\ \sum x_i/0.441]$, a factor of twelve wide.

The exponential example worked so smoothly because $T = \sum X_i/\lambda$ has a distribution that does not depend on λ. Such a function of data and parameter is called a **pivot** [CasellaBerger Def. 9.2.6]. $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ is a pivot for a Gaussian mean, and $(\bar{X} - \mu)/(S/\sqrt{n})$, with the sample standard deviation S, is a pivot with a Student t_{n-1} distribution when σ is unknown. A pivot gives an interval whose coverage is the same for every parameter value, and the coverage can be computed once. Without a pivot, as for the binomial, coverage depends on θ, and we have to check it everywhere (Section 4.12).

One-sided limits. Inverting one-sided tests gives one-sided intervals. To get an *upper* bound for a Gaussian mean with unknown σ, test $\mu = \mu_0$ against $\mu < \mu_0$. The level-α test rejects when $(\bar{x} - \mu_0)/(s/\sqrt{n}) < -t_{n-1,\alpha}$, so it retains when $\mu_0 \leq \bar{x} + t_{n-1,\alpha}s/\sqrt{n}$, and the survivors form $(-\infty,\ \bar{x} + t_{n-1,\alpha}s/\sqrt{n}]$ [CasellaBerger Ex. 9.2.4]. The alternative decides the shape: an alternative of *small* values rejects large μ_0, and what survives is everything below a bound. Upper limits on signal rates in searches are built this way (Section 4.13).

4.10.3 Tests and intervals give the same verdict

Wasserman states the special case of test inversion that practitioners use every day [Wasserman Thm. 10.10]: the size-α Wald test rejects $\theta = \theta_0$ exactly when θ_0 lies outside $\hat{\theta} \pm z_{\alpha/2}\widehat{\mathrm{se}}$. It is (4.51) with $\widehat{\mathrm{se}}$ in place of $\sigma/\sqrt{n}$. Checking whether a null value is inside a 95% interval *is* a test at the 5% level.

The interval carries more information than the verdict, and that answers a common confusion between statistical and scientific significance [Wasserman p. 155]. Figure 4.25 shows two intervals that both exclude θ_0, so both tests reject. In the first, the whole interval sits close to θ_0: the effect is real but tiny, perhaps too small to matter. In the second, the interval is far away: the effect is large. A p-value alone cannot tell these apart; the interval can. A 3×10^{-4} shift in a calibration constant can be "highly significant" with enough data and still irrelevant to every physics result that uses the constant.

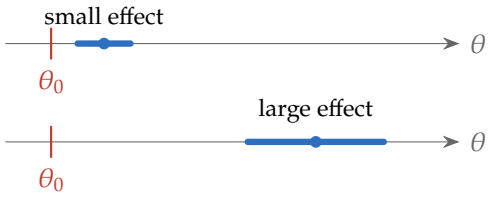


Figure 4.25: Statistical versus scientific significance. Both 95% intervals (blue bars, dots at the estimates) exclude the null value θ_0, so both data sets reject H_0 at the 5% level. In the top row the effect is close to θ_0 and may not matter in practice; in the bottom row it is large. The interval reports the size of the effect; the test only reports that it is not zero.

4.10.4 The interval inherits the test's intentions

The first part of the chapter found that a p-value depends on the stopping rule: the same 7 heads in 24 flips gave different p-values when the experimenter meant to stop at $N = 24$, at $z = 7$, or when time ran out. An interval built from those tests inherits the dependence. Kruschke works out the 95% intervals for the four intentions [Kruschke pp. 318–321]:

intention of the experimenter	95% confidence interval for θ
stop after $N = 24$ trials	[0.126, 0.511]
stop after $z = 7$ successes	[0.126, 0.484]
stop when the time is up	[0.135, 0.497]
test two coins, each with fixed N	[0.110, 0.539]

The data are identical in all four rows. What differs is the cloud of experiments that *could* have happened, and a confidence interval, like a p-value, is a statement about that cloud. For a trigger test this is not academic. A run stopped "when the beam went down" has a random N, and strictly its interval must be computed for that plan. Section 4.17 returns to this point: the Bayesian interval depends on the data only through the likelihood and does not care why the experimenter stopped.

Python. The inversion for the trigger is a few lines of `code/ch04/11_test_inversion.py`.

Listing 4.9: The exact binomial interval by test inversion (`code/ch04/11_test_inversion.py`)

```
# ---------- exact p-value function: two tails, each compared with alpha/2 ----------
def p_lower_tail(t):          # P(Z <= z | theta = t): small when t is too large
    return stats.binom.cdf(z, N, t)

def p_upper_tail(t):          # P(Z >= z | theta = t): small when t is too small
    return stats.binom.sf(z - 1, N, t)

def p_two_sided(t):
    return np.minimum(1.0, 2 * np.minimum(p_lower_tail(t), p_upper_tail(t)))

```

```
# the interval ends are where each one-sided tail equals alpha/2
lo = optimize.brentq(lambda t: p_upper_tail(t) - alpha / 2, 1e-6, z / N)
hi = optimize.brentq(lambda t: p_lower_tail(t) - alpha / 2, z / N, 1 - 1e-6)
# the same numbers from the binomial-beta identity (a check)
lo_beta = stats.beta.ppf(alpha / 2, z, N - z + 1)
hi_beta = stats.beta.ppf(1 - alpha / 2, z + 1, N - z)
assert abs(lo - lo_beta) < 1e-6 and abs(hi - hi_beta) < 1e-6
```

The two functions are the tails in (4.52): `binom.cdf(z, N, t)` is $P(Z \le z)$ and `binom.sf(z - 1, N, t)` is $P(Z \ge z)$ (the survival function is a strict inequality, hence the $z-1$). The p-value function is twice the smaller tail, capped at one. The interval ends come from a root finder: `brentq` solves "tail = $\alpha/2$" on the side of $\hat{\theta}$ where that tail is the one that shrinks. The last lines check the result against a closed form. Integrating the beta density by parts z times shows that $P(Z \ge z \mid \theta)$ equals the probability that a Beta$(z, N-z+1)$ variable is below θ, so θ_L is a beta quantile, and similarly for θ_U; the assertion confirms that the root finder and the identity agree to 10^{-6}.

Watch out: "The parameter is in the interval with probability 95%"

After the experiment the interval is two numbers, $[0.126, 0.511]$, and the efficiency is one number. Either it is inside or it is not; in the frequentist framework the probability is 0 or 1, and we do not know which [Cowan p. 123]. The 95% belongs to the recipe that produced the two numbers. Saying "the efficiency lies in $[0.126, 0.511]$ with probability 95%" is a statement about θ given the data, an inference-direction statement, and it needs a prior (Section 4.17, Chapter 5).

4.11 The Neyman construction: a belt of acceptance regions

When a calorimeter measures the energy of a jet, the shape of its response depends on the true energy θ. At low energies the estimates $\hat{\theta}$ are skewed, while at high energies they have a larger spread. From test beams or simulation we know, for every true θ, the distribution $g(\hat{\theta}; \theta)$ of the estimate. We see one number, $\hat{\theta}_{\text{obs}}$, and want an interval for θ that covers the true energy in, say, 68.3% of repeated measurements, whatever that energy is.

Test inversion (Section 4.10.2) already tells us what to do. For every θ, choose a set of $\hat{\theta}$ values that has probability $1-\alpha$ under $g(\cdot; \theta)$. Then collect the θ whose set contains the observation. Drawing all those sets at once on one graph is Neyman's construction of 1937 [Cowan §9.2], [28]. Every other frequentist interval in this chapter is a special case of it, or a shortcut to it.

The question in words. Given only the family of sampling distributions $g(\hat{\theta}; \theta)$: how do we draw a region in the (data, parameter) plane that yields, for any observed $\hat{\theta}$, an interval with guaranteed coverage? And how does the recipe look for the two workhorses, a Gaussian estimate and a Poisson count?

4.11.1 The construction, step by step

We use the convention of Feldman and Cousins: the measured quantity on the horizontal axis, the parameter on the vertical axis [28].

Step 1. *For one hypothetical θ, pick an acceptance interval.* Along the horizontal line at height θ, choose $[v_\beta(\theta), u_\alpha(\theta)]$ such that the estimate falls above u_α with probability α and

below v_β with probability β [Cowan (9.1)–(9.2)]:

$$\alpha = \int_{u_\alpha(\theta)}^{\infty} g(\hat{\theta};\theta)\,\mathrm{d}\hat{\theta}, \qquad \beta = \int_{-\infty}^{v_\beta(\theta)} g(\hat{\theta};\theta)\,\mathrm{d}\hat{\theta}. \tag{4.56}$$

Why legitimate? We only used $g(\cdot;\theta)$, which we know for every θ by assumption. The data play no role yet.

Step 2. *Repeat for every θ.* The horizontal intervals sweep out a region, the **confidence belt**, bounded by the curves $v_\beta(\theta)$ on the left and $u_\alpha(\theta)$ on the right. For every θ, by construction, $\mathrm{P}\big(v_\beta(\theta) \le \hat{\theta} \le u_\alpha(\theta)\big) = 1 - \alpha - \beta$ [Cowan (9.3)].

Step 3. *Do the experiment, draw a vertical line at $\hat{\theta}_{\text{obs}}$.* The confidence interval $[a, b]$ is the set of θ whose horizontal interval is cut by the line. *Why legitimate?* This is (4.48): the horizontal interval at θ is the acceptance region of a test of θ, and the vertical cut collects the values not rejected.

Why does the result cover? Figure 4.26 shows it. Suppose the truth is θ_{true}, a height on the graph that we do not know.

- The interval contains θ_{true} exactly when the vertical line at $\hat{\theta}_{\text{obs}}$ cuts the horizontal segment at height θ_{true}.
- That happens exactly when $\hat{\theta}_{\text{obs}}$ falls inside that segment.
- By Step 1, the estimate falls inside the segment at the true height with probability $1 - \alpha - \beta$.

So the coverage equals the probability inside the horizontal segment at the true value. We made that $1 - \alpha - \beta$ at every height, so it holds whatever the truth is.

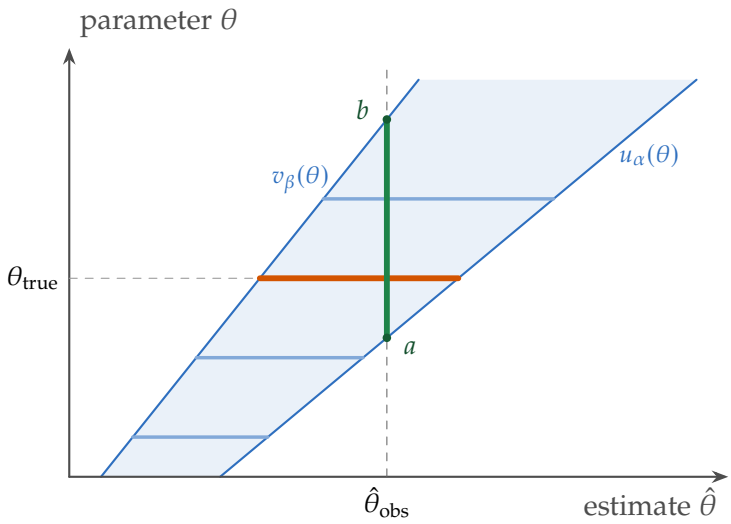


Figure 4.26: A confidence belt. For each hypothetical θ (vertical axis) the horizontal segment holds the estimates that are accepted, with probability $1 - \alpha - \beta$; a few are drawn, and the one at the true value is orange. The vertical line at the observed estimate cuts the belt between a and b (green): that is the confidence interval. The lower end a comes from the right edge of the belt, the upper end b from the left edge. The green interval contains θ_{true} exactly when the observed estimate lies on the orange segment.

Working definition 4.4 (confidence belt)

For each parameter value θ an **acceptance interval** $[v_\beta(\theta), u_\alpha(\theta)]$ of the estimate, with probability $1-\alpha-\beta$ under $g(\hat{\theta};\theta)$. The union over θ is the **confidence belt**. The interval for an observation is the vertical cut of the belt at $\hat{\theta}_{\text{obs}}$. *Operational recipe:* on a grid of θ, compute the acceptance interval from the known (or simulated) distribution of the estimate; for the observed value, keep every grid point whose interval contains it.

The same thing with inverse functions. If the estimator is a good one, both edges increase with θ: a larger true value gives larger estimates. Then they can be inverted, $a(\hat{\theta}) \equiv u_\alpha^{-1}(\hat{\theta})$ and $b(\hat{\theta}) \equiv v_\beta^{-1}(\hat{\theta})$ [Cowan (9.4)], and the coverage follows in two lines. Fix the true θ. Then

$$\mathrm{P}\big(a(\hat{\theta}) \geq \theta\big) \overset{(1)}{=} \mathrm{P}\big(\hat{\theta} \geq u_\alpha(\theta)\big) \overset{(2)}{=} \alpha, \qquad \mathrm{P}\big(b(\hat{\theta}) \leq \theta\big) \overset{(3)}{=} \mathrm{P}\big(\hat{\theta} \leq v_\beta(\theta)\big) \overset{(4)}{=} \beta,$$
$$\mathrm{P}\big(a(\hat{\theta}) < \theta < b(\hat{\theta})\big) \overset{(5)}{=} 1-\alpha-\beta. \tag{4.57}$$

(1) u_α is increasing, so applying it to both sides of $a(\hat{\theta}) \geq \theta$ preserves the inequality, and $u_\alpha(a(\hat{\theta})) = \hat{\theta}$ by the definition of the inverse.
(2) The definition of u_α in (4.56).
(3) The same step for v_β and b.
(4) The definition of v_β.
(5) The two "miss" events cannot happen together, because $a < b$ (the right edge of the belt lies to the right of the left edge), so the probability of a hit is one minus the sum of the two.

The random quantities are a and b: they depend on $\hat{\theta}$. Repeated experiments give different intervals, and a fraction $1-\alpha-\beta$ of them contain the fixed θ [Cowan p. 121].

In practice we find a and b without drawing the belt. By the definition of the inverses, $u_\alpha(a) = \hat{\theta}_{\text{obs}}$ and $v_\beta(b) = \hat{\theta}_{\text{obs}}$. Putting these into (4.56), at $\theta = a$ for the first integral and $\theta = b$ for the second, gives [Cowan (9.9)]

$$\alpha = \int_{\hat{\theta}_{\text{obs}}}^{\infty} g(\hat{\theta}; a)\,\mathrm{d}\hat{\theta}, \qquad \beta = \int_{-\infty}^{\hat{\theta}_{\text{obs}}} g(\hat{\theta}; b)\,\mathrm{d}\hat{\theta}. \tag{4.58}$$

The lower endpoint a is the parameter value for which an estimate as large as the one observed, or larger, would occur with probability α; the upper end b is the value for which an estimate this small or smaller would occur with probability β. Each end is a one-sided test with p-value exactly α (or β). Beware of the names here. The p-value of the test of $\theta = a$ is α, a small number. The confidence level of the limit $\theta \geq a$ is $1-\alpha$, a large one. To keep them apart we say "confidence level" only for the coverage of an interval and "p-value" only for tests [Cowan pp. 122–123].

Central and one-sided. The recipe leaves one choice free: how to split the excluded probability between the two sides. If the total excluded probability is γ, then $\alpha = \beta = \gamma/2$ gives a **central** interval of level $1-\gamma$. $\beta = 0$ gives a lower limit, $\alpha = 0$ an upper limit. A central interval has equal probabilities of missing on either side; it need not be symmetric about the estimate. And the choice must be made *before* looking at the data. Section 4.13 shows what goes wrong when the data decide it.

4.11.2 Gaussian estimate: the error bar is a 68.3% interval

When $g(\hat{\theta};\theta)$ is Gaussian with mean θ and known standard deviation $\sigma_{\hat{\theta}}$, the cumulative distribution is $G(\hat{\theta};\theta) = \Phi\big((\hat{\theta}-\theta)/\sigma_{\hat{\theta}}\big)$ [Cowan (9.10)]. Then (4.58) can be solved in closed form:

$$\alpha \overset{(1)}{=} 1 - \Phi\Big(\frac{\hat{\theta}_{\text{obs}} - a}{\sigma_{\hat{\theta}}}\Big) \implies a \overset{(2)}{=} \hat{\theta}_{\text{obs}} - \sigma_{\hat{\theta}}\,\Phi^{-1}(1-\alpha),$$
$$\beta \overset{(3)}{=} \Phi\Big(\frac{\hat{\theta}_{\text{obs}} - b}{\sigma_{\hat{\theta}}}\Big) \implies b \overset{(4)}{=} \hat{\theta}_{\text{obs}} - \sigma_{\hat{\theta}}\,\Phi^{-1}(\beta) \overset{(5)}{=} \hat{\theta}_{\text{obs}} + \sigma_{\hat{\theta}}\,\Phi^{-1}(1-\beta). \tag{4.59}$$

(1) The upper tail beyond $\hat{\theta}_{\text{obs}}$ of a Gaussian centred on a.
(2) Solve: $\Phi\big((\hat{\theta}_{\text{obs}} - a)/\sigma\big) = 1-\alpha$, apply Φ^{-1} and rearrange.
(3) The lower tail below $\hat{\theta}_{\text{obs}}$ of a Gaussian centred on b.
(4) Apply Φ^{-1} and rearrange.
(5) Symmetry of the Gaussian: $\Phi^{-1}(\beta) = -\Phi^{-1}(1-\beta)$.

With $\alpha = \beta = 0.1587$, $\Phi^{-1}(1-\alpha) = 1$ and the 68.3% central interval is $\hat{\theta}_{\text{obs}} \pm \sigma_{\hat{\theta}}$ [Cowan (9.13)]. This is the precise sense in which a "one-sigma error bar" is a confidence interval. The standard quantiles, which every physicist ends up knowing by heart [Cowan Tables 9.1–9.2]:

central, $1-\gamma$		one-sided, $1-\alpha$	
$\Phi^{-1}(1-\gamma/2)$	$1-\gamma$	$\Phi^{-1}(1-\alpha)$	$1-\alpha$
1	0.6827	1	0.8413
2	0.9544	2	0.9772
3	0.9973	3	0.9987
1.645	0.90	1.282	0.90
1.960	0.95	1.645	0.95
2.576	0.99	2.326	0.99

The number 1.645 appears twice: it is the central 90% quantile and the one-sided 95% one. This coincidence matters when a parameter sits at a physical boundary (Section 4.13).

Physicists use the same convention, the 68.3% central interval, for error bars even when g is not Gaussian. The interval is then no longer symmetric, and the result is quoted as $\hat{\theta}^{+d}_{-c}$, meaning $a = \hat{\theta} - c$ and $b = \hat{\theta} + d$. For example, $5.79^{+0.32}_{-0.25}$ means: the recipe that produced $[5.54, 6.11]$ covers the true value in 68.3% of repeated experiments, with equal chances of missing high and low. It does not mean that θ lies in $[5.54, 6.11]$ with probability 68.3% [Cowan p. 123]. When $\sigma_{\hat{\theta}}$ is itself estimated from the data, the pivot becomes Student's t; for large samples the Gaussian recipe with the estimated $\hat{\sigma}$ is fine, for small ones it is not, and Section 4.12 measures by how much.

4.11.3 Poisson counts: a belt with steps

A count n of rare events is Poisson with mean ν, and the estimate is $\hat{\nu} = n$ [Cowan (9.14)]. *What changes:* the estimate now takes only integer values. So (4.56) cannot be satisfied exactly. The tail probability beyond an integer changes in jumps from one integer to the next, and for a given α there is usually no integer $u_\alpha(\nu)$ whose tail is exactly α. *The fix:* keep the end conditions (4.58), and count the observed value as part of each tail [Cowan (9.15)–(9.16)]:

$$\alpha = \mathrm{P}(n \geq n_{\text{obs}}; a) = 1 - \sum_{n=0}^{n_{\text{obs}}-1} \frac{a^n}{n!} e^{-a}, \qquad \beta = \mathrm{P}(n \leq n_{\text{obs}}; b) = \sum_{n=0}^{n_{\text{obs}}} \frac{b^n}{n!} e^{-b}. \tag{4.60}$$

These can be solved numerically. There is also a closed form, because the Poisson cumulative sum is a χ^2 tail [Cowan (9.17)]:

$$\sum_{n=0}^{k} \frac{\nu^n}{n!} e^{-\nu} = \int_{2\nu}^{\infty} f_{\chi^2}\big(z;\, 2(k+1)\big)\, \mathrm{d}z, \qquad f_{\chi^2}(z;m) = \frac{z^{m/2-1} e^{-z/2}}{2^{m/2}\,\Gamma(m/2)}. \tag{4.61}$$

To prove the identity, we show that the two sides start at the same value and change at the same rate. Call the left side $F(\nu)$ and the right side $H(\nu)$. Both equal one at $\nu = 0$: on the left only the $n = 0$ term survives, and on the right the whole χ^2 density is integrated. Their derivatives are

$$\begin{aligned} F'(\nu) &\overset{(1)}{=} \sum_{n=0}^{k} \Big(\frac{n\nu^{n-1}}{n!} - \frac{\nu^n}{n!} \Big) e^{-\nu} \overset{(2)}{=} \sum_{n=1}^{k} \frac{\nu^{n-1}}{(n-1)!} e^{-\nu} - \sum_{n=0}^{k} \frac{\nu^n}{n!} e^{-\nu} \overset{(3)}{=} -\frac{\nu^k}{k!} e^{-\nu}, \\ H'(\nu) &\overset{(4)}{=} -2\, f_{\chi^2}\big(2\nu;\, 2k+2\big) \overset{(5)}{=} -2\, \frac{(2\nu)^k e^{-\nu}}{2^{k+1}\, k!} \overset{(6)}{=} -\frac{\nu^k}{k!} e^{-\nu}. \end{aligned} \tag{4.62}$$

(1) Product rule on $\nu^n e^{-\nu}$.
(2) Cancel n against $n!$ in the first sum; its $n = 0$ term is zero.
(3) Shift $n - 1 \to n$ in the first sum: it becomes the second sum without its last term, so everything cancels except $-\nu^k e^{-\nu}/k!$.
(4) Fundamental theorem of calculus, with the lower limit 2ν whose derivative is 2.
(5) Insert the density with $m = 2k + 2$: $z^{m/2-1} = (2\nu)^k$, $e^{-z/2} = e^{-\nu}$, $2^{m/2} = 2^{k+1}$, $\Gamma(k+1) = k!$.
(6) Simplify $2 \cdot 2^k / 2^{k+1} = 1$.

The two functions are equal at $\nu = 0$ and have equal derivatives everywhere, so they are equal for every ν. Applying the identity to the two conditions in (4.60) (with $k = n_{\text{obs}} - 1$ for a and $k = n_{\text{obs}}$ for b) turns each into a χ^2 probability up to $2a$ or beyond $2b$, so the limits are quantiles of χ^2 distributions [Cowan (9.18)]:

$$a = \tfrac{1}{2} F^{-1}_{\chi^2}(\alpha;\, 2n_{\text{obs}}), \qquad b = \tfrac{1}{2} F^{-1}_{\chi^2}(1-\beta;\, 2(n_{\text{obs}}+1)), \tag{4.63}$$

where $F^{-1}_{\chi^2}(p;m)$ is the p quantile with m degrees of freedom. These are Garwood's intervals of 1936 [CasellaBerger Ex. 9.2.15]. For n independent Poisson observations of rate λ the sum is Poisson$(n\lambda)$, and the interval for λ is (4.63) divided by n; with $n = 10$ observations summing to 6, the 90% interval is $[0.261,\ 1.184]$.[1]

Example: zero events. If nothing is seen, $n_{\text{obs}} = 0$, there is no lower limit and the upper limit solves $\beta = e^{-b}$, so $b = -\ln\beta$ [Cowan (9.20)]. At 95%, $b = -\ln 0.05 = 2.996 \approx 3$: if the mean were 3, seeing nothing would happen 5% of the time. This "rule of three" is how a search that sees no candidate events quotes its sensitivity. At 90% it is 2.303.

Example: the 90% table. Script `code/ch04/12_neyman_poisson_belt.py` builds the belt by brute force (for each ν on a grid, keep the counts whose two tails both exceed 5%), reads off the intervals, and checks them against (4.63):

n_{obs}	0	1	2	3	4	5	6	7	8	9	10
a	0.00	0.05	0.36	0.82	1.37	1.97	2.61	3.29	3.98	4.70	5.43
b	3.00	4.74	6.30	7.75	9.15	10.51	11.84	13.15	14.43	15.71	16.96

[1] Casella and Berger print 0.262 for the lower end; $F^{-1}_{\chi^2}(0.05;12)/20 = 5.226/20 = 0.2613$.

These are the $\alpha = 0.05$ and $\beta = 0.05$ columns of Cowan's Table 9.3. The belt is in Figure 4.27. Two features distinguish it from the Gaussian one. The edges are staircases, because the acceptance set at each ν is a set of integers. And the probability inside each horizontal set is not 0.90 but somewhere between 0.9002 and 0.9995, depending on ν: adding the observed count to each tail makes each tail at most 5%, so the interval covers at least 90% [Cowan (9.19)]. Discreteness buys safety at the price of intervals that are longer than they need to be for most ν; Section 4.12 maps the over-coverage exactly.

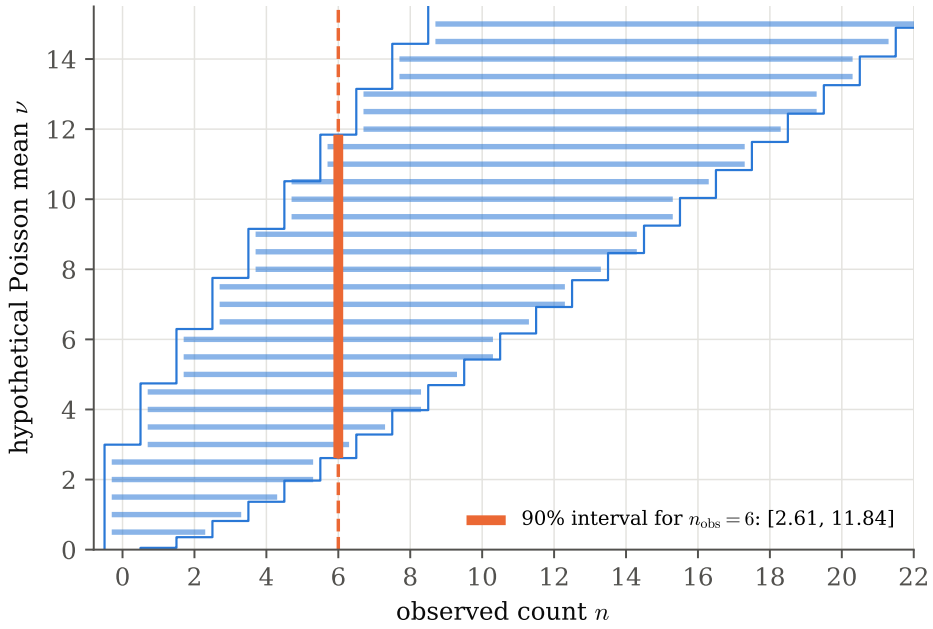


Figure 4.27: The 90% central confidence belt for a Poisson mean, built by brute force. Each horizontal bar is the set of counts accepted at one hypothetical mean ν (bars drawn every 0.5); the thin staircases are the edges of the belt on the fine grid. The dashed vertical line at $n = 6$ cuts the belt between the two ends of the orange bar, the 90% interval for six observed events.

Python. The heart of the brute-force construction:

Listing 4.10: The Poisson belt and its inversion (`code/ch04/12_neyman_poisson_belt.py`)

```
# ---------- the construction: one horizontal acceptance set per hypothetical nu ----------
nu_grid = np.arange(0.0005, 20.0, 0.0005)
n_max = 60
n = np.arange(n_max + 1)
cdf = stats.poisson.cdf(n[None, :], nu_grid[:, None])        # P(N <= n | nu)
sf = 1 - cdf + stats.poisson.pmf(n[None, :], nu_grid[:, None])  # P(N >= n | nu)
# n1(nu): the smallest n with P(N < n) <= a_tail AND P(N < n+1) > a_tail, i.e. the lowest
# count we keep.  Keep n if P(N <= n) > a_tail  and  P(N >= n) > a_tail  (both tails "not too small").
keep = (cdf > a_tail) & (sf > a_tail)
n1 = np.array([n[k].min() for k in keep])
n2 = np.array([n[k].max() for k in keep])
acc_prob = np.array([stats.poisson.pmf(n[k], nu).sum() for k, nu in zip(keep, nu_grid)])

# ---------- the inversion: the interval for n_obs collects every nu whose set contains n_obs ----------
def interval_from_belt(n_obs):
    inside = (n1 <= n_obs) & (n_obs <= n2)
    nus = nu_grid[inside]
    lo = 0.0 if n_obs == 0 else nus.min()
```

```
    return lo, nus.max()
```

`cdf` and `sf` are matrices, one row per hypothetical ν and one column per count, holding $P(N \le n)$ and $P(N \ge n)$. The boolean matrix `keep` marks the counts whose two tails are both larger than 5%: each row is one horizontal acceptance set. Inverting is then a vertical read: `interval_from_belt` collects every ν whose row contains the observed count. The script then asserts that the brute-force interval equals Garwood's formula to within the grid spacing, for $n_{\text{obs}} = 0, \ldots, 10$.

4.11.4 When the estimate is not Gaussian: transform first

For a non-Gaussian estimator we can always build the full belt. Sometimes a change of variables makes that unnecessary. The textbook case is the correlation coefficient r of n pairs from a two-dimensional Gaussian [Cowan §9.5]. Its distribution is skewed, the more so the closer the true ρ is to ± 1, and it is Gaussian only for very large n (Cowan quotes $n \gtrsim 500$).

The trouble is that the spread of r depends on the unknown ρ. Fisher's idea is to transform r so that it does not: a **variance-stabilising transformation** [a function g chosen so that the variance of $g(r)$ no longer depends on ρ]. The variance of r is approximately $(1-\rho^2)^2/n$ [Cowan (5.17)]. The delta method of Chapter 3 (a smooth function of an estimate has variance g'^2 times the estimate's variance) gives $\operatorname{Var} g(r) \approx g'(\rho)^2(1-\rho^2)^2/n$. This is independent of ρ if $g'(\rho) = 1/(1-\rho^2)$:

$$g(\rho) \overset{(1)}{=} \int \frac{\mathrm{d}\rho}{1-\rho^2} \overset{(2)}{=} \frac{1}{2}\int \left(\frac{1}{1+\rho} + \frac{1}{1-\rho}\right)\mathrm{d}\rho \overset{(3)}{=} \frac{1}{2}\ln\frac{1+\rho}{1-\rho} \overset{(4)}{=} \tanh^{-1}\rho. \tag{4.64}$$

(1) Integrate the required derivative.
(2) Partial fractions: $1/(1-\rho^2) = \frac{1}{2}[1/(1+\rho) + 1/(1-\rho)]$.
(3) $\int \mathrm{d}\rho/(1 \pm \rho) = \pm \ln(1 \pm \rho)$; choose the constant so that $g(0) = 0$.
(4) The standard form of the inverse hyperbolic tangent.

So $z = \tanh^{-1} r$ has variance close to $1/n$; a finer calculation gives $1/(n-3)$ [Cowan (9.22)–(9.25)], and z is close to Gaussian much sooner than r. The recipe: transform, build the Gaussian interval for $\zeta = \tanh^{-1}\rho$, transform the ends back with tanh.

Example: $r = 0.5$ from 20 pairs. The point first: the naive Gaussian interval claims a correlation, and the transformed one does not. *Naive.* The naive error $\hat{\sigma}_r = (1-r^2)/\sqrt{n} = 0.168$ makes r look like a three-sigma effect, and the naive 99% interval $[0.068, 0.932]$ excludes zero. *Transformed.* Now $z = \tanh^{-1} 0.5 = 0.549$ and $\sigma_z = 1/\sqrt{n-3} = 1/\sqrt{17} = 0.243$. The 99% interval for ζ is $[-0.075, 1.174]$, and applying tanh to its ends gives $[-0.075, 0.826]$ for ρ: zero is inside [Cowan p. 130]. The evidence for a correlation is much weaker than the naive error suggests. Under $\rho = 0$ the naive Gaussian gives a one-sided probability of 0.14% for $r \ge 0.5$, the z transformation gives 1.2%, and the exact calculation (under $\rho = 0$, $r\sqrt{n-2}/\sqrt{1-r^2}$ is Student t with $n-2$ degrees of freedom, here $t = 2.45$) gives 1.2%.[2] Which interval should we trust? A simulation of 10^5 samples of 20 pairs settles it. At $\rho = 0.5$ the naive 99% interval covers only 0.951 of the time, the Fisher interval 0.989; at $\rho = 0.8$, 0.941 against 0.989.

[2] Cowan quotes 2.3% for the z transformation. We find 1.2% for the one-sided tail, in agreement with the exact t calculation; 2.3% is close to the two-sided value.

4.11.5 Intervals from the likelihood function

The most widely used interval in physics needs no belt at all. In the large sample limit, the maximum-likelihood estimator is Gaussian around the true value with standard deviation $\sigma_{\hat\theta}$, and the likelihood itself becomes a Gaussian function of θ around $\hat\theta$ with the same width [Cowan (9.26)–(9.27)]; the equality of the two widths is the Cramér–Rao bound becoming an equality (Chapter 3). Then

$$\ln \mathcal{L}(\hat\theta \pm N\sigma_{\hat\theta}) \overset{(1)}{=} \ln \mathcal{L}_{\max} - \frac{(N\sigma_{\hat\theta})^2}{2\sigma^2_{\hat\theta}} \overset{(2)}{=} \ln \mathcal{L}_{\max} - \frac{N^2}{2}. \tag{4.65}$$

(1) $\ln \mathcal{L}(\theta) = \ln \mathcal{L}_{\max} - (\theta - \hat\theta)^2/2\sigma^2_{\hat\theta}$ for a Gaussian likelihood, evaluated at $\theta - \hat\theta = \pm N\sigma_{\hat\theta}$.

(2) Cancel $\sigma^2_{\hat\theta}$.

So the 68.3% interval is where $\ln \mathcal{L}$ has dropped by 1/2 from its maximum, the 95.4% interval where it has dropped by 2, and in a least-squares fit, where $\ln \mathcal{L} = -\chi^2/2$, where χ^2 has risen by N^2 [Cowan (9.28)–(9.30)]:

$$\ln \mathcal{L}\big(\hat\theta^{+d}_{-c}\big) = \ln \mathcal{L}_{\max} - \frac{N^2}{2}, \qquad \chi^2\big(\hat\theta^{+d}_{-c}\big) = \chi^2_{\min} + N^2. \tag{4.66}$$

This prescription is used, and works well, even when $\ln \mathcal{L}$ is not a parabola. *Why?* The reason is that the likelihood does not care how we label the parameter. Suppose some monotone reparametrisation $\eta = h(\theta)$ makes the likelihood Gaussian in η. The likelihood is a function on parameter space, and its *values* do not change when we relabel the points: $\mathcal{L}$ at the point called θ equals $\mathcal{L}$ at the same point called $\eta = h(\theta)$. So the points where $\ln \mathcal{L}$ has dropped by 1/2 are the same points in either labelling. In the η labelling they bound an exact 68.3% interval; mapped back, they bound an asymmetric interval in θ with the same coverage. So the recipe uses the good parametrisation without our having to find it. But it is only as good as the assumption that such an h exists, and that is guaranteed only for large samples. That is why the result is called a *likelihood interval*: it approximates the confidence interval, and is not one exactly [Cowan p. 131].

Example: five lifetimes. Take n decay times t_i from an exponential with mean τ. With $\hat\tau = \bar t$ and $u = \hat\tau/\tau$,

$$\ln \mathcal{L}(\tau) - \ln \mathcal{L}_{\max} \overset{(1)}{=} \left(-n\ln\tau - \frac{n\hat\tau}{\tau}\right) - \left(-n\ln\hat\tau - n\right) \overset{(2)}{=} n\left(\ln\frac{\hat\tau}{\tau} - \frac{\hat\tau}{\tau} + 1\right) \overset{(3)}{=} n\,(\ln u - u + 1). \tag{4.67}$$

(1) $\ln \mathcal{L} = \sum_i \ln(\tau^{-1}e^{-t_i/\tau}) = -n\ln\tau - \sum_i t_i/\tau$, and $\sum_i t_i = n\hat\tau$; at the maximum $\tau = \hat\tau$ the second term is $-n$.

(2) Collect the logarithms.

(3) Definition of u.

The drop depends on the data only through u, so the two roots of $n(\ln u - u + 1) = -\frac{1}{2}$ are fixed numbers for each n: for $n = 5$, $u = 0.617$ and $u = 1.516$. The likelihood interval is $[\hat\tau/1.516,\ \hat\tau/0.617]$ for *every* sample of five. For the sample simulated in `code/ch04/14_likelihood_intervals.py` with true $\tau = 1$, $\hat\tau = 1.58$ and the interval is $[1.04, 2.56]$, that is $\hat\tau = 1.58^{+0.98}_{-0.54}$ (Figure 4.28, left). Cowan's own sample gives $0.85^{+0.52}_{-0.30}$ [Cowan Fig. 9.6], the same ratios 1.61 and 0.65 to $\hat\tau$, as the invariance predicts.

Because $u = \hat\tau/\tau$ is a pivot (it is a mean of five exponentials of mean one, a Gamma$(5, 1/5)$ variable, whatever τ is), the coverage of the likelihood interval can be computed exactly: it is

$\mathrm{P}(0.617 \le U \le 1.516) = 0.675$, against a nominal 0.683; a Monte Carlo of 2×10^5 experiments gives 0.674. The exact central interval for our sample, from the pivot $2n\hat{\tau}/\tau \sim \chi^2_{2n}$, is $[1.10, 2.78]$. The symmetric interval $\hat{\tau} \pm \hat{\tau}/\sqrt{n}$ happens to have about the right total coverage at $n = 5$ (0.680), but it distributes its misses badly: it lies entirely below the truth in 0.266 of experiments and entirely above it in 0.053, where a central interval would have 0.159 on each side. The likelihood interval splits them as 0.199 and 0.126. Figure 4.28 (right) shows how both approach the nominal coverage as n grows. With $n = 50$ Cowan's symmetric 1.06 ± 0.15 is fine; with $n = 5$ the asymmetric likelihood interval is the honest summary.

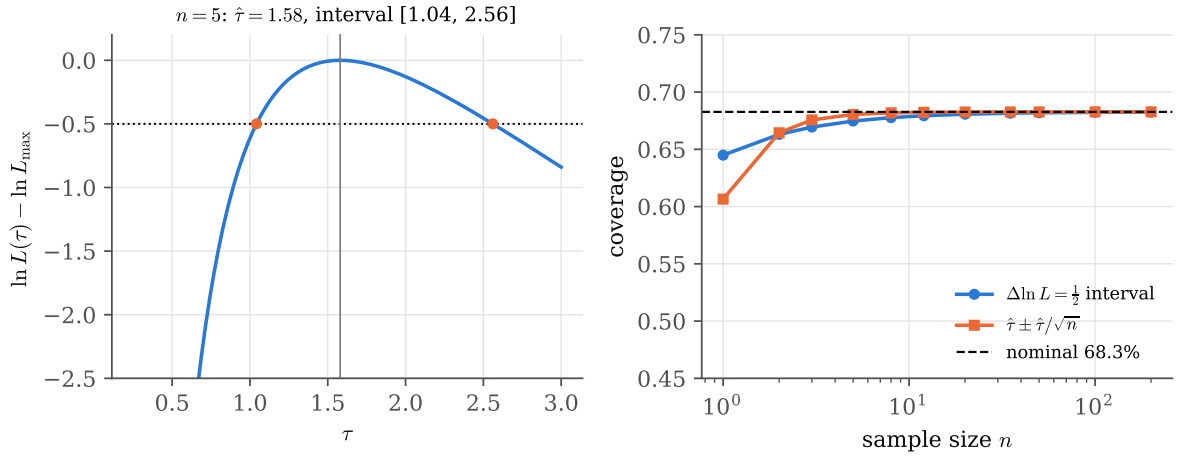


Figure 4.28: Left: the log-likelihood of five simulated decay times, relative to its maximum, as a function of the mean lifetime τ. It is not a parabola: it falls steeply towards small τ and slowly towards large τ. The dotted line is the drop by $1/2$; its two crossings (orange) are the ends of the likelihood interval. Right: the exact coverage of the $\Delta \ln \mathcal{L} = \frac{1}{2}$ interval and of the symmetric $\hat{\tau} \pm \hat{\tau}/\sqrt{n}$, as a function of the sample size; both approach 68.3%.

4.11.6 Several parameters: confidence regions

With m parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$, the large-sample estimators are jointly Gaussian with covariance $\mathbf{C}$, and the likelihood is a Gaussian function of $\boldsymbol{\theta}$ with the same $\mathbf{C}$ [Cowan (9.31)–(9.33)]. Both are governed by the quadratic form

$$Q(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^{\mathsf{T}} \mathbf{C}^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}), \tag{4.68}$$

which is symmetric in its two arguments. The argument for a confidence region takes three steps.

- *Two readings of one form.* As a function of the estimate at fixed truth, the level sets of Q are the ellipses of equal sampling density. As a function of the parameters at fixed estimate, they are the contours of equal likelihood.
- *Its distribution.* At the true $\boldsymbol{\theta}$, Q has a χ^2_m distribution (Chapter 2): whitening the Gaussian vector turns Q into a sum of m squared independent standard normals. So $\mathrm{P}(Q \le Q_\gamma) = 1 - \gamma$ when $Q_\gamma = F^{-1}_{\chi^2}(1 - \gamma; m)$, the $1 - \gamma$ quantile.
- *The symmetry.* "The estimate is within Q_γ of the truth" and "the truth is within Q_γ of the estimate" are the same event.

So the region $\{\boldsymbol{\theta} : Q(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) \le Q_\gamma\}$ contains the truth with probability $1-\gamma$: it is a $1-\gamma$ confidence region. In practice it is found as the contour [Cowan (9.34)–(9.37)]

$$\ln \mathcal{L}(\boldsymbol{\theta}) = \ln \mathcal{L}_{\max} - \frac{Q_\gamma}{2} \qquad \text{or} \qquad \chi^2(\boldsymbol{\theta}) = \chi^2_{\min} + Q_\gamma . \tag{4.69}$$

The thresholds grow with the number of parameters [Cowan Tables 9.4–9.5]:

$1-\gamma$	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$
0.683	1.00	2.30	3.53	4.72	5.89
0.90	2.71	4.61	6.25	7.78	9.24
0.95	3.84	5.99	7.82	9.49	11.1
0.99	6.63	9.21	11.3	13.3	15.1

For $m=1$, $Q_\gamma = N^2$ and the recipe is (4.66). For $m=2$ the familiar threshold no longer works. The $\Delta\chi^2 = 1$ contour, which many plots label "1σ", contains the truth only 39.3% of the time; the 68.3% region needs $\Delta\chi^2 = 2.30$. Figure 4.29 checks this with a straight-line fit to ten points, repeated 2×10^5 times: the truth falls inside the $\Delta\chi^2 = 1$ ellipse in 0.395 of the experiments and inside the $\Delta\chi^2 = 2.30$ ellipse in 0.683. The interval for the slope *alone*, $\hat\theta_2 \pm \sqrt{C_{22}}$, still covers 0.682: a one-parameter question keeps its one-parameter threshold. This is how cosmological constraints are drawn: the 68% and 95% contours in a plane of two parameters, and one-dimensional intervals from the marginal or profile of each (Chapters 9 and 10).

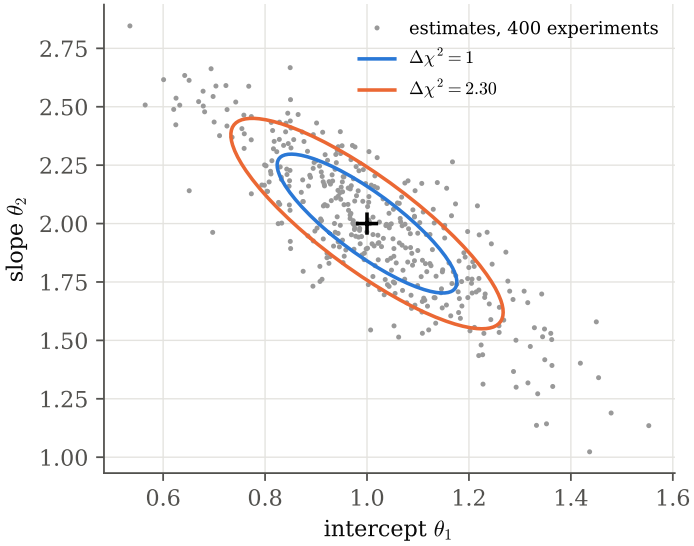


Figure 4.29: Two-parameter confidence regions for a straight-line fit. Grey dots: the fitted (intercept, slope) from 400 simulated experiments with the same true line (cross). The negative tilt is the correlation of the two estimates (correlation coefficient −0.84). The inner ellipse, $\Delta\chi^2 = 1$, holds about 39% of the dots, the outer, $\Delta\chi^2 = 2.30$, about 68%. The same ellipses drawn around any one estimate are its confidence regions.

As in one dimension, the region is random and the parameters are fixed [Cowan p. 135]. And as in one dimension, the χ^2_m calibration is only asymptotic. For strongly non-Gaussian likelihoods the coverage of (4.69) must be checked by simulation, as in Section 4.12.

4.12 Coverage is a property of the procedure

A hundred groups around the world measure the same Gaussian mean, each with $n = 10$ readings of known resolution, and each quotes the 95% interval $\bar{x} \pm 1.96\,\sigma/\sqrt{n}$. Figure 4.30 shows what happens in a simulation where we know the truth. The intervals jump up and down with the data; the true value, the dashed line, stays put; and 5 of the hundred intervals miss it. None of the hundred groups can tell from its own interval whether it is one of the unlucky ones. The 95% is a property of the whole ladder, not of any rung.

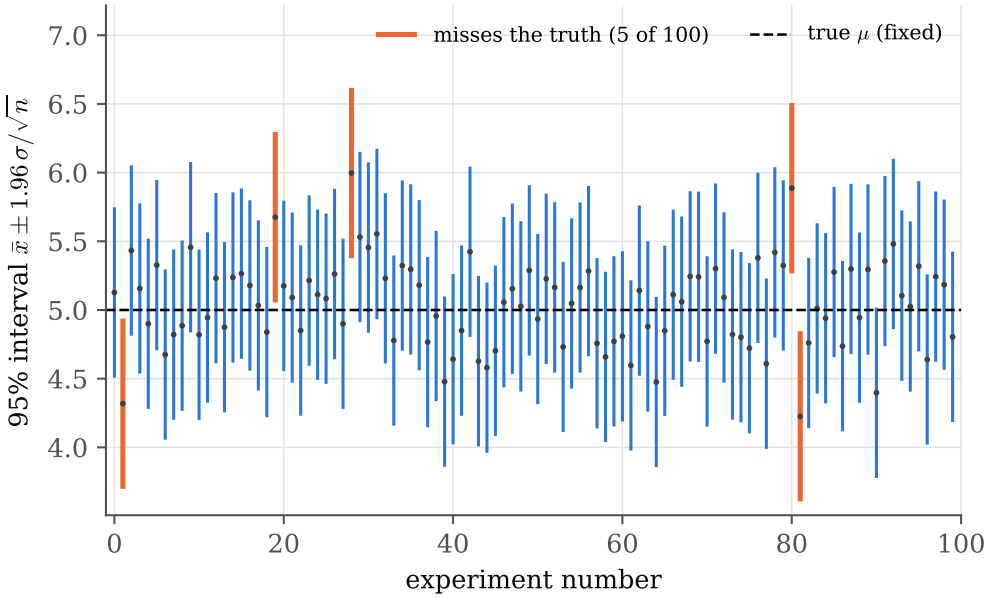


Figure 4.30: A hundred simulated experiments, each quoting a 95% interval for the same Gaussian mean. Each vertical bar is an interval around a sample mean (dot). Across experiments both the dots and the intervals move, while the true mean stays fixed at the dashed line. An orange interval misses that line.

A confidence level is a promise about a *recipe*: apply it to data generated with any θ, and it will contain that θ with the stated probability. We can check by simulation whether a recipe keeps the promise. The check finds three kinds of fault: recipes that cover less often than they claim, recipes that cover more often than needed and so are too wide, and, less obviously, recipes that are right on average but wrong in situations we can recognise from the data.

The question in words.

1. For a given recipe, how does the coverage depend on the true parameter, and what is its worst case?
2. Which common recipes undercover, which overcover, and why?
3. Is the long-run coverage the right thing to know about the one interval we actually hold?

4.12.1 The coverage function

For a recipe $x \mapsto [a(x), b(x)]$ the **coverage function** is $C(\theta) = \mathrm{P}_\theta\big(a(\boldsymbol{X}) \le \theta \le b(\boldsymbol{X})\big)$. For continuous data it is an integral, for counts a sum:

$$C(\theta) = \int \mathbb{1}\big[a(x) \le \theta \le b(x)\big]\, p(x \mid \theta)\, \mathrm{d}x, \qquad C(\nu) = \sum_{n=0}^{\infty} \mathrm{P}(n \mid \nu)\, \mathbb{1}\big[a(n) \le \nu \le b(n)\big]. \tag{4.70}$$

The indicator $\mathbb{1}[\cdot]$ is one when the interval computed from x contains θ and zero otherwise. A recipe is *exact* if $C(\theta) = 1 - \alpha$ everywhere, it **undercovers** where $C(\theta) < 1 - \alpha$ and **overcovers** where $C(\theta) > 1 - \alpha$. A recipe that overcovers somewhere and undercovers nowhere is called **conservative** [28]. The confidence coefficient of Section 4.10.1 is $\inf_\theta C(\theta)$.

Working definition 4.5 (checking coverage)

Operational recipe: choose a grid of true values θ. For each, simulate M experiments, apply the interval recipe to each, and record the fraction $\hat{C}(\theta)$ that contain θ; its Monte Carlo error is $\sqrt{\hat{C}(1 - \hat{C})/M}$. For discrete data with a finite number of outcomes, evaluate the sum (4.70) exactly instead. A recipe is acceptable for a claimed level $1 - \alpha$ if $\hat{C}(\theta) \ge 1 - \alpha$ (within errors) at every θ that physics allows.

The simplest case is a Gaussian mean [CasellaBerger Ex. 9.1.3]. Take four readings from $\mathcal{N}(\mu, 1)$ and the interval $[\bar{X} - 1, \bar{X} + 1]$. *The question:* how often does it contain μ? The mean of four readings has variance $1/4$, so $\bar{X} \sim \mathcal{N}(\mu, 1/4)$, and

$$C(\mu) \overset{(1)}{=} \mathrm{P}(-1 \le \bar{X} - \mu \le 1) \overset{(2)}{=} \mathrm{P}\Big(-2 \le \frac{\bar{X} - \mu}{\sqrt{1/4}} \le 2\Big) \overset{(3)}{=} \mathrm{P}(-2 \le Z \le 2) = 0.9544. \tag{4.71}$$

(1) Rearrange $\bar{X} - 1 \le \mu \le \bar{X} + 1$.
(2) Divide by the standard deviation $\sqrt{1/4} = 1/2$.
(3) The standardised mean is standard normal whatever μ is.

The coverage is the same for every μ because $\bar{X} - \mu$ is a pivot: its distribution does not depend on μ (Section 4.10.2). Most recipes have no pivot, and then the coverage depends on the parameter.

Example: two intervals for a uniform endpoint. Let $X_1, \dots, X_n$ be uniform on $[0, \theta]$ and $Y = \max X_i$ [CasellaBerger Ex. 9.1.6]. Since $Y \le y$ exactly when every $X_i \le y$, $\mathrm{P}(Y \le y) = (y/\theta)^n$, and $T = Y/\theta$ has density $n t^{n-1}$ on $[0, 1]$, whatever θ is. Compare two recipes, $[aY, bY]$ with $1 \le a < b$ and $[Y + c, Y + d]$ with $0 \le c < d$:

$$\begin{aligned} \mathrm{P}_\theta(aY \le \theta \le bY) &\overset{(1)}{=} \mathrm{P}\Big(\frac{1}{b} \le T \le \frac{1}{a}\Big) \overset{(2)}{=} \int_{1/b}^{1/a} n t^{n-1}\, \mathrm{d}t = \Big(\frac{1}{a}\Big)^n - \Big(\frac{1}{b}\Big)^n, \\ \mathrm{P}_\theta(Y + c \le \theta \le Y + d) &\overset{(3)}{=} \mathrm{P}\Big(1 - \frac{d}{\theta} \le T \le 1 - \frac{c}{\theta}\Big) \overset{(4)}{=} \Big(1 - \frac{c}{\theta}\Big)^n - \Big(1 - \frac{d}{\theta}\Big)^n \xrightarrow{\theta \to \infty} 0. \end{aligned} \tag{4.72}$$

(1) Divide $aY \le \theta \le bY$ by θ and invert: $1/b \le Y/\theta \le 1/a$.
(2) The density of T.
(3) Subtract Y and divide by θ (for $\theta \ge d$).
(4) Same integral with the new limits.

The scale recipe has the same coverage for every θ, because $T = Y/\theta$ is a pivot. The shift recipe does not. Its coverage depends on θ and tends to zero for large θ, because a fixed shift becomes

negligible when the endpoint is large. So its worst-case coverage, the confidence coefficient, is zero. Used for, say, the maximum energy of a spectrum, it would be a disastrous recipe, even though for some θ it looks fine.

4.12.2 The Gaussian mean, with and without a known σ

Simulating 200000 experiments of the kind in Figure 4.30 gives a coverage of 0.9498 ± 0.0005 for the 95% interval and 0.6814 for $\bar{x} \pm \sigma/\sqrt{n}$, as it must: this recipe is exact. Things go wrong when σ is not known and is replaced by the sample standard deviation s. The Normal quantile 1.96 assumes the scale is fixed, but s fluctuates from sample to sample. The quantity that does have a fixed distribution is $(\bar{x} - \mu)/(s/\sqrt{n})$, which follows Student's t_{n-1} (from the first part of the chapter). The simulation measures how much the Normal quantile loses:

readings n	3	5	10	30
$\bar{x} \pm 1.96\, s/\sqrt{n}$	0.810	0.879	0.918	0.941
$\bar{x} \pm t_{n-1,0.025}\, s/\sqrt{n}$	0.949	0.951	0.950	0.951

With three readings the "95%" interval covers only 81% of the time; the t quantile restores 95% at every n. The failure is systematic, not a fluctuation, and it is invisible in any single experiment. A calibration run with three repeated measurements, quoted with a Normal quantile, claims more precision than it has.

4.12.3 Counts: over-coverage and a saw-tooth

For a Poisson mean the sum in (4.70) can be done exactly on a fine grid of ν, and Figure 4.31 shows the result for two recipes: Garwood's interval (4.63) and the quick $n \pm z\sqrt{n}$ (the Wald interval for a count). Three things stand out.

First, Garwood's interval never undercovers. Its 90% version has coverage at least 0.903 and on average 0.939 over $0 < \nu < 15$; its 68.3% version has minimum 0.689 and mean 0.782. It is conservative, as (4.60) was designed to be. Feldman and Cousins insist that this is a defect, not a virtue: an interval that claims 68% and delivers 78% is wider than needed, and wider intervals mean less power to exclude false hypotheses [28]. With discrete data some over-coverage is unavoidable; the goal is to have as little of it as possible.

Second, the coverage is a saw-tooth. The reason is in the sum (4.70): count n contributes its probability only if its interval $[a(n), b(n)]$ contains ν. As ν increases past one of the interval ends, one count suddenly stops (or starts) contributing, and $C(\nu)$ jumps [CasellaBerger p. 352]. Between jumps the same counts contribute, and their probabilities change smoothly with ν. So every tooth sits at an entry of the table of interval ends.

Third, the quick recipe fails badly at small counts. At $\nu = 1$ its nominal 90% interval covers only 0.628 of the time (Garwood: 0.981), and as $\nu \to 0$ its coverage drops to 0.020. The reason is visible in the formula: $n = 0$ gives the interval $[0, 0]$, which contains no positive ν, and for small ν zero is the most likely count. The same mechanism ruins the Normal interval for a binomial proportion near 0 or 1.

The script `code/ch04/13_coverage.py` computes these curves with a matrix of Poisson probabilities:

Listing 4.11: Exact coverage of Poisson interval recipes (`code/ch04/13_coverage.py`)

```
def garwood(nobs, cl):
    a = (1 - cl) / 2
    lo = np.where(nobs == 0, 0.0, 0.5 * stats.chi2.ppf(a, 2 * np.maximum(nobs, 1)))
```

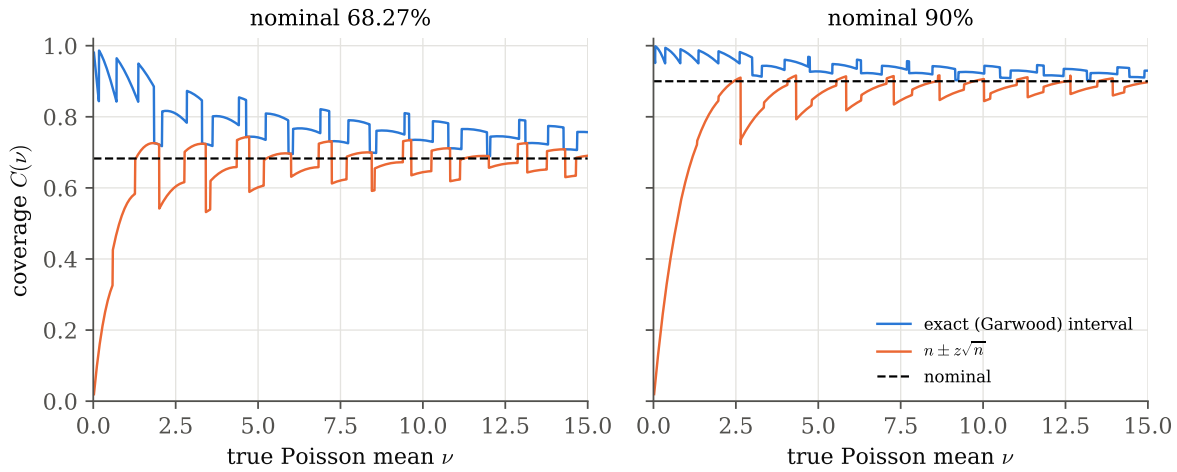


Figure 4.31: Exact coverage of two interval recipes for a Poisson mean, as a function of the true mean. Left: nominal 68.27%, right: nominal 90% (dashed lines). Garwood's exact interval (blue) always lies above the nominal level, with saw-tooth jumps where the true mean crosses an interval end. The quick interval $n \pm z\sqrt{n}$ (orange) undercovers, dramatically so below a mean of about two.

```
    hi = 0.5 * stats.chi2.ppf(1 - a, 2 * (nobs + 1))
    return lo, hi

def wald(nobs, cl):
    zq = stats.norm.isf((1 - cl) / 2)
    return nobs - zq * np.sqrt(nobs), nobs + zq * np.sqrt(nobs)

nmax = 80
nn_all = np.arange(nmax + 1)
nu = np.linspace(0.02, 15, 3000)
pmf = stats.poisson.pmf(nn_all[None, :], nu[:, None])

def coverage(method, cl):
    lo, hi = method(nn_all, cl)
    inside = (lo[None, :] <= nu[:, None]) & (nu[:, None] <= hi[None, :])
    return (pmf * inside).sum(1)

cov_g68, cov_g90 = coverage(garwood, 0.6827), coverage(garwood, 0.90)
cov_w68, cov_w90 = coverage(wald, 0.6827), coverage(wald, 0.90)
```

Each recipe returns the interval ends for all counts $0, \ldots, 80$ at once. The boolean matrix `inside` has one row per true ν and one column per count, and is true where that count's interval contains that ν. Multiplying by the Poisson probabilities and summing along each row is (4.70). A brute-force check at $\nu = 2.7$ draws 200000 counts, builds their intervals and counts hits: the Monte Carlo coverage 0.9796 agrees with the exact 0.9793 (and for the quick recipe the simulation gives 0.7442).

Pointwise versus uniform. The Poisson curve makes concrete a distinction between two meanings of "correct for large samples" [Wasserman §6.5]. Take the Normal interval for a proportion p from n trials, with estimate $\hat{p}$ and coverage $C(p)$. For each *fixed* p, $C(p)$ tends to $1 - \alpha$ as $n \to \infty$: the interval is *pointwise* asymptotic. But at any fixed n, a small enough p makes $\hat{p} = 0$ almost certain; the interval then collapses to $\{0\}$, and the coverage is near zero. So the worst case $\inf_p C(p)$ does not tend to $1 - \alpha$, and the interval is not *uniformly* asymptotic. "Large sample" has to be large compared with something, and near a boundary the something is the number of events, not the number of trials.

4.12.4 Right on average, wrong in a recognisable case

Coverage is an average over all possible data sets. Sometimes the data we hold tell us that we are in a subset of those data sets where the recipe is right much more often, or much less often, than the average. An example due to Berger and Wolpert shows this plainly [Wasserman Ex. 6.14]. Let θ be a fixed number, let X_1, X_2 be independent, each $+1$ or -1 with probability one half, and observe only $Y_i = \theta + X_i$. The "interval", a single point, is

$$C = \begin{cases} \{Y_1 - 1\} & \text{if } Y_1 = Y_2, \\ \{(Y_1 + Y_2)/2\} & \text{if } Y_1 \neq Y_2. \end{cases}$$

There are four equally likely cases:

$$\mathrm{P}_\theta(\theta \in C) \overset{(1)}{=} \tfrac{1}{4}\,\mathbb{1}[(\theta+1)-1=\theta] + \tfrac{1}{4}\,\mathbb{1}[(\theta-1)-1=\theta] + \tfrac{2}{4}\,\mathbb{1}[\tfrac{1}{2}(2\theta)=\theta] \overset{(2)}{=} \tfrac{1}{4} + 0 + \tfrac{1}{2} = \tfrac{3}{4}. \tag{4.73}$$

(1) $(X_1, X_2) = (+1, +1)$: both Y equal $\theta + 1$, and $C = \{\theta\}$. $(-1, -1)$: both equal $\theta - 1$, and $C = \{\theta - 2\}$. $(+1, -1)$ or $(-1, +1)$: the Y differ and their average is θ.
(2) Evaluate the indicators.

So C is a 75% confidence set for every θ.

What we observe. Now suppose the data are $Y_1 = 15$, $Y_2 = 17$. *The question.* Which of the four cases could have produced these data? The two Y differ, so $(+1, +1)$ and $(-1, -1)$ are ruled out; only the mixed cases remain, and in both of them the average of the Y is θ. So $C = \{16\}$ is *certainly* right. Had we instead seen $Y_1 = Y_2$, only the first two cases could have produced it, and the recipe is right in one of them and wrong in the other: right half the time. A simulation of 4×10^5 repetitions in `code/ch04/11_test_inversion.py` confirms it: overall coverage 0.7498, coverage 1.0000 when $Y_1 \neq Y_2$ (a fraction $1 - 0.5009$ of the time), and 0.5005 when $Y_1 = Y_2$.

The lesson: "75%" is a statement about the recipe, not about this data set. Calling $\{16\}$ a 75% confidence set is still correct, because that is what the recipe guarantees [Wasserman p. 93]. A probability statement about θ given these particular Y is a different kind of statement, the one Bayesian inference makes (Section 4.17). Frequentists handle the problem by *conditioning*: they compute the coverage only over data sets that share a feature of the observed one, a feature that carries no information about θ but does tell us how precise the experiment turned out to be (here, whether $Y_1 = Y_2$). In physics the classic case is a measurement whose resolution depends on an event-by-event quantity: the relevant error bar is the one for the resolution actually obtained.

How this is used in real research: Coverage studies: verification and research

A coverage simulation can play two different roles. When an analytic result exists, it *verifies*: the t interval must cover 95% at every n, the Garwood interval must never undercover, and a simulation that disagrees has a bug. When no analytic result exists, the same simulation *is* the method. Experiments at the LHC build Neyman belts from toy Monte Carlo when the test statistic has no known distribution; neutrino-oscillation fits construct Feldman–Cousins regions from millions of pseudo-experiments; cosmological analyses check whether a posterior interval from an MCMC chain (Chapter 9) covers the input parameters of simulated skies. In every case the recipe is (4.70) evaluated by brute force.

4.13 Upper limits near a physical boundary

The squared mass of a neutrino can be estimated from a measured energy and momentum as $\widehat{m^2} = E^2 - p^2$. The true m^2 cannot be negative, but the estimate can: both E^2 and p^2 carry measurement errors, and when the true mass is tiny, their difference comes out negative about half the time [Cowan §9.8]. The same happens to any parameter estimated as a difference, $\hat{\theta} = x - y$: a signal rate estimated as "counts minus expected background", a dark-matter cross-section, a tensor-to-scalar ratio. Physics says $\theta \geq 0$. The data, fluctuating, do not know that.

A search that sees nothing reports an *upper limit*. Near the boundary every recipe met so far runs into trouble. The interval can come out entirely in the forbidden region, or be empty, or be so short that it claims more than the experiment can know. Each fix trades one good property for another, and seeing what each one gives up shows what a limit really promises.

The question in words.

1. What does the classical recipe give when the estimate is negative, and is it wrong?
2. Which alternatives exist (shifting the estimate, a Bayesian prior on $\theta \geq 0$, Feldman and Cousins' ordering, CL_s), and what does each do to the coverage?
3. Why must the choice between "upper limit" and "two-sided interval" be made before the data are seen?

4.13.1 The classical limit can be negative

Let $\hat{\theta}$ be Gaussian with mean θ and known $\sigma_{\hat{\theta}}$, with $\theta \geq 0$ known a priori. The one-sided classical upper limit at confidence level $1 - \beta$ is, from (4.59),

$$\theta_{\mathrm{up}} = \hat{\theta}_{\mathrm{obs}} + \sigma_{\hat{\theta}}\, \Phi^{-1}(1 - \beta) \tag{4.74}$$

[Cowan (9.39)]. With $\sigma_{\hat{\theta}} = 1$ and $\hat{\theta}_{\mathrm{obs}} = -2$, the 95% limit is $-2 + 1.645 = -0.355$. The experiment "excludes" every physical value of θ, including zero.

Is the recipe wrong? No. It covers the true value in 95% of experiments, whatever that value is. If $\theta = 0$, an estimate below -1.645 happens 5% of the time, and those are exactly the experiments whose limit is negative. We have simply landed in the 5%. *Is the result useful?* Also no. We knew before the experiment that $\theta \geq 0$, and so certainly that $\theta > -0.355$. Two tempting repairs make it worse. Raising the confidence level to 99% moves the limit to 0.326, a limit three times smaller than the experiment's resolution, quoted at very high confidence; and choosing the level 97.725% gives $\theta_{\mathrm{up}} = 10^{-5}$, which is absurd [Cowan p. 137].

Whatever limit we quote, the first rule is to report the estimate itself, $\hat{\theta}_{\mathrm{obs}} \pm \sigma_{\hat{\theta}}$, even when it is negative. The reason is combination. Unbiased estimates from many experiments average to the truth only if the negative ones are kept; setting them to zero pushes every average upwards. When $\hat{\theta}$ is far from Gaussian, the whole likelihood function should be published so that others can combine it [Cowan p. 137].

4.13.2 Three repairs: shift, Bayes, and flip-flop

Shift the estimate to zero. A common practice replaces a negative estimate by zero before applying (4.74) [Cowan (9.40)]: $\theta_{\mathrm{up}} = \max(\hat{\theta}_{\mathrm{obs}}, 0) + \sigma_{\hat{\theta}}\Phi^{-1}(1 - \beta)$. At 95% the limit is then

never below $1.645\,\sigma_{\hat\theta}$ (1.645 for our example), so it is never smaller than about the resolution. The shifted limit is never below the classical one, so it contains the truth at least as often: it overcovers. In Figure 4.32 (right) its coverage is 1 for all $\theta \le 1.645$, because every experiment then quotes a limit above θ.

A Bayesian limit with a prior that knows $\theta \ge 0$. In the Bayesian approach the boundary is built in by a prior that vanishes for $\theta < 0$. With a flat prior on $\theta \ge 0$ [Cowan (9.41)–(9.44)] and $\sigma_{\hat\theta} = 1$, the posterior is the Gaussian likelihood cut off at zero and renormalised. Write $x = \hat\theta_{\text{obs}}$ and φ for the standard normal density. The upper limit solves

$$1-\beta \overset{(1)}{=} \frac{\int_0^{\theta_{\text{up}}} \varphi(\theta - x)\,\mathrm{d}\theta}{\int_0^{\infty} \varphi(\theta - x)\,\mathrm{d}\theta} \overset{(2)}{=} \frac{\Phi(\theta_{\text{up}} - x) - \Phi(-x)}{1 - \Phi(-x)} \overset{(3)}{=} \frac{\Phi(\theta_{\text{up}} - x) - 1 + \Phi(x)}{\Phi(x)},$$

$$\Longrightarrow \quad \theta_{\text{up}} \overset{(4)}{=} x + \Phi^{-1}\big(1 - \beta\,\Phi(x)\big). \tag{4.75}$$

(1) Bayes' theorem with a flat prior on $\theta \ge 0$: the posterior is the likelihood $\varphi(\theta - x)$ normalised over the physical region [Cowan (9.43)].
(2) Substitute $t = \theta - x$; the integrals are differences of the normal CDF.
(3) Symmetry, $\Phi(-x) = 1 - \Phi(x)$, in numerator and denominator.
(4) Multiply by $\Phi(x)$, solve for $\Phi(\theta_{\text{up}} - x) = 1 - \beta\Phi(x)$, and apply Φ^{-1}.

For $x = -2$ this gives $\theta_{\text{up}} = 1.05$. For large positive x, $\Phi(x) \to 1$ and the limit tends to the classical $x + 1.645$. The Bayesian limit is always positive and always above the classical one (Cowan's Fig. 9.8, our Figure 4.32, left).

It means something different. It is the value that θ exceeds with posterior probability 5%, given the prior. So it depends on the prior, and the flat prior has two weaknesses. First, it says that $0 < \theta < 1$ is as plausible before the experiment as $10^{40} < \theta < 10^{40} + 1$. Second, it changes when we relabel the parameter: a flat prior on m^2 is not flat on m. The prior $\pi(\theta) \propto 1/\theta$, flat in $\ln\theta$, gives limits invariant under $\theta \to \theta^k$, but the integrals diverge unless the likelihood vanishes at $\theta \to 0$, which a Gaussian likelihood does not [Cowan (9.45) and p. 139]. In practice the flat prior is the common choice, with its assumptions stated.

The Bayesian limit was not built to have coverage, but we can still measure its coverage. Figure 4.32 (right) shows that for this problem it overcovers near the boundary and approaches 95% from above for large θ (minimum 0.950).

Flip-flopping. Feldman and Cousins describe a third habit, common enough that they gave it a name [28]. Physicist X decides after seeing the data: "if the result is less than 3σ, I quote a 90% upper limit; if it is more, I quote a 90% central interval; and if it is negative, I pretend it was zero." Each of the two recipes alone covers 90%. The mixture does not. *The question:* take $\sigma = 1$ and a true mean $\mu = 2$; how often does the mixture contain 2? Here x is the measured value. Take the two branches one at a time.

- The upper-limit branch ($x < 3$) quotes $[0, \max(x, 0) + 1.28]$, which contains 2 when $x \ge 0.72$.
- The central branch ($x \ge 3$) quotes $[x - 1.645, x + 1.645]$, which contains 2 when $x \le 3.645$.

So

$$C(2) \overset{(1)}{=} \mathrm{P}(0.72 \le x < 3) + \mathrm{P}(3 \le x \le 3.645) \overset{(2)}{=} \mathrm{P}(-1.28 \le x - 2 \le 1.645)$$
$$\overset{(3)}{=} \Phi(1.645) - \Phi(-1.28) \overset{(4)}{=} 0.95 - 0.10 = 0.85. \tag{4.76}$$

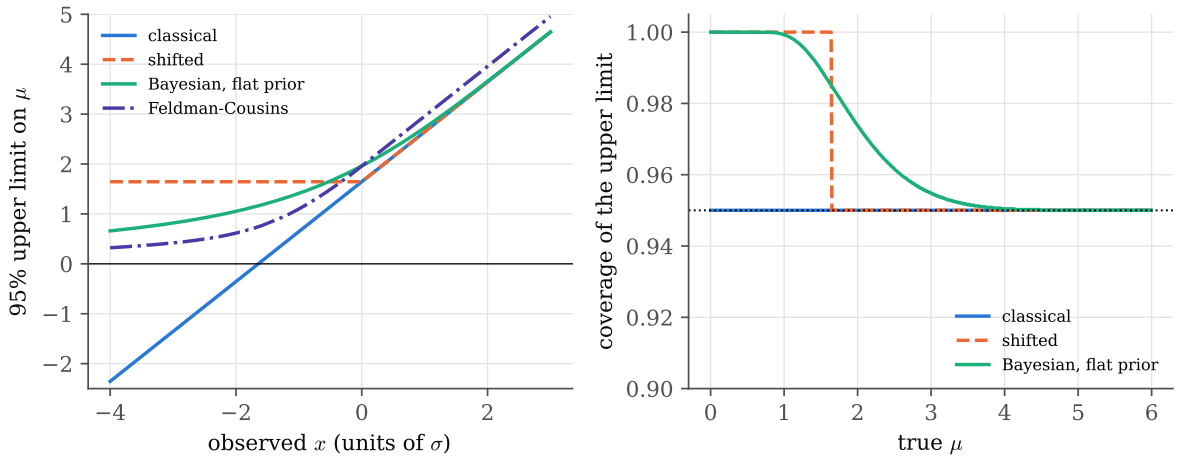


Figure 4.32: Left: 95% upper limits on a non-negative Gaussian mean as a function of the observed estimate, in units of the resolution. The classical limit (blue) goes negative for estimates below -1.645; the shifted limit (orange dashed) is flat at 1.645 for negative estimates; the flat-prior Bayesian limit (green) and the Feldman–Cousins upper end (violet) stay positive and fall slowly. All agree for large estimates. Right: the coverage of three of the recipes as a function of the true value. The classical limit covers exactly 95%, the other two more near the boundary.

(1) The two branches, each with its own condition for covering $\mu = 2$.
(2) The two ranges join into one; subtract the true value 2.
(3) $x - 2$ is standard normal.
(4) The standard quantiles of Section 4.11.2.

The quoted "90%" intervals cover 85% of the time. The missing 5% are the measurements above 3.645. The upper-limit recipe alone would have given them a limit above 2, but the switch hands them a central interval that misses. The script finds the same 0.850 at $\mu = 2$, and that this is the minimum; the coverage stays at 85% from $\mu \approx 1.3$ to $\mu \approx 4.3$ (Figure 4.35, left). In belt language (Figure 4.33, dashed), switching recipes at $x = 3$ bends the belt so that the horizontal acceptance intervals at intermediate μ hold only 85% of the probability.

Watch out: Deciding the form of the result after seeing the data

Choosing between an upper limit and a two-sided interval according to how the data came out destroys the coverage of both, even though each recipe is correct alone. The rule of the Neyman construction is that the acceptance regions are fixed before the measurement. If we want the result to switch automatically from a limit to an interval as the signal grows, the switch must be part of the construction itself, which is what the unified approach below does.

The strict alternative, always quoting the central interval, has its own problem. With $\sigma = 1$ and $x = -1.8$, the 90% central interval has upper end $x + 1.645 = -0.155$, so it contains no physical value at all. The interval is *empty* [28]. It is still a correct member of an ensemble with 90% coverage, but not one anybody wants to publish.

4.13.3 The unified approach of Feldman and Cousins

The Neyman construction (Section 4.11.1) leaves one choice open: *which* values of the measurement to put in the acceptance interval at each μ, so that it holds probability $1 - \alpha$. Central intervals leave out equal tails on both sides; upper limits leave out only the lowest values,

keeping everything above a cut. Feldman and Cousins choose by a likelihood ratio.

Key idea: Likelihood-ratio ordering

At each μ, rank the possible outcomes x by

$$R(x) = \frac{P(x \mid \mu)}{P(x \mid \mu_{\text{best}})},$$

where μ_{best} is the *physically allowed* value of μ that makes x most probable. Add outcomes to the acceptance interval in order of decreasing R until their total probability reaches $1 - \alpha$ [28, (4.1)]. Then build the belt and read the intervals as usual. The coverage is right by construction, the interval is never empty, and the result turns by itself from an upper limit into a two-sided interval as the data move away from the boundary.

Why this ordering? R compares how well μ explains x with how well the best physical value explains it. An outcome with R close to one is one that μ explains nearly as well as any allowed value could. Putting such outcomes in the acceptance set first means that we reject μ only for data that some other physical μ explains much better. This is the likelihood-ratio test of the first part of the chapter, with the physical boundary written into the maximisation, inverted into an interval.

Gaussian with $\mu \geq 0$. With $P(x \mid \mu) = \varphi(x - \mu)$, the best physical value is $\mu_{\text{best}} = \max(0, x)$. Then

$$R(x) \overset{(1)}{=} \begin{cases} \varphi(x-\mu)/\varphi(0) = e^{-(x-\mu)^2/2}, & x \geq 0, \\ \varphi(x-\mu)/\varphi(x) = e^{-(x-\mu)^2/2 + x^2/2} = e^{\,x\mu - \mu^2/2}, & x < 0, \end{cases} \tag{4.77}$$

(1) For $x \geq 0$, $\mu_{\text{best}} = x$ and $\varphi(0) = 1/\sqrt{2\pi}$; for $x < 0$, $\mu_{\text{best}} = 0$, and expanding $-(x-\mu)^2/2 + x^2/2 = x\mu - \mu^2/2$ [28, (4.2)–(4.3)].

For $x \geq 0$, R decreases with the distance from μ, so the acceptance interval is centred, as in a central interval. For $x < 0$, R decreases as x becomes more negative, but only linearly in the exponent, much more slowly than the Gaussian: negative outcomes are penalised less, because no physical μ explains them well. The acceptance interval $[x_1, x_2]$ at each μ is found numerically from $R(x_1) = R(x_2)$ and $\int_{x_1}^{x_2} \varphi(x - \mu)\mathrm{d}x = 1 - \alpha$ [28, (4.4)].

Where does the limit turn into a two-sided interval? The acceptance set at $\mu = 0$ decides it. There $R(x) = 1$ for all $x < 0$ and $R(x) = e^{-x^2/2}$ for $x \geq 0$, so every negative x is ranked first (probability $1/2$), followed by x from 0 upwards until the total is 0.90: $\Phi(x_2) = 0.90$, $x_2 = 1.28$. For an observation above 1.28, $\mu = 0$ is excluded and the interval is two-sided; below it, the interval starts at zero and is an upper limit. The brute-force belt in `code/ch04/15_feldman_cousins.py` reproduces Feldman and Cousins' Table X at 90%:

x_0	-2	-1	0	1	2	3
$[\mu_1, \mu_2]$	$[0.00, 0.40]$	$[0.00, 0.81]$	$[0.00, 1.64]$	$[0.00, 2.64]$	$[0.58, 3.64]$	$[1.37, 4.64]$

Here x_0 is the observed measurement and $[\mu_1, \mu_2]$ its interval. At $x_0 = 0$ the 90% upper limit is 1.64: the classical 95% limit rather than the 90% one (1.28). The reason is that at $\mu \approx 1.6$ the unified acceptance interval is already nearly central, so its left edge sits 1.645 below μ, and 1.645 is also the one-sided 95% quantile (Section 4.11.2). That is the price of repairing flip-flopping: near the boundary the unified limits are somewhat looser [28]. In exchange no interval is empty,

the limits for very negative x_0 shrink only slowly (as $1/|x_0|$), and the coverage is 90% for every μ (between 0.8998 and 0.9002 on our grid; Figure 4.35, left).

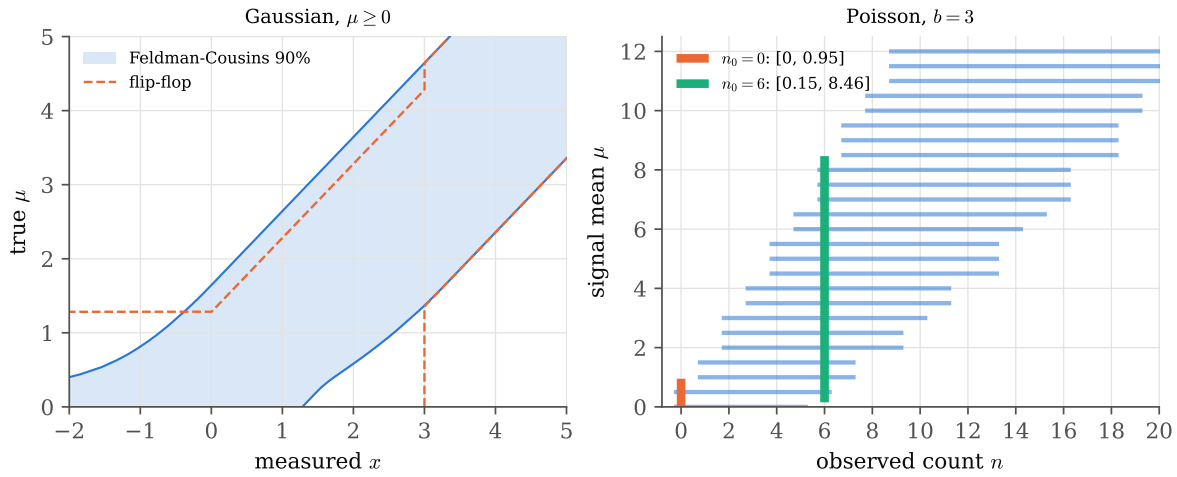


Figure 4.33: Left: the 90% Feldman–Cousins belt for a Gaussian measurement of a non-negative mean (shaded, solid edges) and the belt implied by flip-flopping (dashed). Below $x \approx 1.28$ the unified belt reaches down to $\mu = 0$, so the interval is an upper limit; above, it is two-sided. The flip-flop belt jumps at $x = 3$ where the recipe switches. Right: the Feldman–Cousins acceptance sets for a Poisson count on a known background $b = 3$ (bars every 0.5 in μ), with the intervals for $n_0 = 0$ (orange) and $n_0 = 6$ (green).

4.13.4 Counting on a known background

The most common version of the problem in particle physics is a count n of candidate events, of which a known mean b is expected from background and an unknown mean $\mu \geq 0$ from signal [Cowan §9.9]. The total is Poisson with mean $\mu + b$,

$$P(n \mid \mu) = \frac{(\mu + b)^n}{n!}\, e^{-(\mu+b)}, \tag{4.78}$$

and the naive estimate is $\hat{\mu} = n - b$, which is negative whenever fewer events than the expected background are seen [Cowan (9.47)–(9.48)]. The classical limits are the background-free ones shifted down by b [Cowan (9.50)]. For $b = 3$ and no events observed, the 90% upper limit is $2.303 - 3 = -0.70$: negative. The central 90% interval is $[0,\ 3.00 - 3]$, with an upper end -0.004 below zero: empty [28].

The ordering at one value of μ. Feldman and Cousins work one row of their construction by hand, $b = 3$ and $\mu = 0.5$ [28, Table I]. For each n the best physical signal is $\mu_{\text{best}} = \max(0, n - b)$, and $R = P(n \mid 0.5)/P(n \mid \mu_{\text{best}})$:

n	0	1	2	3	4	5	6	7	8	9	10
$P(n \mid \mu)$	0.030	0.106	0.185	0.216	0.189	0.132	0.077	0.039	0.017	0.007	0.002
μ_{best}	0	0	0	0	1	2	3	4	5	6	7
R	0.607	0.708	0.826	0.963	0.966	0.753	0.480	0.259	0.121	0.050	0.018
rank	6	5	3	2	1	4	7	8	9	10	11

Adding counts in the order of the last row, the total first reaches 0.90 with rank 7, so the acceptance set at $\mu = 0.5$ is $n \in [0, 6]$, with probability 0.935 (more than 0.90, because of

discreteness). Look at $n = 0$. Its probability under $\mu = 0.5$ is only 0.030. But no physical μ makes it much more probable: the best is $\mu = 0$, with 0.050. So $R = 0.607$, high in the ranking, and $n = 0$ is accepted. This is what keeps the interval for $n_0 = 0$ from being empty.

The intervals. Repeating the ranking on a grid of μ and reading vertically gives the Feldman–Cousins intervals; Figure 4.33 (right) shows the belt. The code is short:

Listing 4.12: The Feldman–Cousins acceptance sets for a Poisson count on background (`code/ch04/15_feldman_cousins.py`)

```
def fc_poisson_sets(mu_grid, b, cl=CL):
    """acceptance sets [n1, n2] for each mu, ranked by R = P(n|mu)/P(n|mu_best)."""
    mub = np.maximum(nn - b, 0.0)
    pbest = stats.poisson.pmf(nn, mub + b)
    n1 = np.empty(mu_grid.size, int); n2 = np.empty(mu_grid.size, int)
    for k, mu in enumerate(mu_grid):
        p = stats.poisson.pmf(nn, mu + b)
        order = np.argsort(-(p / pbest), kind="stable")
        cum = np.cumsum(p[order])
        keep = order[: np.searchsorted(cum, cl) + 1]
        n1[k], n2[k] = keep.min(), keep.max()
    return n1, n2
```

For every n the array `pbest` holds $P(n \mid \mu_{\text{best}})$; for each hypothetical μ the counts are sorted by decreasing R (`argsort` of $-R$), their probabilities accumulated, and the set cut where the running sum first reaches the confidence level. The minimum and maximum of the kept counts are the ends of the horizontal set. For $b = 3$ at 90%:

n_0	0	1	2	3	4	5	6	7	8	9	10
FC μ_1	0.00	0.00	0.00	0.00	0.00	0.00	0.15	0.90	1.51	1.88	2.64
FC μ_2	0.95*	1.88	3.04	4.42	5.59	6.99	8.46	9.53	10.99	12.29	13.50
classical	−0.70	0.89	2.32	3.68	4.99	6.27	7.53	8.77	9.99	11.21	12.41
CL_s	2.30	2.84	3.52	4.36	5.34	6.44	7.60	8.80	10.00	11.21	12.41

The first two rows agree with Feldman and Cousins' Table IV to the grid precision, with one exception marked *. For $n_0 = 0$ the raw construction gives 0.95; Feldman and Cousins publish 1.08. The difference is a deliberate correction: because n is discrete, the raw upper end is not always a decreasing function of b, and they lengthen intervals to make it so. Applying their rule (replace $\mu_2(b)$ by the largest value at any larger background) on a fine grid of b gives 1.07. For $n_0 \le 5$ the interval starts at zero: an upper limit. From $n_0 = 6$ on it is two-sided. The switch happens by itself.

CL_s and the Bayesian limit. The classical limit in the third row goes negative for small n_0. The reason: when the background fluctuates down, few events are seen, and every signal looks excluded. The LEP and Tevatron experiments adopted a modification designed against exactly this [27]. In a counting experiment, a small n is what argues against a signal, so the two relevant tail probabilities are those of seeing n_0 or fewer events, with and without the signal:

$$\text{CL}_{s+b}(\mu) = P(n \le n_0 \mid \mu + b), \qquad \text{CL}_b = P(n \le n_0 \mid b). \tag{4.79}$$

The classical recipe excludes μ when $\text{CL}_{s+b} < \alpha$.

Working definition 4.6 (CL_s)

The ratio $CL_s(\mu) = CL_{s+b}(\mu)/CL_b$ (in the notation of [27], $p_{s+b}/(1-p_b)$). A signal μ is excluded at confidence level $1-\alpha$ if $CL_s(\mu) < \alpha$. Dividing by $CL_b \le 1$ makes exclusion harder exactly when the data look background-poor [when even background alone rarely gives so few events].

For $n_0 = 0$, $CL_s = e^{-(\mu+b)}/e^{-b} = e^{-\mu}$, and the 90% limit is $\mu = \ln 10 = 2.30$ for *every* b. Seeing nothing on a background of three events excludes no more than seeing nothing on no background. That is the intended behaviour: an experiment should not get a better limit because its background fluctuated low. Figure 4.34 contrasts the three recipes as a function of b. The classical limit for $n_0 = 0$ falls as $2.30 - b$ and crosses zero at $b = 2.30$. The Feldman–Cousins limit also falls, from 2.44 at $b = 0$ towards about one: it is a correct confidence interval, and with a larger background, zero events really are rarer for every μ. CL_s stays at 2.30.

The CL_s limit is the flat-prior Bayesian limit of Cowan's (9.53)–(9.54). With a flat prior on $\mu \ge 0$, the posterior probability above μ is

$$P(\mu_s > \mu \,|\, n_0) \overset{(1)}{=} \frac{\int_\mu^\infty (\mu'+b)^{n_0} e^{-(\mu'+b)}\,\mathrm{d}\mu'}{\int_0^\infty (\mu'+b)^{n_0} e^{-(\mu'+b)}\,\mathrm{d}\mu'} \overset{(2)}{=} \frac{\int_{\mu+b}^\infty t^{n_0} e^{-t}\,\mathrm{d}t/n_0!}{\int_b^\infty t^{n_0} e^{-t}\,\mathrm{d}t/n_0!}$$
$$\overset{(3)}{=} \frac{\sum_{n=0}^{n_0} (\mu+b)^n e^{-(\mu+b)}/n!}{\sum_{n=0}^{n_0} b^n e^{-b}/n!} \overset{(4)}{=} CL_s(\mu). \tag{4.80}$$

(1) Posterior tail with the likelihood (4.78); the $1/n_0!$ cancels.
(2) Substitute $t = \mu' + b$, and divide top and bottom by $n_0!$.
(3) The gamma tail equals the Poisson cumulative sum, $\int_x^\infty t^k e^{-t}\mathrm{d}t/k! = \sum_{n=0}^k x^n e^{-x}/n!$: this is (4.61) with $z = 2t$, since $f_{\chi^2}(2t; 2k+2)\,\mathrm{d}(2t) = t^k e^{-t}\mathrm{d}t/k!$.
(4) Definition (4.79).

So setting the posterior tail to α and setting $CL_s = \alpha$ give the same limit; the script finds the two limits equal to within 4.300×10^{-12}. For $b = 0$ both reduce to the classical limit [Cowan (9.55)], because then $CL_b = 1$. So for a counting experiment without background the frequentist and the flat-prior Bayesian limits are the same number, and there is no choice to make [Cowan p. 142].

What each recipe does to the coverage. Figure 4.35 (right) shows the exact coverage for $b = 3$. All three recipes keep 90% as a minimum (Feldman–Cousins 0.900, classical 0.900, CL_s 0.901) and all have the Poisson saw-tooth. They differ in how much they overcover. The classical limit, negative values and all, is a correct 90% upper limit; its failures include every experiment that quotes a negative limit. CL_s overcovers most near $\mu = 0$, where it always covers, because it never excludes any μ below 2.30. It is conservative by design: not a confidence interval in the strict Neyman sense, but a limit with guaranteed coverage that a lucky downward fluctuation cannot improve. Feldman–Cousins is closest to exact.

Feldman and Cousins point out one more advantage of their construction [28, §IV.C]. The interval is never empty, even when the observation is improbable for *every* μ (say, no events at all on a background of fifteen). In that case a non-empty interval should not reassure anyone; the right response is a separate goodness-of-fit question, "how probable is so few events even for $\mu = 0$?". Their tables flag such entries. The interval answers "which μ are compatible with the data, given the model"; it does not test the model.

In practice the unified approach is standard for neutrino-oscillation results, CL_s for exclusion limits at the LHC, and in both communities the Bayesian flat-prior limit is quoted as a cross-check. Whatever the choice, it must be stated, and fixed before the data are looked at [Cowan p. 137].

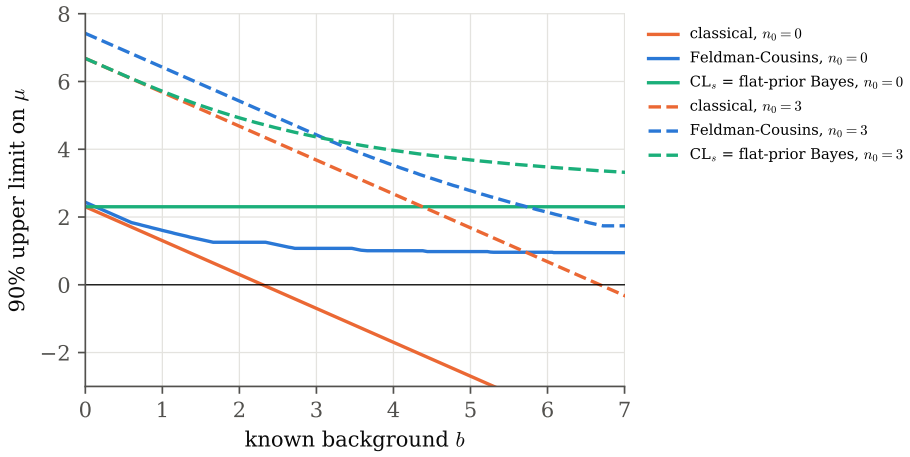


Figure 4.34: 90% upper limits on a signal mean for $n_0 = 0$ (solid) and $n_0 = 3$ (dashed) observed events, as a function of the known background. The classical limit (orange) falls linearly and becomes negative; the Feldman–Cousins limit (blue) falls but stays positive; CL$_s$, equal to the flat-prior Bayesian limit (green), does not improve when the background grows: for $n_0 = 0$ it is flat at 2.30.

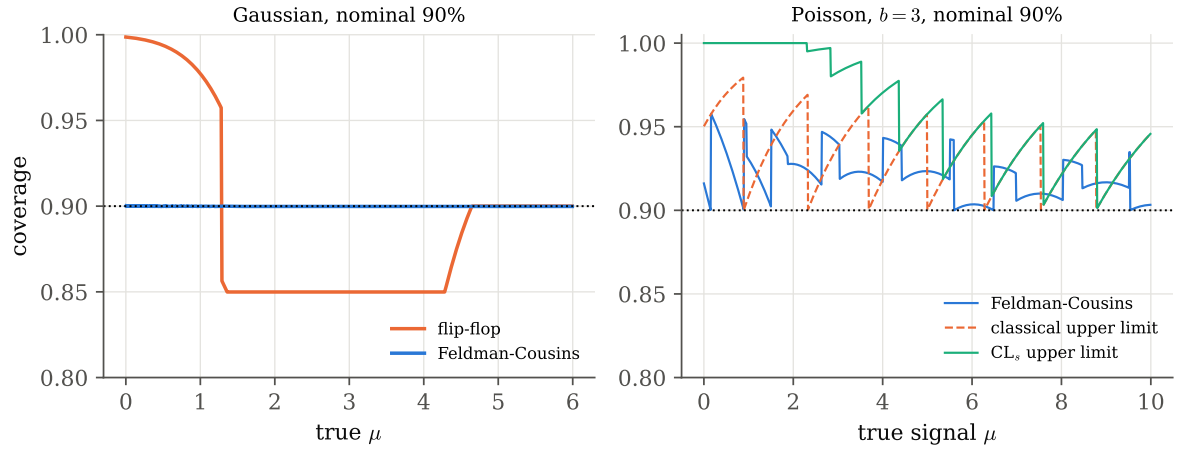


Figure 4.35: Coverage near a boundary. Left: Gaussian measurement of a non-negative mean, nominal 90%. The flip-flop recipe (orange) drops to 85% over a wide range of true values; the Feldman–Cousins intervals (blue, under the dotted line) cover 90% everywhere. Right: Poisson count on $b = 3$. Feldman–Cousins, classical upper limit and CL$_s$ all stay above 90%, with saw-tooth over-coverage; CL$_s$ overcovers most at small signals.

4.14 Many tests at once and the look-elsewhere effect

A microarray measures the activity of 2638 genes in two groups of patients, and for each gene we can test whether the two groups differ [Wasserman p. 165]. A particle search looks for a resonance at every mass between 105 and 155 GeV. A survey tests the isotropy of the sky in hundreds of patches. In all three cases each single test has a small false-alarm probability α. But there are many tests, so the chance that *some* test fires by luck is much larger. This is the **multiple testing** problem; in a particle search it is called the **look-elsewhere effect**. The first part of the chapter met it as one of the reasons for the 5σ convention. How much the false-alarm rate grows can be computed, and it can be kept under control.

The question in words.

1. If we run m tests, how likely is at least one false rejection, and how do we keep that under control?
2. When we expect some real effects among many tests, is "no false rejection at all" the right goal, or is "few false rejections among those we claim" better?
3. When the search is over a continuous parameter, like a mass, how many independent tests is it, and how do we convert the local p-value of the biggest bump into a global one?

4.14.1 The family-wise error rate and Bonferroni

Suppose we run c tests, all their null hypotheses are true, and the tests are independent, each with level α. *The question:* how likely is at least one false alarm? This probability, which Kruschke calls the *experimentwise* error rate [Kruschke §11.4.1], is

$$\alpha_{\mathrm{EW}} \overset{(1)}{=} 1 - \mathrm{P}(\text{no false alarm}) \overset{(2)}{=} 1 - (1-\alpha)^c. \tag{4.81}$$

(1) Complement.
(2) Independence: each test stays quiet with probability $1-\alpha$.

With nine groups there are $c = 9 \times 8/2 = 36$ pairwise comparisons, and at $\alpha = 0.05$ the chance of at least one false alarm is 0.84. An analyst who compares everything with everything is almost sure to find something.

The simplest cure is to make each test stricter. From here on we write m for the number of tests (the c above) and P_i for the p-value of test i. The **Bonferroni** method rejects null $H_{0,i}$ only if $P_i < \alpha/m$ [Wasserman p. 166]. Then, whatever the dependence between the tests, the probability of at least one false rejection, the **family-wise error rate**, is at most α [Wasserman Thm. 10.24]. Let R_i be the event "null i is true and falsely rejected", and R the event that at least one null is falsely rejected:

$$\mathrm{P}(R) \overset{(1)}{=} \mathrm{P}\Big(\bigcup_i R_i\Big) \overset{(2)}{\le} \sum_i \mathrm{P}(R_i) \overset{(3)}{\le} \sum_{i=1}^{m} \frac{\alpha}{m} = \alpha. \tag{4.82}$$

(1) At least one false rejection means one of the R_i happened.
(2) The union bound: the probability of a union never exceeds the sum of the probabilities (overlaps are counted more than once on the right). No independence is needed.
(3) Under a true null the p-value is uniform (or, for discrete data, valid), so $\mathrm{P}(P_i < \alpha/m) \le \alpha/m$; nulls that are false contribute no false rejections at all.

For the gene example, $0.05/2638 = 1.9 \times 10^{-5}$ is the threshold [Wasserman Ex. 10.25]. For the 36 comparisons the per-comparison level is 0.00139, and the actual family-wise rate for independent tests is $1-(1-0.00139)^{36} = 0.0488$, just below the target. The exact level for independent tests, $1-(1-\alpha)^{1/c} = 0.00142$ (the Šidák correction), is hardly different; Bonferroni's advantage is that it holds for any dependence.

The price is power. Bonferroni makes even one false rejection unlikely, and with thousands of tests the threshold becomes so small that real effects are missed. There is a second objection [Kruschke pp. 326–327]: the correction depends on how many comparisons the analyst *intended* to make, and on whether a comparison was planned or prompted by the data, so two analysts with the same data can reach different conclusions.

Watch out: The post hoc comparison

Comparing the two groups that happen to be furthest apart, after looking, is not one test: it is the largest of all possible comparisons, and its null distribution is that of a maximum. The same is true of the highest bin in a histogram, the best of several selection cuts, and the mass where a bump appeared. The p-value must be computed for the procedure actually followed, including the looking.

4.14.2 The false discovery rate and Benjamini–Hochberg

In a screening problem we *expect* some effects to be real. Then the useful promise is about the list of discoveries: most of what we claim should be true. To state this, sort the outcomes of the m tests into a table [Wasserman Table 10.2]:

	H_0 not rejected	H_0 rejected	total
H_0 true	U	V	m_0
H_0 false	T	S	m_1
total	$m - R$	R	m

V counts false discoveries, S true ones, $R = V + S$ all rejections. The **false discovery proportion** is FDP $= V/R$ (zero if $R = 0$), and the **false discovery rate** is its expectation, FDR $= \mathbb{E}(\text{FDP})$ [Wasserman p. 166]. Bonferroni controls $\mathrm{P}(V \geq 1)$; the FDR asks only that V be a small fraction of R on average.

The Benjamini–Hochberg procedure [Wasserman (10.8)]:

1. Sort the p-values, $P_{(1)} < \cdots < P_{(m)}$.
2. Find the largest i with $P_{(i)} < i\alpha/(C_m m)$, where $C_m = 1$ for independent tests and $C_m = \sum_{j=1}^{m} 1/j$ otherwise.
3. Reject every null with $P_i \leq P_{(i)}$ for that i.

Then FDR $\leq (m_0/m)\alpha \leq \alpha$, whatever the number of true nulls and whatever the distribution of the p-values of the false ones [Wasserman Thm. 10.26].

Why the sloping line $i\alpha/m$? Take the case $C_m = 1$, and follow the estimate of the false discovery proportion step by step.

- Suppose we reject every null whose p-value is below a threshold t.
- The m_0 true nulls have uniform p-values, so about $m_0 t$ of them fall below t. That is the expected number of false discoveries.
- If the threshold is the i-th smallest p-value, we make $R = i$ rejections, so the false discovery proportion is about $m_0 t/i$.
- We do not know m_0, so we replace it by its upper bound m. Requiring $mt/i \leq \alpha$ gives $t \leq i\alpha/m$: exactly the BH line.

So the procedure picks the most generous threshold whose estimated false discovery proportion is still at most α. The theorem shows that this heuristic holds exactly for the expectation.

Example: ten p-values. Wasserman's ten ordered p-values are 0.00017, 0.00448, 0.00671, 0.00907, 0.01220, 0.33626, 0.39341, 0.53882, 0.58125, 0.98617 [Wasserman Ex. 10.28]. At $\alpha = 0.05$, Bonferroni rejects below 0.005: the first 2. BH compares $P_{(i)}$ with $0.005\,i$: $0.00017 < 0.005$, $0.00448 < 0.010$, $0.00671 < 0.015$, $0.00907 < 0.020$, $0.01220 < 0.025$, but $0.33626 > 0.030$, and none of the later ones gets under the line. So BH rejects the first 5, the same as uncorrected testing here.

Example: a thousand tests. In `code/ch04/16_multiple_testing.py`, $m = 1000$ one-sided z-tests are run, of which $m_1 = 100$ have a real shift of 3 standard deviations, and the whole family is repeated 2000 times. On average:

procedure	true discoveries S	false discoveries V	FDR	$\mathrm{P}(V \geq 1)$
uncorrected, $P < 0.05$	91.2	45.1	0.330	1.000
Bonferroni	18.6	0.05	0.0024	0.045
Benjamini–Hochberg	60.5	3.0	0.0456	0.944

Each procedure buys something different. Uncorrected testing finds most real effects, but a third of its "discoveries" are false. Bonferroni makes almost no false discoveries and finds fewer than a fifth of the real effects. BH finds three times as many as Bonferroni, with a false discovery rate of 0.0456, at the bound $(m_0/m)\alpha = 0.045$ of the theorem. It almost always makes *some* false discovery, which it never promised not to do. When all nulls are true, $V = R$ and $\mathrm{FDR} = \mathrm{P}(V \geq 1)$, so BH controls the family-wise rate too: in 5000 all-null families it fired in 0.0462 of them, Bonferroni in 0.0456. Figure 4.36 shows the procedure on one family.

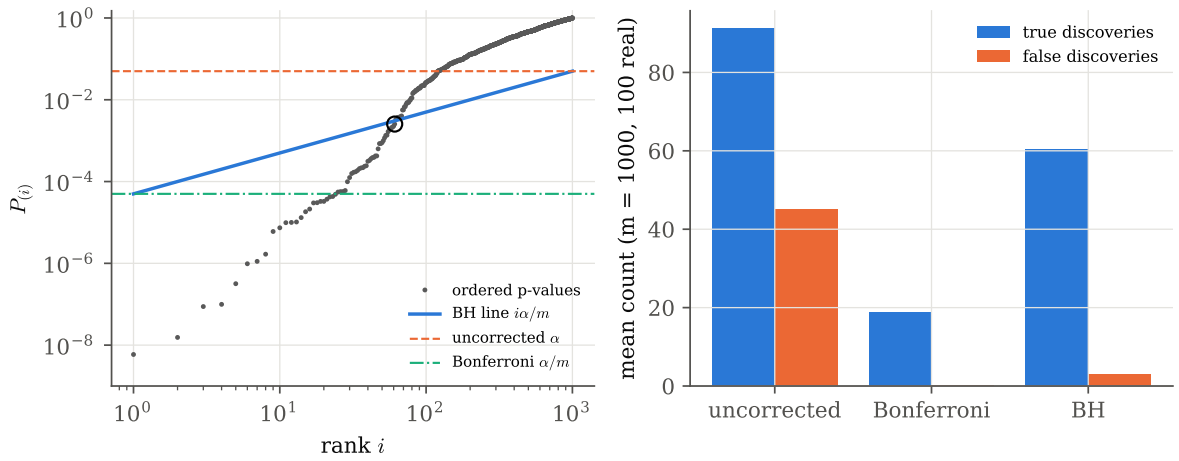


Figure 4.36: Left: the thousand p-values of one simulated family, sorted, on logarithmic axes. The real effects make the steep tail at small rank. The BH threshold is the last p-value below the sloping line $i\alpha/m$ (circled); the horizontal lines are the uncorrected and Bonferroni thresholds. Right: mean numbers of true and false discoveries of the three procedures over 2000 families.

Key idea: Two error rates for many tests

The family-wise error rate $\mathrm{P}(V \geq 1)$ is controlled by Bonferroni, $P_i < \alpha/m$, for any dependence. The false discovery rate $\mathbb{E}(V/R)$ is controlled by Benjamini–Hochberg, which rejects below the last crossing of the line $i\alpha/m$. Use the first when a single false claim is costly (a discovery), the second when a list of candidates will be followed up (a screening).

4.14.3 The look-elsewhere effect in a mass scan

A resonance search is a multiple test with a continuum of tests: one for every possible mass. (In this section m is a mass, no longer a number of tests.) *What is done.* We do not know the mass, so for every hypothesised mass m we fit a signal of that mass and compute the likelihood-ratio statistic $q_0(m)$ of the first part of the chapter. Its square root $Z(m) = \sqrt{q_0(m)}$ is the **local significance**, and $p_{\text{loc}} = \Phi(-Z)$ the **local p-value**. We then report the largest excess, $q_{\max} = \max_m q_0(m)$.

Why the usual theory fails. The mass exists only under the alternative: with no signal, the background does not depend on m, so the null hypothesis has no mass to fit. That is exactly the situation where Wilks' theorem does not apply to the maximum [15]. *The question* is then: under the background-only hypothesis, how is the maximum over the scan distributed?

Working definition 4.7 (local and global p-value, trials factor)

The **local p-value** of the biggest bump is the probability, under the background only, of an excess at least as large *at its mass*. The **global p-value** is the probability of an excess at least as large *anywhere in the search range*, $p_{\text{glob}} = \mathrm{P}(q_{\max} \geq q_{\text{obs}} \,|\, H_0)$. The **trials factor** is $p_{\text{glob}}/p_{\text{loc}}$. *Operational recipe:* simulate background-only data sets, scan each one exactly as the data were scanned, record $q_{\max}$; p_{glob} is the fraction with $q_{\max} \geq q_{\text{obs}}$.

A simulated search. The script `code/ch04/17_bump_scan.py` generates a diphoton-like spectrum of 3000 background events in sixty 1 GeV bins from 100 to 160 GeV, falling exponentially with decay length 30 GeV, and searches for a Gaussian peak of width $\sigma_m = 1.5$ GeV at 201 masses between 105 and 155 GeV in steps of 0.25 GeV. For simplicity the background shape and normalisation are taken as known; in the Gross–Vitells toy the normalisation is fitted, which changes the numbers, not the method. At each mass the signal strength $\hat{\mu} \geq 0$ maximises the binned Poisson likelihood, and

$$q_0(m) = 2\Big[\sum_i n_i \ln \frac{b_i + \hat{\mu} s_i(m)}{b_i} - \hat{\mu}\Big], \qquad \hat{\mu} \geq 0, \tag{4.83}$$

where $s_i(m)$ is the fraction of a unit signal at mass m falling in bin i, so that $\hat{\mu}$ counts signal events and $\sum_i s_i = 1$.

Listing 4.13: The scan: one-sided likelihood ratio at every mass, for many spectra at once (`code/ch04/17_bump_scan.py`)

```
def q0_scan(n):
    """n: (T, bins) counts.  Returns q0 (T, M) and mu_hat (T, M), one-sided (mu >= 0)."""
    n = n[:, None, :].astype(float)                    # (T, 1, bins)
    s = S[None, :, :]                                  # (1, M, bins)
    bb = b[None, None, :]
    g0 = (n * s / bb).sum(-1) - 1.0                    # d lnL / d mu at mu = 0 (sum s = 1)
    mu = np.maximum((s * (n - bb) / bb).sum(-1) / (s * s / bb).sum(-1), 0.0)   # linear
    start
    for _ in range(12):                                # Newton on a concave function
        lam = bb + mu[..., None] * s
        g = (n * s / lam).sum(-1) - 1.0
        h = -(n * s * s / lam**2).sum(-1)
        mu = np.maximum(mu - g / h, 0.0)
    mu = np.where(g0 > 0, mu, 0.0)
    lam = bb + mu[..., None] * s
    lnr = (n * np.log(lam / bb)).sum(-1) - mu          # lnL(mu) - lnL(0)
    return np.maximum(2 * lnr, 0.0), mu
```

The function takes a batch of spectra and returns q_0 at every mass for each of them. `g0` is the slope of $\ln \mathcal{L}$ at $\mu = 0$. If it is negative, the likelihood falls as soon as any signal is added, so the best non-negative signal is zero and $q_0 = 0$. Otherwise Newton's method climbs the log-likelihood, which is concave in μ, starting from the linear estimate $\sum s_i(n_i - b_i)/b_i \,/\, \sum s_i^2/b_i$; twelve iterations are plenty. Arrays of shape (spectra, masses, bins) let numpy do all 500×201 fits of a batch at once.

We took one background-only spectrum as "the data", choosing, among a hundred generated ones, the spectrum whose largest local excess was closest to 3σ. Figure 4.37 shows it. The best bump is at 142.50 GeV with $\hat{\mu} = 42$ events, $q_0 = 10.79$, local significance $Z_{\text{loc}} = 3.28$ and local p-value $p_{\text{loc}} = 5.10 \times 10^{-4}$. There is, of course, no particle.

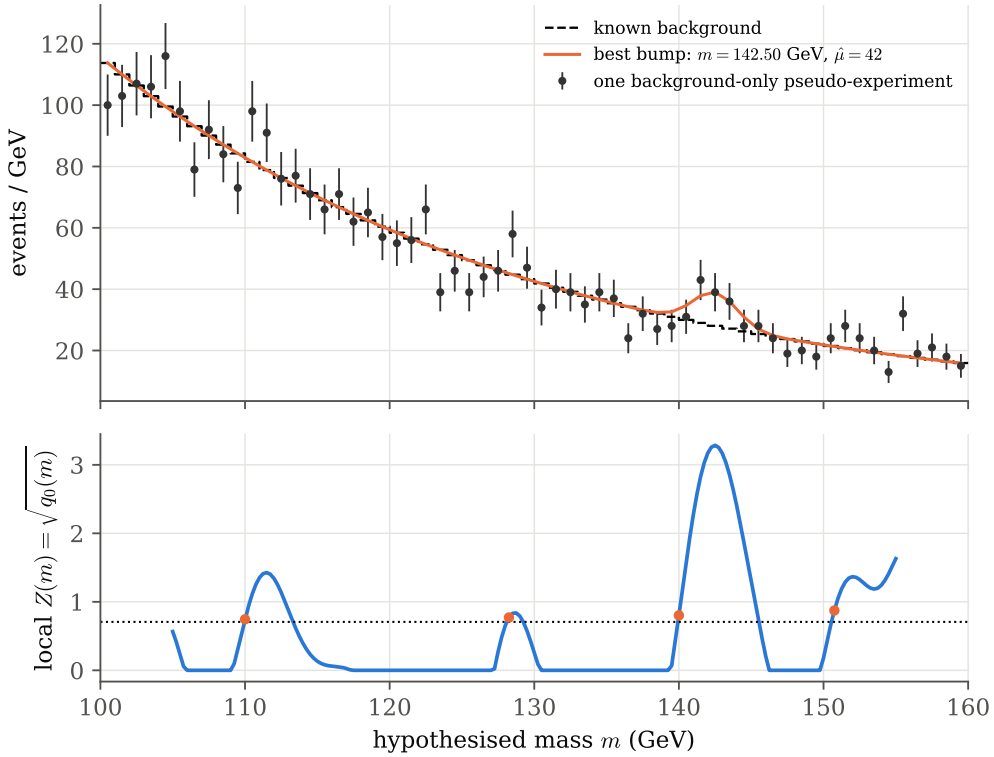


Figure 4.37: Top: a background-only pseudo-experiment (points), the known background (dashed) and the best fitted bump (orange). Bottom: the local significance $\sqrt{q_0}$ as a function of the hypothesised mass. It is zero wherever the data dip below the background, and the fluctuations are about as wide as the mass resolution. The dotted line is the reference level $\sqrt{0.5}$ used to count upcrossings (orange dots).

At one fixed mass the results of the first part of the chapter still hold. At 125 GeV, over 20000 background-only spectra, $q_0 = 0$ in 0.508 of them (the half-χ^2 spike at zero) and $\mathrm{P}(q_0 > 4) = 0.0234$, against $\Phi(-2) = 0.0228$. *Over all masses* the picture changes: a fraction 0.0159 ± 0.0009 of the background-only spectra have $q_{\max} \geq 10.79$. So

$$p_{\text{loc}} = 5.10 \times 10^{-4}\ (Z_{\text{loc}} = 3.28), \qquad p_{\text{glob}} = 0.0159\ (Z_{\text{glob}} = 2.15), \qquad \text{trials factor} \approx 31.3.$$

A "3.28σ bump" is a 2.15σ effect once we remember where we looked. A local 3σ excess somewhere in this range appears in 0.040 of background-only spectra, one in twenty-five.

4.14.4 The global p-value without a million toys

The simulation needed 20000 scans to measure a global p-value of 0.0159. For a 5σ claim, $p \approx 3 \times 10^{-7}$, it would need of order 10^7 scans, each a full fit [15]. Gross and Vitells, building on a bound by Davies, found a way around this.

The event. The largest excess is above some level c: $q_{\max} > c$. *How can it happen?* There are only two ways. Either q_0 is already above c at the lower end of the range, m_{lo}, or $q_0(m)$ crosses the level c upwards at least once somewhere in the scan. Call $N(c)$ the number of such **upcrossings**. Then

$$\mathrm{P}(q_{\max} > c) \overset{(1)}{\leq} \mathrm{P}\big(q_0(m_{\text{lo}}) > c\big) + \mathrm{P}\big(N(c) \geq 1\big) \overset{(2)}{\leq} p_{\text{loc}}(c) + \mathbb{E}[N(c)], \tag{4.84}$$

(1) The union bound over the two ways of exceeding c.
(2) $\mathrm{P}(N \geq 1) = \sum_{k \geq 1} \mathrm{P}(N = k) \leq \sum_k k\, \mathrm{P}(N = k) = \mathbb{E}\, N$; the first term is the local p-value at a fixed mass.

and for large c the bound becomes an equality, since two upcrossings of a high level are rare [15, (1)]. *The remaining question:* how does $\mathbb{E}[N(c)]$ depend on c? Treat $Z(m)$ as a smooth Gaussian process of unit variance, a random curve along the mass axis. The rate at which it crosses a level u upwards is proportional to the density of Z at u, that is to $e^{-u^2/2}$ (Rice's formula). The proportionality constant depends on how quickly $Z(m)$ forgets its value as m changes, here on the mass resolution, but not on u. With $q_0 = Z^2$ and $u = \sqrt{c}$, this gives $\mathbb{E}[N(c)] \propto e^{-c/2}$. So a count of upcrossings at one low reference level c_0 fixes the whole curve [15, (3)]:

$$p_{\text{glob}}(c) \;\approx\; p_{\text{loc}}(c) + \mathbb{E}[N(c_0)]\, e^{-(c-c_0)/2}. \tag{4.85}$$

At a low level upcrossings are frequent, so a handful of background-only scans measures $\mathbb{E}[N(c_0)]$ well. With $c_0 = 0.5$ and only the first 100 toys we count $\mathbb{E}[N(c_0)] = 2.78$ (all 20000 give 2.804), and (4.85) at $c = 10.79$ gives $p_{\text{glob}} \approx 0.0167$, against 0.0159 ± 0.0009 from the full simulation. Figure 4.38 (left) compares the whole curves.

The trials factor grows with the significance. Divide (4.85) by $p_{\text{loc}} = \Phi(-Z)$, with $c = Z^2$, and use the large-Z tail estimate $\Phi(-Z) \approx e^{-Z^2/2}/(\sqrt{2\pi}\, Z)$ from the first part of the chapter. Writing $\mathcal{N}_1 = \mathbb{E}[N(c_0)]e^{c_0/2}$, so that $\mathbb{E}[N(c)] = \mathcal{N}_1 e^{-c/2}$,

$$\frac{p_{\text{glob}}}{p_{\text{loc}}} \overset{(1)}{=} 1 + \frac{\mathcal{N}_1 e^{-Z^2/2}}{\Phi(-Z)} \overset{(2)}{\approx} 1 + \mathcal{N}_1 e^{-Z^2/2} \cdot \sqrt{2\pi}\, Z\, e^{Z^2/2} \overset{(3)}{=} 1 + \sqrt{2\pi}\, \mathcal{N}_1 Z. \tag{4.86}$$

(1) Divide (4.85) by p_{loc} and insert $c = Z^2$.
(2) The Gaussian tail estimate.
(3) The exponentials cancel.

The trials factor is not a fixed "number of independent masses". It grows linearly with the local significance [15, (12)]. (Gross and Vitells write $1 + \sqrt{\pi/2}\, \mathcal{N} Z$ for the two-sided statistic, whose level crossings are twice as many, $\mathcal{N} = 2\mathcal{N}_1$: the same formula.) Here $\mathcal{N}_1 = 3.60$, so a local 3σ bump carries a trials factor of about 28 and our 3.28σ bump about 31.3, matching the toys (Figure 4.38, right). *Why does it grow?* A higher level is crossed only near the top of a fluctuation, and the part of the fluctuation above the level gets narrower as the level rises. A pre-chosen mass must land on that narrow top, so the local p-value falls faster than the chance that the top appears somewhere in the range. For a local 5σ excess in this search, (4.85) gives $p_{\text{glob}} = 1.36 \times 10^{-5}$, a global significance of 4.20σ, from $p_{\text{loc}} = 2.87 \times 10^{-7}$: no simulation of 10^7 scans needed.

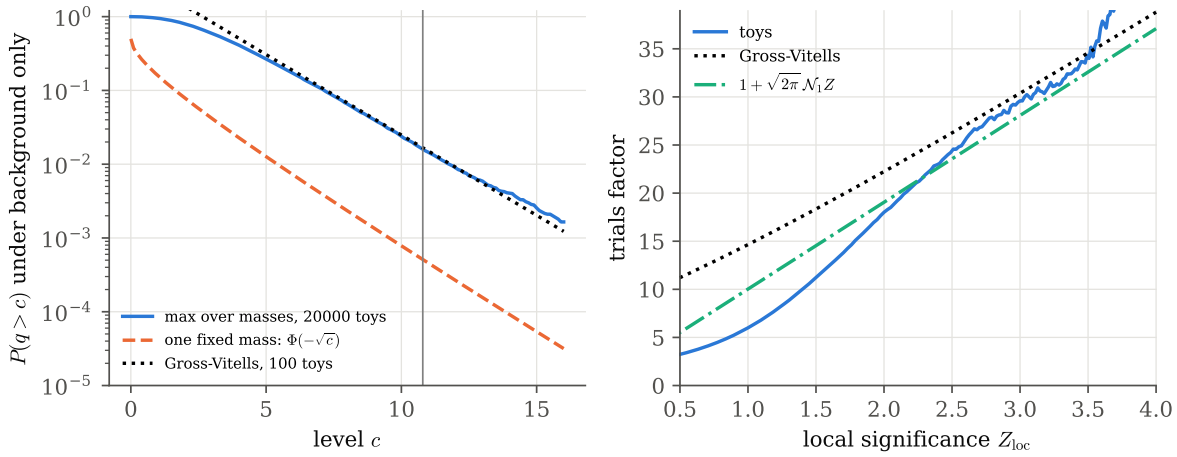


Figure 4.38: Left: the probability, with background only, that the largest excess exceeds a level c anywhere in the scan (blue, 20 000 toys), compared with the local probability at one fixed mass (orange dashed) and the Gross–Vitells formula calibrated on 100 toys (black dotted). The vertical line marks the observed bump. Right: the trials factor, global over local p-value, as a function of the local significance, with the Gross–Vitells prediction and its asymptotic straight line.

What "elsewhere" means. The global p-value refers to the range of the scan, and that range must be fixed before the data are seen. But other decay channels, other signal widths and other analyses in the same experiment are also places one could have looked. There is no unique answer to how far "elsewhere" extends. That is why experiments quote the local significance together with the global one for a stated range.

How this is used in real research: How searches report the look-elsewhere effect

A search at a collider quotes the local p-value of the largest excess, and a global p-value over a stated mass range, obtained from background-only toys where affordable and from the upcrossing formula (4.85) otherwise. The test statistic is the profile likelihood ratio of [27] with the nuisance parameters (background shape, energy scale) fitted in every toy. The same formula applies to any scan over a parameter that exists only under the signal hypothesis: the position of a localised feature in the CMB power spectrum, the frequency of an oscillation in a Galactic-centre spectrum, the sky position and frequency of a gravitational-wave template.

The Bayesian language of Chapter 5 handles the same problem differently. The unknown mass gets a prior spread over the search range. The evidence for a signal is then automatically diluted by the fraction of the range that is compatible with the bump. This dilution, the *Occam factor*, plays the role of the trials factor.

Two things were held simple in this scan: the background shape was known, and the only question was discovery. A real search fits the background together with its systematic uncertainties, and when nothing is found it must say which signal strengths are excluded. Section 4.15 builds the likelihood with nuisance parameters that the scan's $q_0(m)$ then uses, and Section 4.16 turns the same profile likelihood ratio into upper limits, CL_s and the expected band.

4.15 A search as a likelihood: templates and nuisance parameters

In the summer of 2012 two experiments at the LHC showed the same kind of plot: a histogram of the invariant mass of photon pairs, a smooth background falling from left to right, and a small bump near 125 GeV. The question put to that histogram was simple to say. Is there a signal in it, and how big is it? The size is quoted as a **signal strength** μ [the number of signal events seen, divided by the number the Standard Model predicts], so $\mu = 0$ means "background only" and $\mu = 1$ means "the predicted Higgs boson".

But the number of signal events the Standard Model "predicts" in that histogram is not known exactly. It depends on how many collisions were recorded (the luminosity), on how often a real photon pair passes the selection (the efficiency), and on where exactly a 125 GeV peak lands on the measured mass axis (the photon energy scale). Each of these was measured in a separate calibration, each to a few per cent. The background level under the peak is not predicted at all; it has to be read off the data. So the honest statement of the experiment's knowledge is not one number for the prediction but a prediction that depends on several uncertain inputs, each of which has been measured somewhere. Writing all of that as one likelihood is what turns a histogram into a measurement of μ, and the same construction serves any search: a dark-matter recoil spectrum, a line in an X-ray spectrum, an excess of gravitational-wave triggers above the noise.

The question in words.

1. What is the probability of the whole histogram, given a signal strength and everything else the prediction depends on?
2. What is a "systematic uncertainty" in this language, and why does it become a parameter with a measurement of its own?
3. How do we remove those extra parameters to speak about μ alone, and how does this compare with the profile likelihood of Chapter 3 and with the Bayesian marginalisation of Chapter 5?

4.15.1 The probability of a histogram

Start with what is recorded. Each selected event has a measured mass x, and the events are sorted into bins, $n_1, \dots, n_N$ events in bins $1, \dots, N$. What is the probability of this list of counts?

The inputs. Two physical assumptions carry everything, the same ones that gave the Poisson distribution in Chapter 2: the collisions that produce selected events happen independently of each other, one at a time, and at a rate that does not change during the run. Then the total number of selected events N_{tot} is Poisson with some mean ν_{tot}. One more input, also physical: each event lands in bin i with a probability p_i that does not depend on the other events.

The question: the bins split one Poisson stream into N pieces; what is the joint probability of the counts in the pieces? Given the total, the counts are multinomial (each of the N_{tot} events chooses its bin independently). So, with $N_{\text{tot}} = \sum_i n_i$ and $\nu_i = \nu_{\text{tot}} p_i$,

$$\begin{aligned} \mathrm{P}(n_1, \dots, n_N) &\overset{(1)}{=} \mathrm{P}(N_{\text{tot}})\ \mathrm{P}(n_1, \dots, n_N \mid N_{\text{tot}}) \overset{(2)}{=} \frac{\nu_{\text{tot}}^{N_{\text{tot}}} e^{-\nu_{\text{tot}}}}{N_{\text{tot}}!}\ \frac{N_{\text{tot}}!}{n_1! \cdots n_N!}\ p_1^{n_1} \cdots p_N^{n_N} \\ &\overset{(3)}{=} \prod_{i=1}^{N} \frac{(\nu_{\text{tot}} p_i)^{n_i}}{n_i!}\, e^{-\nu_{\text{tot}} p_i} \overset{(4)}{=} \prod_{i=1}^{N} \frac{\nu_i^{n_i}}{n_i!}\, e^{-\nu_i}. \end{aligned} \tag{4.87}$$

(1) The product rule: first the total, then how it is shared among the bins.
(2) Poisson for the total; multinomial for the sharing.
(3) The $N_{\text{tot}}!$ cancel; write $\nu_{\text{tot}}^{N_{\text{tot}}} = \prod_i \nu_{\text{tot}}^{n_i}$ and $e^{-\nu_{\text{tot}}} = \prod_i e^{-\nu_{\text{tot}} p_i}$, using $\sum_i n_i = N_{\text{tot}}$ and $\sum_i p_i = 1$.
(4) Name the mean of bin i, $\nu_i = \nu_{\text{tot}} p_i$.

The joint probability factorises: the bin counts are *independent* Poisson variables, each with its own mean. Nothing about the bins being "far apart" was needed; the independence comes from the Poisson total.

Signal plus background. In each bin the events come from two independent sources, signal and background. The sum of two independent Poisson counts is Poisson with the summed mean (Chapter 2), so the mean of bin i is

$$\nu_i = \mu\, s_i + b_i, \qquad s_i = S\, f_i, \quad b_i = B\, g_i. \tag{4.88}$$

Here S and B are the expected total numbers of signal (for $\mu = 1$) and background events, and f_i, g_i are the fractions of each falling in bin i [27, (2)–(4)].

Working definition 4.8 (template)

A **template** is the predicted histogram of one process: the fractions f_i (signal) or g_i (background) of its events expected in each bin [a shape, normalised to one, usually made by running the detector simulation on many simulated events]. The analysis fits the data as a sum of templates with free or constrained normalisations. *Operationally:* histogram the simulated events of each process with the binning of the data, divide by their number.

Example: the search we will follow. Our pseudo-experiment has sixty 1 GeV bins from 100 to 160 GeV. The background template g_i is a falling exponential of decay length 30 GeV with $B = 3000$ expected events; the signal template f_i is a Gaussian peak at 125 GeV with the mass resolution $\sigma_m = 1.5$ GeV and $S = 60$ expected events. The data were generated with $\mu = 1$ by the library `code/ch04/lib_templates.py`. With every input known, the likelihood would be (4.87) with (4.88) inserted, a function of μ alone, exactly the likelihood that was scanned over masses in Section 4.14.3.

4.15.2 Systematic uncertainties become nuisance parameters

Now let the inputs be uncertain, as they are. Take the three of our example one at a time, and ask of each: *how does it change the predicted counts, and who measured it?*

- *The background normalisation* β. The background is βB instead of B. Nobody predicts β well; the data measure it, from the bins away from the peak (the **sidebands**) or from a separate **control region** chosen to contain no signal.
- *The signal yield* κ, luminosity times efficiency relative to its nominal value. The signal is $\mu\kappa S$ instead of μS. A calibration measured κ: its result was $a_\kappa = 1$ with an error of $\sigma_\kappa = 0.10$.
- *The energy scale* δ. A mis-calibration shifts every measured photon energy, so the peak sits at $125 + \delta$ and the signal template becomes $f_i(125 + \delta)$. The scale was calibrated on the well-known Z boson peak: $a_\delta = 0$ with $\sigma_\delta = 0.5$ GeV.

Each of these is a parameter the predicted counts depend on and that we do not want to report: a **nuisance parameter** (Chapter 3). The first one is different from the other two in kind. It changes the background, which the histogram itself measures. The other two change the signal,

which the histogram cannot separate from μ: a larger κ and a smaller μ give the same peak. Only the calibration can tell them apart.

What should be done with κ? There are three candidates. We could fix it at 1; that pretends the calibration was perfect, and Chapter 3 showed what this costs: intervals too short by the factor $\sqrt{1-\rho^2}$ and coverage below its promise. We could leave it free with no information; then μ and κ are completely degenerate and μ is not measured at all. Or we could use what the calibration actually found. The calibration is a measurement like any other. Its result a_κ is a random number whose distribution depends on the true κ: Gaussian, with mean κ and standard deviation σ_κ. The probability of the calibration result, read as a function of κ, is its likelihood. The calibration and the search are independent measurements, so the probability of both results is the product of their probabilities. This gives the complete likelihood

$$\mathcal{L}(\mu, \boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{\big(\mu\, s_i(\boldsymbol{\theta}) + b_i(\boldsymbol{\theta})\big)^{n_i}}{n_i!}\, e^{-(\mu s_i(\boldsymbol{\theta}) + b_i(\boldsymbol{\theta}))} \;\times\; \prod_k p\big(a_k \mid \theta_k\big), \tag{4.89}$$

with $\boldsymbol{\theta} = (\beta, \kappa, \delta)$ the nuisance parameters and one factor $p(a_k \mid \theta_k)$ for each auxiliary measurement [27, (6)], [29]. For a calibration quoted as $a_k \pm \sigma_k$,

$$p(a_k \mid \theta_k) = \frac{1}{\sqrt{2\pi}\,\sigma_k} \exp\!\left[-\frac{(a_k - \theta_k)^2}{2\sigma_k^2}\right], \tag{4.90}$$

a **constraint term**. If the auxiliary measurement is itself a count, a control region with m events expected to be $\tau\beta B$, the factor is Poisson instead; Section 4.15.4 works that case out.

Watch out: A constraint term is a likelihood, not a prior

The Gaussian factor (4.90) looks like a prior on θ_k, and in a Bayesian analysis it is combined with one. But in (4.89) it is the probability of an observed number, the calibration result a_k, given the parameter. In pseudo-experiments, therefore, a_k must be *generated* afresh together with the histogram (the calibration would have come out differently too), not held at its observed value [27].

A nuisance parameter can change the *shape* of a template, not only its size: the energy scale moves the peak. Real analyses rarely have a formula like $f_i(125 + \delta)$. They have the nominal template and two shifted ones, made by re-running the simulation with the calibration moved up and down by one standard deviation, and they interpolate between them bin by bin [29]. With θ measured in units of its calibration error, the simplest choice is

$$f_i(\theta) = f_i^0 + \theta\,\big(f_i^+ - f_i^0\big) \quad (\theta \ge 0), \qquad f_i(\theta) = f_i^0 + \theta\,\big(f_i^0 - f_i^-\big) \quad (\theta < 0), \tag{4.91}$$

so that $\theta = 0, \pm 1$ give back the three simulated templates, and the constraint term is a standard normal in θ.

Figure 4.39 shows how the pieces of (4.89) hang together: every parameter predicts some measurements, every measurement contributes one factor, and the likelihood is their product.

4.15.3 Removing the nuisance parameters: fix, profile or integrate

We have met this problem twice before, and it is worth recalling both times concretely. In Chapter 3, in the section on nuisance parameters and the profile likelihood, a counting experiment saw 20 events in a signal region and 8 in a control region of the same size. The estimated signal was $\hat{s} = 12$. Holding the background fixed at its best value gave the error $^{+4.811}_{-4.145}$, letting it adjust (profiling) gave $^{+5.47}_{-5.192}$, and in simulated repetitions only the profile interval

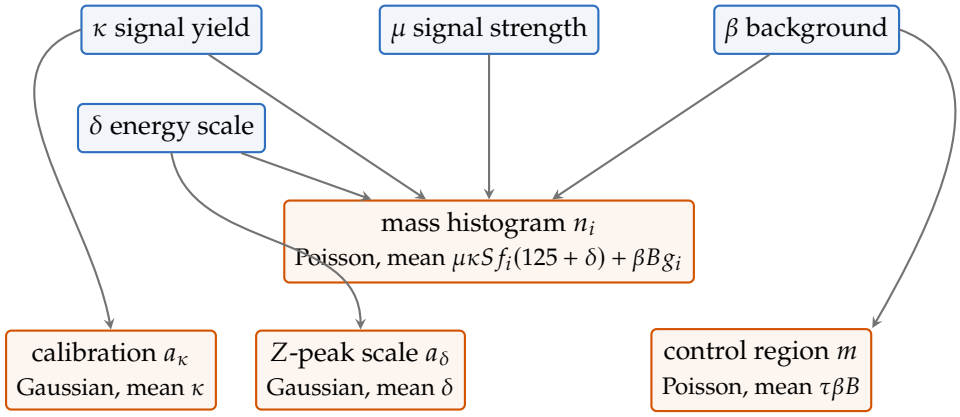


Figure 4.39: The structure of a search likelihood. Parameters (blue) predict the measurements (orange). The signal strength μ is predicted only by the main histogram, where it always appears multiplied by the signal yield κ; the calibrations measure κ and the energy scale δ on their own, and the background normalisation β is measured by the histogram's sidebands and, optionally, by a control region. The likelihood is the product of one factor per measurement.

covered the truth 68% of the time (0.6849 against 0.6088). In Chapter 5, in the section on nuisance parameters, the Bayesian removed a background level by integrating the joint posterior over it, and found that integrating and maximising agree when the likelihood is Gaussian, and differ by a "volume" factor (how many values of the nuisance parameter fit almost as well) when it is not. The search likelihood (4.89) is the same problem with three differences: there are several nuisance parameters, their information comes from separate auxiliary measurements, and the goal is not only an interval but the test statistics of Section 4.16.

The frequentist tool is the same as in Chapter 3: the **profile likelihood ratio**

$$\lambda(\mu) = \frac{\mathcal{L}\big(\mu, \hat{\hat{\boldsymbol{\theta}}}(\mu)\big)}{\mathcal{L}(\hat{\mu}, \hat{\boldsymbol{\theta}})}, \qquad 0 \le \lambda(\mu) \le 1, \tag{4.92}$$

where $\hat{\mu}, \hat{\boldsymbol{\theta}}$ maximise $\mathcal{L}$ over everything, and $\hat{\hat{\boldsymbol{\theta}}}(\mu)$ maximises it over the nuisance parameters with μ held fixed [27, (7)]. The double hat is read "the best nuisance parameters for this μ". The numerator can never exceed the denominator, because the denominator is maximised over more.

Set the three ways of removing $\boldsymbol{\theta}$ side by side.

- *Fix* $\boldsymbol{\theta}$ at $\hat{\boldsymbol{\theta}}$: the conditional likelihood. It keeps only the statistical fluctuation of the histogram. Its width is the "statistical error".
- *Profile*: at each μ, let $\boldsymbol{\theta}$ take its best value. The width of $-2\ln\lambda(\mu)$ is the total error.
- *Integrate*: give $\boldsymbol{\theta}$ a prior (here flat in β, and the calibrations' Gaussians, which are then genuinely priors) and integrate the joint posterior over it, as in Chapter 5.

What is the same? For a Gaussian likelihood all three are Gaussians in μ; profiling and integrating give the same width, the full variance $V_{\mu\mu}$, and fixing gives the smaller $1/F_{\mu\mu}$ (Chapter 3). *What differs?* The meaning: the profile curve is a likelihood ratio whose sampling distribution we will compute, the marginal is a posterior density that needs a prior. And the shape, as soon as the likelihood is not Gaussian, through the volume factor.

How large is the systematic part of the error? Analyses quote it separately, by subtraction. For one nuisance parameter, Chapter 3 found $1/F_{\mu\mu} = V_{\mu\mu}(1-\rho^2)$, with ρ the correlation of $\hat{\mu}$ and

$\hat\theta$, so

$$\sigma^2_{\text{syst}} \equiv \sigma^2_{\text{tot}} - \sigma^2_{\text{stat}} \overset{(1)}{=} V_{\mu\mu} - \frac{1}{F_{\mu\mu}} \overset{(2)}{=} V_{\mu\mu} - V_{\mu\mu}(1-\rho^2) = \rho^2\, V_{\mu\mu}. \tag{4.93}$$

(1) The profile width is $\sqrt{V_{\mu\mu}}$, the conditional width $\sqrt{1/F_{\mu\mu}}$.
(2) The identity of Chapter 3.

The "systematic error" is the part of the variance of $\hat\mu$ that it shares with the nuisance parameter through their correlation. It adds to the statistical one in quadrature by construction, not by assumption. With several nuisance parameters the same subtraction is used, and ρ^2 becomes the fraction of the variance of $\hat\mu$ that all of them together account for.

Example: the template fit. The script `code/ch04/19_hep_templates.py` maximises (4.89) for our pseudo-data. The heart of it is the negative log-likelihood:

Listing 4.14: The search likelihood: Poisson bins times two Gaussian constraint terms (`code/ch04/lib_templates.py`)

```
def expected(mu, theta, m=125.0):
    beta, kappa, delta = theta
    return mu * kappa * S0 * sig_template(m + delta) + beta * B0 * G


def nll(mu, theta, n, m=125.0, aux=(A_KAPPA, A_DELTA)):
    """-ln L(mu, theta) up to a constant: Poisson bins times Gaussian constraint terms."""
    nu = expected(mu, theta, m)
    if np.any(nu <= 0):
        return 1e12
    beta, kappa, delta = theta
    pois = np.sum(nu - n * np.log(nu))
    cons = 0.5 * ((kappa - aux[0]) / SIG_KAPPA) ** 2 + 0.5 * ((delta - aux[1]) / SIG_DELTA)
    ** 2
    return pois + cons
```

`expected` builds the mean counts (4.88) with the three nuisance parameters inserted; `nll` returns $-\ln\mathcal{L}$ without the constants $\ln n_i!$ and $\ln\sqrt{2\pi}\sigma_k$, which cancel in every ratio. A numerical minimiser (`scipy.optimize.minimize`) does the rest, over all four parameters for the denominator of (4.92), over the three nuisance parameters at each fixed μ for the numerator. The global fit gives

$$\hat\mu = 1.07, \qquad \hat\beta = 0.972, \qquad \hat\kappa = 1.00, \qquad \hat\delta = -0.14\,\text{GeV}.$$

Why is $\hat\kappa$ exactly the calibration value? Because the histogram sees only the product $\mu\kappa$. Whatever the product the histogram prefers, the fit can reach it with any κ by moving μ, so it sets κ where its own measurement, the calibration, wants it. The energy scale, in contrast, is pulled a little by the data: the peak in the histogram sits slightly below 125 GeV, and the fit moves it by a fraction of the calibration error.

Figure 4.40 (right) shows the three curves. Read at $-2\ln(\mathcal{L}/\mathcal{L}_{\max}) = 1$:

$$\text{fixed: } \hat\mu^{+0.30}_{-0.32}, \qquad \text{profile: } \hat\mu^{+0.34}_{-0.34}, \qquad \text{marginal: } 1.04^{+0.36}_{-0.32}.$$

So $\sigma_{\text{stat}} = 0.31$, $\sigma_{\text{tot}} = 0.34$, and by (4.93) $\sigma_{\text{syst}} = 0.15$. Most of the systematic part is the 10% yield calibration: since the histogram measures $\mu\kappa$, a 10% error on κ is a 10% error on μ, about $0.1\,\hat\mu \approx 0.11$, and $\sqrt{0.31^2 + 0.11^2} \approx 0.33$. Profiling and marginalising nearly agree: the marginal peaks slightly lower, at 1.04, with almost the same width, and the largest difference between the two curves inside the plotted interval is 0.27 in $-2\ln$. That is the volume effect of Chapter 5

at work, small here because the likelihood is close to Gaussian. The sizes are the lesson. A 1.5 GeV resolution and 3000 background events make the statistical error dominant; a ten times larger data set would shrink σ_{stat} by $\sqrt{10}$ and leave σ_{syst} untouched, and the calibration would then limit the measurement.

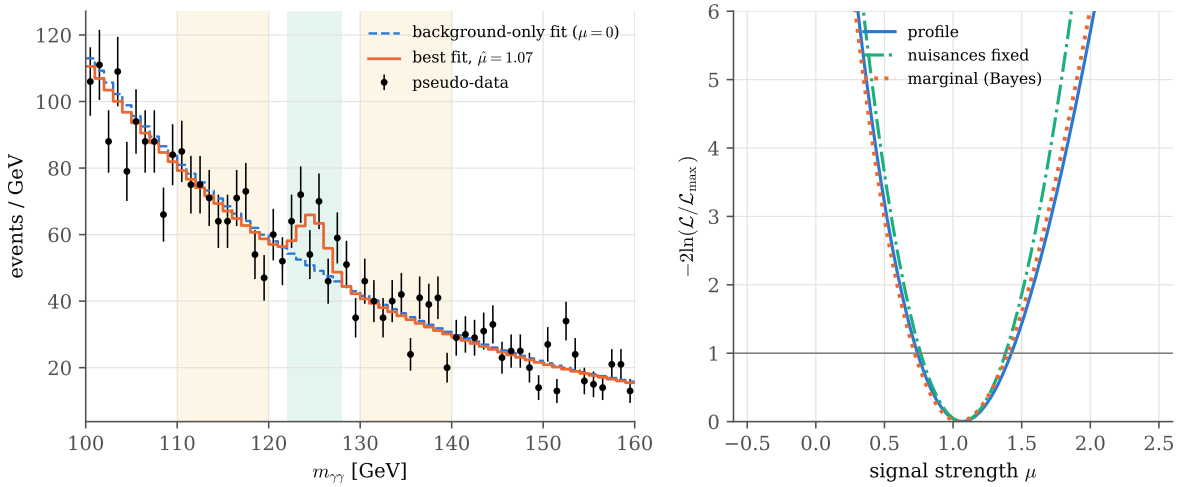


Figure 4.40: Left: the pseudo-data (points), the best background-only fit (blue dashed) and the best signal-plus-background fit (orange). The green band is the signal window and the yellow bands the sidebands used by the counting version of the search. Right: three ways of turning the four-parameter likelihood into a curve for μ. Fixing the nuisance parameters gives the narrowest curve (the statistical error); profiling them gives the total error; integrating them with priors (Bayes) gives nearly the same curve, slightly shifted towards smaller μ. The grey line at one marks the 68% interval.

4.15.4 The smallest search: a signal window and its sidebands

Every piece of the template likelihood can be seen working, and solved by hand, in a stripped-down version of the same search. Keep only two numbers from the histogram: the count n in a signal window around the peak, 122–128 GeV, and the count m in two sidebands, 110–120 and 130–140 GeV (Figure 4.40, left). In the window we expect $\mu s + b$ events, where $s = 57.3$ is the nominal signal falling in the window and b the unknown background. In the sidebands we expect τb, where $\tau = 3.531$ is the ratio of the background template integrated over the sidebands to its integral over the window. The shape is taken as known; only its normalisation b is not. The sidebands are the control region and b is the nuisance parameter. The likelihood is two Poisson factors [27, (90)],

$$\ln \mathcal{L}(\mu, b) = n \ln(\mu s + b) - (\mu s + b) + m \ln(\tau b) - \tau b + \text{const}, \tag{4.94}$$

the on/off likelihood of Chapter 3 with k renamed τ and the signal written as μs.

The global maximum. Setting the two derivatives to zero,

$$\begin{aligned} \frac{\partial \ln \mathcal{L}}{\partial \mu} &\overset{(1)}{=} s\left(\frac{n}{\mu s + b} - 1\right) = 0 \;\Rightarrow\; \hat{\mu} s + \hat{b} = n, \\ \frac{\partial \ln \mathcal{L}}{\partial b} &\overset{(2)}{=} \frac{n}{\mu s + b} - 1 + \frac{m}{b} - \tau = 0 \;\Rightarrow\; \hat{b} = \frac{m}{\tau}, \qquad \hat{\mu} = \frac{n - m/\tau}{s}. \end{aligned} \tag{4.95}$$

(1) Only the first two terms of (4.94) contain μ.

(2) By the first equation the first two terms cancel, leaving $m/b = \tau$; then $\hat{\mu} s = n - \hat{b}$.

The sidebands supply an estimate of the background, scaled down to the window by τ, and the signal is what is left over. With $n = 365$ and $m = 1047$, $\hat{b} = 296.5$ and $\hat{\mu} = 1.20$. We allow $\hat{\mu} < 0$ here: it is whatever value maximises the likelihood, as long as the mean count $\mu s + b$ stays positive. Section 4.16 needs this.

The best background for a given μ. Hold μ fixed and solve only the second equation of (4.95):

$$\begin{aligned} 0 &\overset{(1)}{=} n\,b - b(\mu s + b) + m(\mu s + b) - \tau b(\mu s + b) \\ &\overset{(2)}{=} -(1+\tau)\,b^2 + \big[n + m - (1+\tau)\mu s\big]\,b + m\mu s, \\ \hat{\hat{b}}(\mu) &\overset{(3)}{=} \frac{A}{2(1+\tau)} + \sqrt{\frac{A^2 + 4(1+\tau)\,m\,\mu s}{4(1+\tau)^2}}, \qquad A = n + m - (1+\tau)\mu s. \end{aligned} \tag{4.96}$$

(1) Multiply $\partial \ln \mathcal{L}/\partial b = 0$ by $b(\mu s + b) > 0$.
(2) Collect powers of b.
(3) The quadratic formula; for $\mu > 0$ the constant term $m\mu s$ is positive, so the roots have opposite signs and only the + root is a background.

This is [27, (93)], and the same quadratic as in Chapter 3. At $\mu = 0$ the square root equals $A/(2(1+\tau))$ and $\hat{\hat{b}}(0) = (n+m)/(1+\tau)$: with no signal, the window and the sidebands are pooled into one background measurement. Here $\hat{\hat{b}}(0) = 311.7$, more than $\hat{b}$, because the pooled estimate also uses the excess in the window.

The discovery statistic. For $\hat{\mu} > 0$, $q_0 = -2\ln\lambda(0)$. At the global maximum $\hat{\mu}s + \hat{b} = n$ and $\tau\hat{b} = m$; at $\mu = 0$ the window mean is $\hat{\hat{b}}(0)$ and the sideband mean $\tau\hat{\hat{b}}(0)$, and the two linear terms add up to $(1+\tau)\hat{\hat{b}}(0) = n + m$. So

$$\begin{aligned} q_0 &\overset{(1)}{=} 2\Big[\big(n\ln n - n + m\ln m - m\big) - \big(n\ln\hat{\hat{b}}(0) + m\ln\tau\hat{\hat{b}}(0) - (n+m)\big)\Big] \\ &\overset{(2)}{=} 2\Big[n\ln\frac{n}{\hat{\hat{b}}(0)} + m\ln\frac{m}{\tau\,\hat{\hat{b}}(0)}\Big]. \end{aligned} \tag{4.97}$$

(1) (4.94) at the two maxima, using the values just listed.
(2) The $-(n+m)$ terms cancel; combine the logarithms.

The numbers give $q_0 = 11.28$, a significance $Z = \sqrt{q_0} = 3.36$ (Section 4.16.2 shows why the square root is the significance even with a nuisance parameter). Compare with two neighbours.

- *The background known exactly*, $b = 302.1$: the formula of the first part of the chapter, $q_0 = 2[n\ln(n/b) - (n-b)]$, gives $Z = 3.50$.
- *The full template fit* of Section 4.15.3, with three nuisance parameters: $Z = \sqrt{q_0} = 3.53$.

One pseudo-experiment is a poor judge of which method is better: all three values of Z move with the luck of the same data. The expected significances, computed with the Asimov data set of Section 4.16.3, do not: 2.80 for the window with sidebands, 3.20 with the background known, and 3.32 for the template fit. Measuring the background instead of knowing it costs significance. The template fit wins it back, because it uses every bin, each weighted by its own ratio of signal to background, instead of a window with a hard edge.

The size of the cost is easiest to see in the Gaussian limit. $\hat{\mu} = (n - m/\tau)/s$ is a difference of independent counts, so its variance is about $(n + m/\tau^2)/s^2$, an error of 0.37; with the background known it would be $\sqrt{n}/s = 0.33$. The control region adds the term $m/\tau^2 \approx b/\tau$: a control region τ times larger than the window adds a fraction $1/\tau$ of the background variance.

The lesson in one line: a systematic uncertainty is a parameter with a measurement of its own, and the search likelihood is the product of the main measurement and those auxiliary measurements.

4.16 Discovery and exclusion with nuisance parameters: q_0, $\tilde{q}_\mu$, CL_s and the Brazil band

The discovery statistic q_0 has appeared twice already. In the first part of the chapter, a count N on a *known* background b gave $q_0 = 2[N\ln(N/b) - (N-b)]$ for $N > b$ and zero otherwise; with background only it is half the time zero and otherwise χ^2_1, so its tail is $\Phi(-\sqrt{q_0})$ and the significance is simply $Z = \sqrt{q_0}$. For Cowan's peak of 11 events on an expected 3.2 this gave $Z = 3.40$, against 3.29 from the exact Poisson tail. In Section 4.14.3 the same q_0 was computed at every hypothesised mass, again with the background known, and the largest one needed the look-elsewhere correction. The statistic is the same now; three things are new. The background and the calibrations are nuisance parameters, profiled in q_0, and we need to know whether $Z = \sqrt{q_0}$ survives that. We need a second statistic for *excluding* a signal, and the distribution of both statistics when the signal is present, to say what an experiment can expect. And we need a rule for exclusion that does not reward an experiment for a lucky downward fluctuation.

The question in words.

1. With nuisance parameters, which number measures how badly the data disagree with "no signal", and which with "a signal of strength μ"?
2. What are their distributions, without simulating millions of experiments?
3. Before looking at the data, what discovery significance and what limit should the experiment expect, and how far from it may the observed values reasonably fall?
4. Why do the LHC experiments exclude with CL_s rather than with the plain p-value?

4.16.1 Three statistics built from the profile likelihood ratio

All three statistics are built from the profile likelihood ratio $\lambda(\mu)$ of (4.92): the best likelihood with μ fixed, nuisance parameters at their best, divided by the best likelihood overall. A value near one says that μ fits about as well as anything; a small value says that it does not. The basic statistic is $t_\mu = -2\ln\lambda(\mu)$, large when μ fits badly. The physical question decides which fluctuations should count against μ, and that gives the three variants [27, §2].

Discovery: testing $\mu = 0$. A signal can only add events. If the data show fewer events than the background predicts ($\hat{\mu} < 0$), that may be evidence of a badly modelled background, but it is no evidence for a signal. So

$$q_0 = \begin{cases} -2\ln\lambda(0), & \hat{\mu} \geq 0, \\ 0, & \hat{\mu} < 0, \end{cases} \tag{4.98}$$

[27, (12)], the one-sided statistic of the first part of the chapter, now with nuisance parameters inside λ.

Exclusion: testing a signal $\mu > 0$. To set an upper limit we ask whether the data have too *few* events for μ. Data with $\hat{\mu} > \mu$ (more signal-like than μ itself) are not evidence against μ when the question is an upper limit, so they count as perfect agreement. And since $\mu \geq 0$ is physical,

the reference fit should not use a negative signal: if $\hat{\mu} < 0$, the best *physical* fit is $\mu = 0$. Taking these one at a time gives the statistic

$$\tilde{q}_\mu = \begin{cases} -2\ln \dfrac{\mathcal{L}\big(\mu, \hat{\hat{\boldsymbol{\theta}}}(\mu)\big)}{\mathcal{L}\big(0, \hat{\hat{\boldsymbol{\theta}}}(0)\big)}, & \hat{\mu} < 0, \\ -2\ln \dfrac{\mathcal{L}\big(\mu, \hat{\hat{\boldsymbol{\theta}}}(\mu)\big)}{\mathcal{L}\big(\hat{\mu}, \hat{\boldsymbol{\theta}}\big)}, & 0 \le \hat{\mu} \le \mu, \\ 0, & \hat{\mu} > \mu, \end{cases} \tag{4.99}$$

[27, (16)]. Replacing $\hat{\mu}$ by the best physical value when $\hat{\mu} < 0$ is the move of Feldman and Cousins, whose ordering used $\mu_{\text{best}} = \max(0, x)$ (Section 4.13.3). (The variant q_μ without the tilde keeps $\hat{\mu}$ in the denominator in all cases; in the approximation of the next subsection both give the same limits.)

Figure 4.41 draws the two statistics as functions of $\hat{\mu}$, in the approximation derived next, for a signal μ two standard deviations from zero. Both are monotonic where they are not zero: q_0 grows as $\hat{\mu}$ rises, $\tilde{q}_\mu$ grows as $\hat{\mu}$ falls. That monotonicity is what makes their distributions easy.

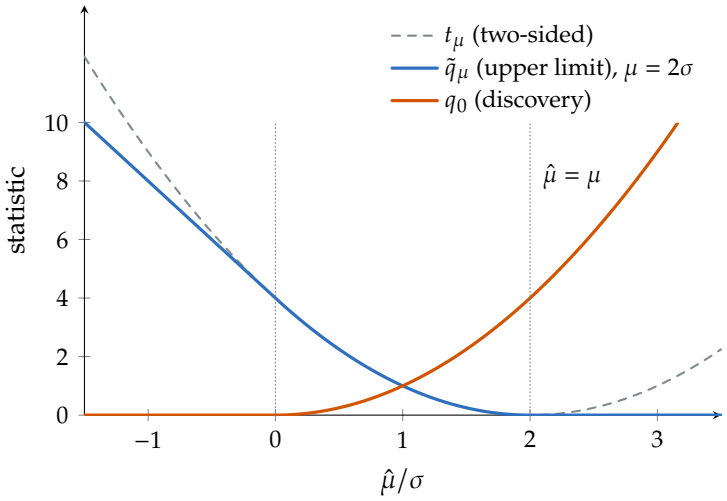


Figure 4.41: The test statistics as functions of the estimated signal strength, in units of its standard deviation, for a tested signal $\mu = 2\sigma$. The discovery statistic q_0 (orange) is zero for a deficit and grows as $(\hat{\mu}/\sigma)^2$ for an excess. The upper-limit statistic $\tilde{q}_\mu$ (blue) is zero for $\hat{\mu} > \mu$, follows the two-sided t_μ (grey dashed) between 0 and μ, and for $\hat{\mu} < 0$ continues as a straight line, because the reference fit stops at the physical boundary.

4.16.2 Wald's approximation and the half-χ^2

The question: if the true signal strength is μ', how is $\tilde{q}_\mu$ (or q_0) distributed? The answer rests on one approximation, due to Wald [27, (17)]: for a large data sample,

$$-2\ln \lambda(\mu) \approx \frac{(\mu - \hat{\mu})^2}{\sigma^2}, \qquad \hat{\mu} \sim \mathcal{N}(\mu', \sigma^2), \tag{4.100}$$

where $\sigma^2 = V_{\mu\mu}$ is the variance of $\hat{\mu}$ *including* the nuisance parameters. Both halves come from Chapter 3. Near its maximum every smooth log-likelihood is quadratic, and maximising the quadratic over the nuisance directions leaves a parabola in μ with curvature $1/V_{\mu\mu}$: that is the profile result of Chapter 3 (the section on nuisance parameters), which gives

$\ln \mathcal{L}_p(\mu) - \ln \mathcal{L}_{\max} = -(\mu - \hat{\mu})^2/2V_{\mu\mu}$, and -2 times this is (4.100). And the maximum-likelihood estimator of a large sample is Gaussian about the true value with that variance (Chapter 3). The corrections are of relative order $1/\sqrt{N_{\text{events}}}$. The nuisance parameters enter only through σ, which they make larger. That is the whole reason the formulas of the first part of the chapter survive with nuisance parameters.

The distribution of q_0. With (4.100) at $\mu = 0$, $-2\ln\lambda(0) = \hat{\mu}^2/\sigma^2$, so $q_0 = \hat{\mu}^2/\sigma^2$ for $\hat{\mu} \geq 0$ and 0 otherwise. *The question:* what is $\mathrm{P}(q_0 \leq x)$ for $x > 0$? The event "$q_0 \leq x$" happens in two ways: $\hat{\mu} < 0$ (then $q_0 = 0$), or $0 \leq \hat{\mu} \leq \sigma\sqrt{x}$. Together this is just $\hat{\mu} \leq \sigma\sqrt{x}$:

$$
\begin{aligned}
F(q_0 \,|\, \mu') &\overset{(1)}{=} \mathrm{P}\big(\hat{\mu} \leq \sigma\sqrt{q_0}\big) \overset{(2)}{=} \Phi\Big(\frac{\sigma\sqrt{q_0} - \mu'}{\sigma}\Big) = \Phi\Big(\sqrt{q_0} - \frac{\mu'}{\sigma}\Big), \\
f(q_0 \,|\, \mu') &\overset{(3)}{=} \Big[1 - \Phi\Big(\frac{\mu'}{\sigma}\Big)\Big]\delta(q_0) + \frac{1}{2\sqrt{2\pi q_0}} \exp\Big[-\frac{1}{2}\Big(\sqrt{q_0} - \frac{\mu'}{\sigma}\Big)^2\Big].
\end{aligned}
\tag{4.101}
$$

(1) The two ways combined.
(2) Standardise: $(\hat{\mu} - \mu')/\sigma$ is standard normal.
(3) The jump of F at $q_0 = 0$, $\Phi(-\mu'/\sigma)$, is a point mass; for $q_0 > 0$ differentiate with the chain rule, $\mathrm{d}\sqrt{q_0}/\mathrm{d}q_0 = 1/(2\sqrt{q_0})$, and $\Phi' = \varphi$.

For $\mu' = 0$ this is $\frac{1}{2}\delta(q_0)$ plus half a χ^2_1 density [27, (48)–(49)]; its p-value is $p_0 = 1 - F(q_0 \,|\, 0) = 1 - \Phi(\sqrt{q_0})$, and the significance $Z_0 = \Phi^{-1}(1 - p_0) = \sqrt{q_0}$ [27, (53)]. The result of the first part of the chapter, now with any number of profiled nuisance parameters.

The distribution of $\tilde{q}_\mu$. Insert (4.100) into the three cases of (4.99). The middle case is $(\mu - \hat{\mu})^2/\sigma^2$ directly. In the first, the reference is the fit at $\mu = 0$ instead of the global one:

$$
\begin{aligned}
-2\ln\frac{\mathcal{L}(\mu, \hat{\hat{\theta}}(\mu))}{\mathcal{L}(0, \hat{\hat{\theta}}(0))} &\overset{(1)}{=} -2\ln\frac{\mathcal{L}(\mu, \hat{\hat{\theta}}(\mu))}{\mathcal{L}(\hat{\mu}, \hat{\theta})} + 2\ln\frac{\mathcal{L}(0, \hat{\hat{\theta}}(0))}{\mathcal{L}(\hat{\mu}, \hat{\theta})} \\
&\overset{(2)}{=} \frac{(\mu - \hat{\mu})^2}{\sigma^2} - \frac{\hat{\mu}^2}{\sigma^2} \overset{(3)}{=} \frac{\mu^2 - 2\mu\hat{\mu}}{\sigma^2}.
\end{aligned}
\tag{4.102}
$$

(1) Multiply and divide by the global maximum: $-2\ln\lambda(\mu) + 2\ln\lambda(0)$.
(2) (4.100) at μ and at 0.
(3) Expand the square; the $\hat{\mu}^2$ cancel.

This is the straight line of Figure 4.41, matching the parabola at $\hat{\mu} = 0$ where both equal μ^2/σ^2 [27, (62)].

The question: if the truth is μ', how probable is a value of $\tilde{q}_\mu$ at least as large as the observed $\tilde{q}_{\text{obs}}$? Since $\tilde{q}_\mu$ falls as $\hat{\mu}$ rises (where it is not zero), "$\tilde{q}_\mu \geq \tilde{q}_{\text{obs}}$" is the same event as "$\hat{\mu} \leq \hat{\mu}_{\text{obs}}$", where $\hat{\mu}_{\text{obs}}$ is the estimate that produces $\tilde{q}_{\text{obs}}$. So

$$
\mathrm{P}\big(\tilde{q}_\mu \geq \tilde{q}_{\text{obs}} \,|\, \mu'\big) = \mathrm{P}\big(\hat{\mu} \leq \hat{\mu}_{\text{obs}} \,|\, \mu'\big) = \Phi\Big(\frac{\hat{\mu}_{\text{obs}} - \mu'}{\sigma}\Big).
\tag{4.103}
$$

To express this through the statistic itself, invert each branch. For $\tilde{q}_{\text{obs}} \leq \mu^2/\sigma^2$ (the parabola), $\hat{\mu}_{\text{obs}} = \mu - \sigma\sqrt{\tilde{q}_{\text{obs}}}$; for $\tilde{q}_{\text{obs}} > \mu^2/\sigma^2$ (the line), $\hat{\mu}_{\text{obs}} = (\mu^2 - \sigma^2\tilde{q}_{\text{obs}})/(2\mu)$. Inserting each into (4.103),

$$
1 - F(\tilde{q}_\mu \,|\, \mu') \overset{(1)}{=} \begin{cases} \Phi\big(-\sqrt{\tilde{q}_\mu} + (\mu - \mu')/\sigma\big), & \tilde{q}_\mu \leq \mu^2/\sigma^2, \\ \Phi\Big(-\dfrac{\tilde{q}_\mu - (\mu^2 - 2\mu\mu')/\sigma^2}{2\mu/\sigma}\Big), & \tilde{q}_\mu > \mu^2/\sigma^2, \end{cases}
\tag{4.104}
$$

(1) First branch: $(\mu - \sigma\sqrt{\tilde{q}} - \mu')/\sigma$. Second branch: $(\mu^2 - \sigma^2\tilde{q} - 2\mu\mu')/(2\mu\sigma)$, then divide top and bottom by σ.

which is [27, (65)]. For $\mu' = \mu$, the p-value of the tested signal is $p_\mu = 1 - \Phi(\sqrt{\tilde{q}_\mu})$ on the parabola: the half-χ^2 again.

The upper limit. The 95% upper limit is the μ whose p-value is $\alpha = 0.05$. By (4.103) with $\mu' = \mu$, $p_\mu = \Phi((\hat{\mu} - \mu)/\sigma)$ on *both* branches, and setting it equal to α gives

$$\mu_{\text{up}} = \hat{\mu} + \sigma\,\Phi^{-1}(1 - \alpha) \tag{4.105}$$

[27, (69)]: the classical one-sided limit (4.74) of Section 4.13.1, with σ now including the systematics. It inherits that limit's flaw. When $\hat{\mu} < -1.645\,\sigma$ the limit is negative, and it excludes every physical signal including none at all.

Toys against the formulas. How large must the sample be? The formulas are tested by the script `code/ch04/20_cls_limits.py` on the smallest case available, the window-and-sideband experiment of Section 4.15.4 with Cowan et al.'s numbers $s = 10$, $b = 10$, $\tau = 1$ [27, §5.1]: ten signal events on ten background events, the background measured in a control region of the same size. The statistics follow from (4.96)–(4.97) with no numerical fitting:

Listing 4.15: Profiled background, q_0 and $\tilde{q}_\mu$ for the counting experiment with a control region, for many toys at once (`code/ch04/20_cls_limits.py`)

```
    def bhh(self, mu, n, m):          # profiled background, Cowan et al. (93)
        s, t = self.s, self.tau
        a = n + m - (1 + t) * mu * s
        return np.maximum(a / (2 * (1 + t)) + np.sqrt(a**2 + 4 * (1 + t) * m * mu * s) / (2
    * (1 + t)), 1e-12)

    def lnL(self, mu, b, n, m):
        nu = np.maximum(mu * self.s + b, 1e-300)
        return n * np.log(nu) - nu + m * np.log(self.tau * b) - self.tau * b

    def mle(self, n, m):
        return (n - m / self.tau) / self.s, m / self.tau

    def lnL_hat(self, n, m):          # at the MLE: mu s + b = n, tau b = m (0 ln 0 = 0)
        with np.errstate(divide="ignore", invalid="ignore"):
            return np.where(n > 0, n * np.log(np.maximum(n, 1e-300)), 0) - n + \
                   np.where(m > 0, m * np.log(np.maximum(m, 1e-300)), 0) - m

    def q0(self, n, m):
        mu_h, _ = self.mle(n, m)
        q = 2 * (self.lnL_hat(n, m) - self.lnL(0.0, self.bhh(0.0, n, m), n, m))
        return np.where(mu_h > 0, np.maximum(q, 0), 0.0)

    def qt(self, mu, n, m):           # q~_mu, Cowan et al. (16)
        mu_h, _ = self.mle(n, m)
        l_mu = self.lnL(mu, self.bhh(mu, n, m), n, m)
        l_ref = np.where(mu_h >= 0, self.lnL_hat(n, m), self.lnL(0.0, self.bhh(0.0, n, m),
    n, m))
        return np.where(mu_h > mu, 0.0, np.maximum(2 * (l_ref - l_mu), 0.0))
```

`bhh` is (4.96); `lnL_hat` is the log-likelihood at the global maximum, where the means equal the counts; `q0` and `qt` are (4.98) and (4.99), case by case, with `np.where` choosing the branch for every toy. With 400 000 toys for each hypothesis (Figure 4.42): with background only, $q_0 = 0$ in a fraction 0.546 of the toys (above one half because the counts are integers), and $\mathrm{P}(q_0 \geq 9) = 0.001572$, against $\Phi(-3) = 1.35 \times 10^{-3}$; with the signal present, the toys follow the

shifted density of (4.101), with σ taken from the Asimov data set of the next subsection. Even with twenty events, the formulas describe the bulk of the distributions; the far tails, and the discreteness, are where toys are still needed.

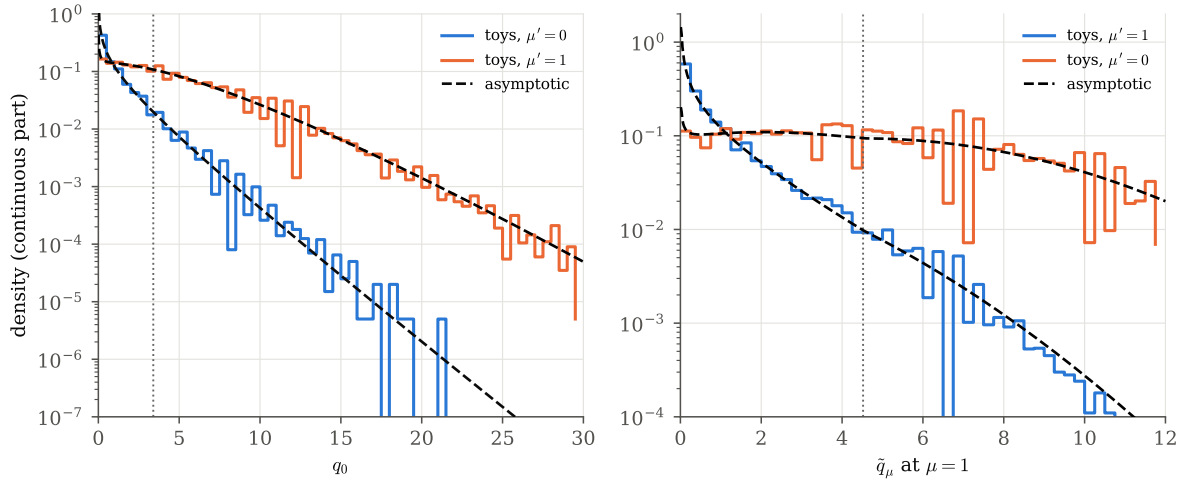


Figure 4.42: Distributions of the test statistics for a counting experiment with $s = 10$, $b = 10$ and a control region of the same size, from toys (steps) and from the asymptotic formulas (dashed). Left: q_0 with background only (blue, the half-χ^2) and with the signal present (orange). Right: $\tilde{q}_\mu$ for $\mu = 1$, when the signal is present (blue, again a half-χ^2) and when it is absent (orange). The dotted lines are the Asimov values, which estimate the medians of the orange distributions. The jagged steps come from the integer counts.

4.16.3 The Asimov data set: what an experiment should expect

Before an experiment runs, its designers want to know its *sensitivity*: if the signal is there, what significance will it typically reach, and if not, what limit will it typically set? "Typically" means the median over repetitions. Simulating thousands of experiments answers this; one artificial data set answers it too.

Working definition 4.9 (Asimov data set)

The **Asimov data set** for a strength μ' is the data set in which every count equals its expectation, $n_i = \mu' s_i + b_i$, and every auxiliary measurement equals the nominal value of its nuisance parameter [no statistical fluctuation anywhere, the counts are allowed to be non-integer] [27, (24)–(25)]. *Operationally:* fill the histogram with the model prediction and fit it like data.

Why do the estimators return the true values on it? The maximum of $\ln \mathcal{L}$ is where every derivative vanishes. For the Poisson part of (4.89), $\ln \mathcal{L} = \sum_i [n_i \ln \nu_i - \nu_i] + \text{const}$, so for any parameter θ_j (including μ)

$$\frac{\partial \ln \mathcal{L}}{\partial \theta_j} \overset{(1)}{=} \sum_i \left(\frac{n_i}{\nu_i} - 1 \right) \frac{\partial \nu_i}{\partial \theta_j} \overset{(2)}{=} 0 \quad \text{when } n_i = \nu_i \text{ for all } i, \tag{4.106}$$

(1) Differentiate $n_i \ln \nu_i - \nu_i$ with the chain rule.

(2) Every bracket vanishes. The Gaussian constraint terms vanish in the same way when $a_k = \theta_k$.

So the true parameters are a maximum [27, (23)]: fitted on the Asimov data set, $\hat{\mu} = \mu'$ and $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$.

Why does it give the median? By (4.100), q_0 and $\tilde{q}_\mu$ are monotonic functions of $\hat{\mu}$, and $\hat{\mu}$ is Gaussian with mean and median μ'. The median of a monotonic function is that function of the median. And the value of the statistic at $\hat{\mu} = \mu'$ is exactly what the Asimov data set gives. So

$$\operatorname{med}[Z_0 \,|\, \mu'] = \sqrt{q_{0,\mathrm{A}}}, \qquad \sigma_{\mathrm{A}}^2 = \frac{(\mu - \mu')^2}{q_{\mu,\mathrm{A}}}, \tag{4.107}$$

where the subscript A marks a statistic evaluated on the Asimov data set [27, (29)–(30), (79)]. The second formula reads (4.100) backwards: on the Asimov data set $\hat{\mu} = \mu'$, so $-2\ln\lambda_{\mathrm{A}}(\mu) = (\mu - \mu')^2/\sigma^2$ can be solved for σ. This is how the σ in all the formulas is obtained in practice: one fit to one artificial data set, nuisance parameters included.

The expected significance of a counting experiment. Take a count with known background b and nominal signal s, and evaluate $q_0 = 2[n\ln(n/b) - (n-b)]$ on the Asimov count $n = s + b$:

$$Z_{\mathrm{A}}^2 = q_{0,\mathrm{A}} \overset{(1)}{=} 2\left[(s+b)\ln\frac{s+b}{b} - s\right] \overset{(2)}{=} 2\left[(s+b)\ln\left(1+\frac{s}{b}\right) - s\right],$$
$$Z_{\mathrm{A}} = \sqrt{2\left[(s+b)\ln\left(1+\frac{s}{b}\right) - s\right]} \overset{(3)}{\approx} \frac{s}{\sqrt{b}}\left(1 - \frac{s}{6b}\right). \tag{4.108}$$

(1) The known-background q_0 at $n = s + b$, where $n - b = s$.
(2) $(s+b)/b = 1 + s/b$.
(3) The expansion $q_{0,\mathrm{A}} = (s^2/b)[1 - s/3b + \dots]$ worked out in the first part of the chapter, and $\sqrt{1-u} \approx 1 - u/2$.

This is the same algebra as in the first part of the chapter; what is new is its meaning, the *median* significance, which (4.107) supplies. The familiar $s/\sqrt{b}$ is its small-signal limit and overestimates it. For $s = b = 10$, $Z_{\mathrm{A}} = 2.78$.

With the background measured in a control region, the Asimov counts are $n = s + b$ and $m = \tau b$, and (4.97) with $\hat{\hat{b}}(0) = (s + b + \tau b)/(1+\tau)$ gives

$$Z_{\mathrm{A}}^2 = 2\left[(s+b)\ln\frac{(1+\tau)(s+b)}{s+(1+\tau)b} + \tau b\ln\frac{(1+\tau)\,b}{s+(1+\tau)b}\right]. \tag{4.109}$$

For $s = b = 10$ and $\tau = 1$ this is $Z_{\mathrm{A}} = 1.84$, against 2.78 with the background known: measuring the background in a region no larger than the signal region costs a third of the expected significance. As $\tau \to \infty$ the control region measures b perfectly: the second logarithm is $\ln[1 - s/(s + (1+\tau)b)] \approx -s/(\tau b)$, so the second term tends to $-s$, the first to $(s+b)\ln(1+s/b)$, and (4.108) returns. The toys agree: their median q_0 under the signal is 3.40, the same as the Asimov value 3.40 to the digits shown. (Here the expected counts are whole numbers, so the Asimov data set is itself a possible outcome of the experiment.)

4.16.4 CL_s: not excluding what we cannot see

The limit (4.105) has the defect met at the start of Section 4.13: a downward fluctuation of the background makes it negative. There is a sharper way to see what is wrong, and it is the reason CL_s exists.

The event. An experiment with almost no sensitivity to some signal (because the signal is tiny compared with the background, $\mu/\sigma \ll 1$) reports that it excludes that signal at 95% confidence. *What made it exclude?* The data cannot tell signal from no signal, so it can only have been a downward fluctuation of the background. *How often does that happen when there is no signal?* The

rule "exclude μ if $p_\mu < \alpha$" with $p_\mu = \Phi((\hat{\mu} - \mu)/\sigma)$ excludes when $\hat{\mu} < \mu - \sigma\Phi^{-1}(1-\alpha)$. With background only, $\hat{\mu} \sim \mathcal{N}(0, \sigma^2)$:

$$\mathrm{P}(\text{exclude } \mu \,|\, 0) \overset{(1)}{=} \mathrm{P}\Big(\frac{\hat{\mu}}{\sigma} < \frac{\mu}{\sigma} - \Phi^{-1}(1-\alpha)\Big) \overset{(2)}{=} \Phi\Big(\frac{\mu}{\sigma} - \Phi^{-1}(1-\alpha)\Big) \xrightarrow{\mu/\sigma \to 0} \alpha. \tag{4.110}$$

(1) Divide the exclusion condition by σ.
(2) $\hat{\mu}/\sigma$ is standard normal.

An experiment blind to a signal still excludes it in 5% of runs. That is correct coverage (the signal is absent, so excluding it is no error) but it is not a statement any physicist wants to publish: the exclusion says nothing about the signal, only about the luck of the background. The counting experiment of our toys, with a weak signal $s = 1$ on $b = 10$ and $\tau = 1$, has $\sigma = 4.47$ at $\mu = 1$; in background-only toys it excludes $\mu = 1$ in a fraction 0.082 of the runs (the formula: 0.078), not far above the 5% of a completely blind experiment, although one signal event on ten background events measured this way is all but invisible.

Working definition 4.10 (CL_s)

For a tested signal μ and the statistic $\tilde{q}_\mu$, let $\mathrm{CL}_{s+b} = \mathrm{P}(\tilde{q}_\mu \geq \tilde{q}_{\mathrm{obs}} \,|\, \mu)$ be the p-value of the signal hypothesis and $\mathrm{CL}_b = \mathrm{P}(\tilde{q}_\mu \geq \tilde{q}_{\mathrm{obs}} \,|\, 0)$ the same tail computed with background only [how background-poor the data look, judged by the same statistic]. Then

$$\mathrm{CL}_s = \frac{\mathrm{CL}_{s+b}}{\mathrm{CL}_b},$$

and μ is excluded at confidence level $1 - \alpha$ if $\mathrm{CL}_s < \alpha$ [30, 31]. In the notation of [27], $p_\mu/(1 - p_b)$.

This is the same ratio as (4.79) of Section 4.13.4, where the statistic was the count itself; the profile likelihood ratio generalises it to any number of bins and nuisance parameters. *Why does dividing help?* Take the two extremes.

- *No sensitivity.* The signal changes nothing, so both tails are computed from nearly the same distribution: $\mathrm{CL}_{s+b} \approx \mathrm{CL}_b$ and $\mathrm{CL}_s \approx 1$. The signal is never excluded. In the weak-signal toys, CL_s excluded $\mu = 1$ in none of the runs (a fraction 0.0000).
- *Good sensitivity.* Data that disfavour the signal are typical for the background, CL_b is near one, and $\mathrm{CL}_s \approx \mathrm{CL}_{s+b}$. Nothing changes.

In between, dividing by $\mathrm{CL}_b \leq 1$ raises the ratio exactly when the data look background-poor, which is when the plain p-value would have been lucky. The price is over-coverage: a CL_s limit is above the classical one and covers the true value in more than 95% of repetitions, as Figure 4.35 showed for counts. CL_s is a ratio of p-values, not a p-value, and not a probability of the signal.

The asymptotic CL_s limit. By (4.103), $\mathrm{CL}_{s+b} = \Phi((\hat{\mu} - \mu)/\sigma)$ and, with $\mu' = 0$, $\mathrm{CL}_b = \Phi(\hat{\mu}/\sigma)$. The limit solves $\mathrm{CL}_s = \alpha$:

$$\Phi\Big(\frac{\hat{\mu} - \mu_{\mathrm{up}}}{\sigma}\Big) = \alpha\,\Phi\Big(\frac{\hat{\mu}}{\sigma}\Big) \implies \mu_{\mathrm{up}} = \hat{\mu} + \sigma\,\Phi^{-1}\Big(1 - \alpha\,\Phi\Big(\frac{\hat{\mu}}{\sigma}\Big)\Big), \tag{4.111}$$

using $\Phi^{-1}(\alpha u) = -\Phi^{-1}(1 - \alpha u)$. Compare (4.75) of Section 4.13.2: it is the same formula. The asymptotic CL_s limit on a Gaussian estimate is the flat-prior Bayesian upper limit, just as for counts the CL_s limit equalled the flat-prior Bayesian limit (4.80). Two different questions, a

ratio of frequentist tails and a posterior probability with the boundary built into the prior, give the same number. That coincidence is part of why CL_s was accepted: it behaves like a Bayesian limit near the boundary and like a classical one far from it.

For the counting experiment with $s = b = 10$, $\tau = 1$ and an observation of $n = 14$ events in the signal region and $m = 9$ in the control region, the 95% CL_s limit computed with toys (tables of $\tilde{q}_\mu$ under μ and under 0 on a grid of μ, 40 000 toys each) is $\mu_{\mathrm{up}} = 1.37$; the asymptotic formulas, with $\sigma = \sigma_{\mathrm{A}}(\mu)$ from (4.107), give 1.37.

4.16.5 The expected limit and the Brazil band

A limit means little without knowing what the experiment should have achieved. If there is no signal, $\hat{\mu} \sim \mathcal{N}(0, \sigma^2)$, and the CL_s limit (4.111) is an increasing function of $\hat{\mu}$. *The question:* what limit does an experiment with no signal typically report, and in what range? Since the limit is monotonic in $\hat{\mu}$, its quantiles are the limits at the quantiles of $\hat{\mu}$. The median is at $\hat{\mu} = 0$; the $\pm N$ standard-deviation quantiles at $\hat{\mu} = N\sigma$. Putting $\hat{\mu} = N\sigma$ into (4.111),

$$\mu_{\mathrm{up}}(N) \overset{(1)}{=} N\sigma + \sigma\,\Phi^{-1}\big(1 - \alpha\,\Phi(N)\big), \qquad \mu_{\mathrm{up}}(0) \overset{(2)}{=} \sigma\,\Phi^{-1}(1 - \alpha/2) = 1.96\,\sigma, \tag{4.112}$$

(1) (4.111) with $\hat{\mu}/\sigma = N$.
(2) $\Phi(0) = \frac{1}{2}$, and $\Phi^{-1}(0.975) = 1.96$ for $\alpha = 0.05$.

For $N = -2, -1, 0, 1, 2$ the bracket is 1.05, 1.41, 1.96, 2.73 and 3.66 times σ. The band is lopsided: a downward fluctuation of the background cannot push the CL_s limit much below σ, while an upward one raises it freely. (The plain limit (4.105) would have given $1.645\sigma + N\sigma$, down to $-0.36\,\sigma$ at $N = -2$.) Since $\sigma = \sigma_{\mathrm{A}}(\mu)$ depends on the μ being tested, each line of (4.112) is solved for μ_{up} numerically [27, (88)–(89)].

Toys against Asimov. For the counting experiment with $s = b = 10$, $\tau = 1$, the script computes the CL_s limit of 4000 background-only experiments with toys and takes the quantiles:

	-2σ	-1σ	median	$+1\sigma$	$+2\sigma$
toys	0.44	0.62	0.90	1.28	1.71
Asimov, (4.112)	0.48	0.65	0.92	1.32	1.86

For an experiment of twenty events the asymptotic band is within a few per cent in the middle and within ten per cent at the edges, a little conservative throughout. In the tails the integer counts matter, and the toys are the reference.

The Higgs-style plots. For the template search of Section 4.15, the script repeats the whole procedure at every hypothesised mass from 110 to 150 GeV: the observed CL_s limit on μ from the pseudo-data, the expected median and bands from the Asimov background-only data set at that mass, and the local p_0 with its Asimov expectation for $\mu = 1$ (Figure 4.43). The green and yellow bands that give the plot its nickname, the **Brazil band**, are (4.112). The expected limit falls from 0.79 to 0.39 across the range, because the background falls with mass while the signal does not. Away from 125 GeV the observed limit wanders inside the bands, as a background-only experiment should, and $\mu = 1$ is excluded at 34 of the 41 masses. Near 125 GeV the observed limit rises far above the band (1.68 against an expected 0.60): the data there refuse to exclude a signal, and the right panel says why. The local significance at 125 GeV is 3.53σ, close to the 3.32σ expected for $\mu = 1$. It is local: a search over this range must correct it for the look-elsewhere effect of Section 4.14.3.

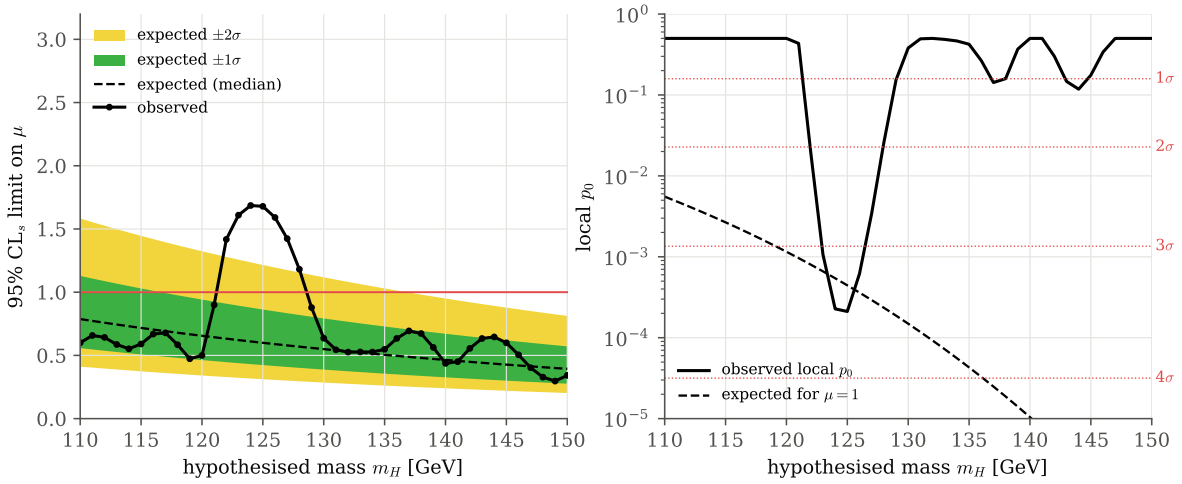


Figure 4.43: The template search as the LHC experiments show it. Left: the 95% CL_s upper limit on the signal strength at each hypothesised mass (solid), with the median expected for background only (dashed) and the $\pm1\sigma$ (green) and $\pm2\sigma$ (yellow) bands. Masses where the solid line is below the thin horizontal line at $\mu = 1$ are excluded for the nominal signal. Right: the local p-value of the background-only hypothesis (solid) and its median expectation if the nominal signal is present (dashed); the dotted lines mark 1 to 4σ.

Key idea: The LHC recipe in four lines

Write the search as a likelihood with nuisance parameters constrained by their auxiliary measurements. Test "no signal" with q_0 and report $Z = \sqrt{q_0}$, local and global. Exclude a signal when $\mathrm{CL}_s = \mathrm{CL}_{s+b}/\mathrm{CL}_b < 0.05$, using $\tilde{q}_\mu$. Show the observed limit against the expected one and its Brazil band, both from the Asimov data set, and check with toys where counts are small.

How this is used in real research: How the experiments compute it

The likelihoods of real searches have hundreds of bins and hundreds of nuisance parameters, with templates interpolated as in (4.91). They are written in standard formats (HistFactory, with the RooStats, `pyhf` and CMS `Combine` tools) and minimised numerically; the asymptotic formulas of [27] give q_0, the CL_s limit and the Brazil band from the observed and Asimov fits alone, and toys are run where counts are small or near the boundary [29]. Our two scripts verify the formulas against toys on a case where both are cheap. In research the same machinery is the instrument: the toys are the reference whenever the formulas are in doubt, and published likelihoods let others combine results, as Section 4.13.1 urged.

The lesson in one line: with nuisance parameters profiled, the significance is still $\sqrt{q_0}$ and the limit still comes from a half-χ^2, and CL_s divides by the background's own p-value so that a signal is never excluded by luck.

4.17 Confidence intervals and credible intervals

Ten one-minute runs of a low-background detector record a total of six events. Two analysts report a 90% interval for the rate λ per run. The first, using Garwood's construction, reports $[0.261,\ 1.184]$. The second, using Bayes' theorem with a mild prior, reports $[0.299,\ 1.077]$ [CasellaBerger Exs. 9.2.15–9.2.16]. The numbers are close. The statements are not. The first is a confidence interval: it comes from a recipe that covers the true rate in at least 90% of repeated

experiments. The second is a **credible interval**: given these data and the prior, it contains the rate with probability 90%. The first is a prediction-direction statement, about which data each rate would produce. The second is an inference-direction statement, about which rates could have produced the data we saw. Bayes' theorem is the bridge between them. Comparing the two on the same data shows what each one claims and what it needs; Chapter 5 develops the Bayesian side in full.

The question in words.

1. What does each interval say, in words, and what is random in each?
2. Does a credible interval cover like a confidence interval, and does a confidence interval hold the posterior probability it claims?
3. When do the two coincide?

4.17.1 A credible set

In the Bayesian framework θ has a probability distribution, first the prior $\pi(\theta)$ and, after the data, the posterior $\pi(\theta \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid \theta)\pi(\theta)$. Any set A of parameter values has a **credible probability** $\mathrm{P}(\theta \in A \mid \boldsymbol{x}) = \int_A \pi(\theta \mid \boldsymbol{x})\,\mathrm{d}\theta$ [CasellaBerger (9.2.18)], and a set with credible probability $1-\alpha$ is a $1-\alpha$ credible set. The different name keeps the two kinds of probability apart. As with confidence intervals, many sets have the required probability, so we must choose one. Two common choices are the *equal-tailed* interval, with posterior probability $\alpha/2$ beyond each end, and the **highest density interval** (HDI) [every point inside has higher posterior density than every point outside], which makes $\pi(a \mid \boldsymbol{x}) = \pi(b \mid \boldsymbol{x})$ and is the shortest interval with the required probability [Cowan (9.42)], [Kruschke §11.3.2].

Example: the Poisson rate. Take n observations from Poisson(λ) with sum $y = \sum x_i$, and a gamma prior with shape a and scale b, density $\propto \lambda^{a-1}e^{-\lambda/b}$. Then

$$\pi(\lambda \mid \boldsymbol{x}) \overset{(1)}{=\!\propto} \Big(\prod_i \lambda^{x_i}e^{-\lambda}\Big)\lambda^{a-1}e^{-\lambda/b} \overset{(2)}{=} \lambda^{a+y-1}e^{-\lambda(n+1/b)} \overset{(3)}{=} \mathrm{Gamma}\Big(a+y,\ \frac{b}{nb+1}\Big), \tag{4.113}$$

(1) Bayes' theorem: likelihood of independent Poisson counts times prior; factors not involving λ are dropped.
(2) Collect powers of λ and the exponentials.
(3) This is the gamma density with shape $a+y$ and rate $n+1/b$, that is scale $1/(n+1/b) = b/(nb+1)$ [CasellaBerger (9.2.19)].

A gamma variable of shape k and scale s, multiplied by $2/s$, is χ^2 with $2k$ degrees of freedom, so $2(nb+1)\lambda/b \sim \chi^2_{2(a+y)}$ under the posterior, and the equal-tailed interval has χ^2 quantiles for ends [CasellaBerger (9.2.20)]. With $a = b = 1$, $n = 10$, $y = 6$: $22\lambda \sim \chi^2_{14}$, and the 5% and 95% quantiles 6.571 and 23.685 give $[0.299, 1.077]$. The credible interval is shorter than the confidence interval and pulled towards zero. Both effects come from the prior. It adds information, which narrows the interval. And it favours small rates (its density $e^{-\lambda}$ falls with λ), which pulls the interval down.

4.17.2 Judging each interval by the other's standard

Each interval keeps its own promise exactly. *The question:* how does each do when judged by the other's promise [CasellaBerger Ex. 9.2.17]?

Coverage of the credible interval. Fix a true λ. Simulate the total of the ten runs, $y \sim \text{Poisson}(10\lambda)$, many times (or sum exactly over y), build the credible interval each time and count hits. Figure 4.44 (left) shows the result. For small λ near the prior's bulk the credible interval covers about 90%, sometimes more, sometimes less (minimum 0.727 on $0 < \lambda < 3$). For large λ the prior's pull towards small values becomes a bias, and the coverage collapses: 0.844 at $\lambda = 5$, 0.620 at $\lambda = 20$, 0.302 at $\lambda = 50$, tending to zero [CasellaBerger (9.2.22)]. A credible interval is not a confidence interval.

Credible probability of the confidence interval. Fix the data, compute the posterior, and ask how much posterior probability lies inside the Garwood interval. At $y = 6$ it is 0.947; it decreases steadily with y, reaching 0.868 at $y = 50$, and tends to zero for $y \to \infty$ unless the prior scale is $b = 1/n$ (Figure 4.44, right). A confidence interval is not a credible interval either.

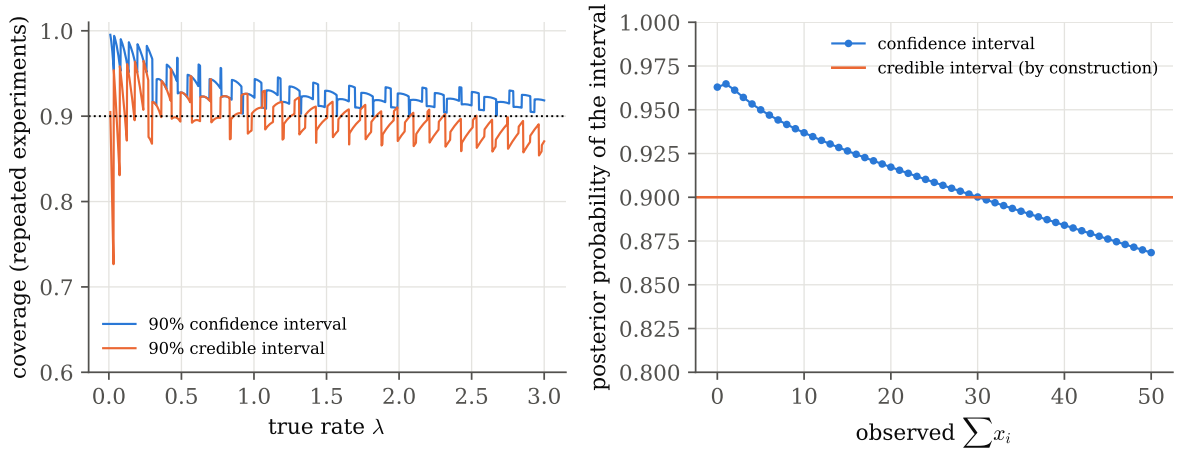


Figure 4.44: Two 90% intervals for a Poisson rate from ten observations, judged by each other's criterion. Left: the frequentist coverage as a function of the true rate. The confidence interval (blue) never drops below 90%; the credible interval with a gamma(1, 1) prior (orange) scatters around 90% at small rates and falls away as the rate grows past the prior's range. Right: the posterior probability contained in each interval, as a function of the observed total. The credible interval holds 90% by construction; the confidence interval holds more for small totals and less for large ones.

Example: a Gaussian mean with a Gaussian prior. Here the calculation can be done in closed form [CasellaBerger Ex. 9.2.18]. Let $X_1, \dots, X_n$ be $\mathcal{N}(\theta, \sigma^2)$ and the prior $\mathcal{N}(\mu, \tau^2)$, and write $\gamma = \sigma^2/(n\tau^2)$ for the ratio of the data variance of the mean to the prior variance. The posterior is Gaussian with mean $\delta = (\gamma\mu + \bar{x})/(1+\gamma)$ and variance $V = \sigma^2/(n(1+\gamma))$, and the $1-\alpha$ credible interval is $\delta \pm z_{\alpha/2}\sqrt{V}$. Its coverage at a fixed true θ, with $Z = \sqrt{n}(\bar{X} - \theta)/\sigma \sim \mathcal{N}(0,1)$ and $\Delta = \sqrt{n}(\theta - \mu)/\sigma$, is

$$
\begin{aligned}
\mathrm{P}_\theta\big(|\theta - \delta| \le z_{\alpha/2}\sqrt{V}\big) &\overset{(1)}{=} \mathrm{P}_\theta\left(\left|\frac{\gamma(\theta-\mu) - (\bar{X}-\theta)}{1+\gamma}\right| \le z_{\alpha/2}\frac{\sigma}{\sqrt{n(1+\gamma)}}\right) \\
&\overset{(2)}{=} \mathrm{P}\big(|\gamma\Delta - Z| \le \sqrt{1+\gamma}\, z_{\alpha/2}\big) \\
&\overset{(3)}{=} \Phi\big(\gamma\Delta + \sqrt{1+\gamma}\, z_{\alpha/2}\big) - \Phi\big(\gamma\Delta - \sqrt{1+\gamma}\, z_{\alpha/2}\big).
\end{aligned} \tag{4.114}
$$

(1) $\theta - \delta = [(1+\gamma)\theta - \gamma\mu - \bar{X}]/(1+\gamma) = [\gamma(\theta-\mu) - (\bar{X}-\theta)]/(1+\gamma)$.
(2) Multiply both sides by $(1+\gamma)\sqrt{n}/\sigma$ and use the definitions of Z and Δ.
(3) Z is standard normal: the probability that it lies within $\sqrt{1+\gamma}\, z_{\alpha/2}$ of $\gamma\Delta$.

So the coverage depends on how far the truth is from the prior mean, measured by Δ in units of the standard error $\sigma/\sqrt{n}$. If the truth sits at the prior mean, $\Delta = 0$, the interval overcovers (for $\gamma = 1$ and 95%, the coverage is 0.994), because the prior's information is correct. If the truth is two or four standard errors away from the prior mean, the coverage is 0.780 or 0.110 (Figure 4.45, right). As $\tau \to \infty$, $\gamma \to 0$ and (4.114) becomes $\Phi(z_{\alpha/2}) - \Phi(-z_{\alpha/2}) = 1 - \alpha$ for every θ: with a flat prior the credible interval is the confidence interval $\bar{x} \pm z_{\alpha/2}\sigma/\sqrt{n}$, number for number. This coincidence, like the counting-experiment coincidence of Cowan's (9.55), is special to location parameters and flat priors; it is why the two schools rarely argue about a well-measured Gaussian quantity.

4.17.3 The coin again: the HDI does not depend on the stopping rule

Kruschke returns to $z = 7$ heads in $N = 24$ flips with a prior beta(11, 11), a coin that looks fair [Kruschke §11.3.2]. The posterior is beta(11 + 7, 11 + 17) and its 95% HDI is $[0.254, 0.531]$; with a uniform prior it would be $[0.141, 0.483]$. Figure 4.45 (left) shows both posteriors with the four confidence intervals of Section 4.10.4 underneath. The four confidence intervals differ; the HDI is the same for all four. *Why?* The Bayesian interval sees the data only through the likelihood $\theta^7(1-\theta)^{17}$, and that is the same whether the experimenter meant to stop at $N = 24$, at $z = 7$, or when time ran out. The confidence interval also depends on which cloud of imaginary outcomes the plan would have produced. Kruschke lists three advantages of the HDI [Kruschke pp. 324–325]: it is a direct statement about the credibility of values of θ; it does not depend on sampling and testing intentions; and it responds, openly, to prior knowledge. The price is the prior itself, which must be stated and defended.

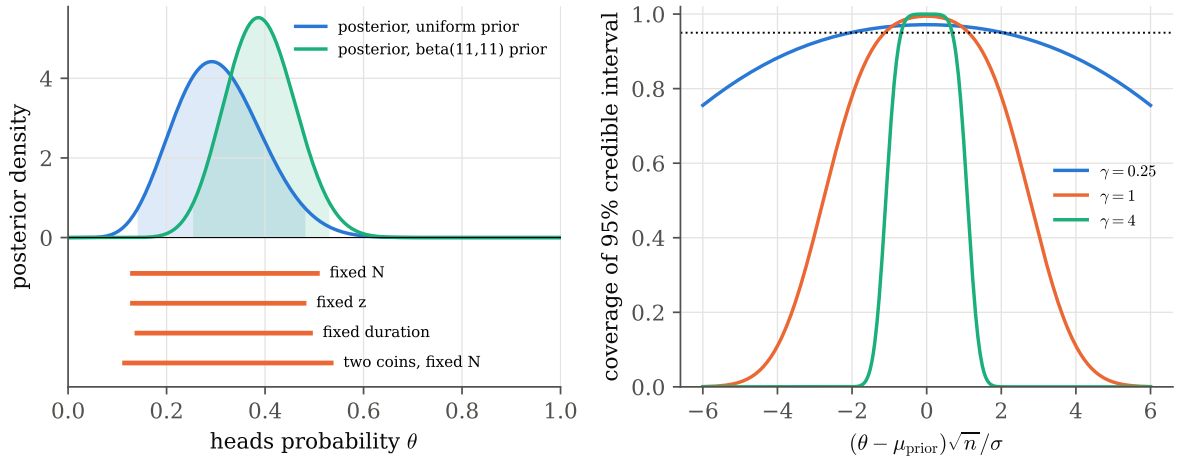


Figure 4.45: Left: posteriors for 7 heads in 24 flips with a uniform (blue) and a beta(11, 11) (green) prior; the shaded parts are the 95% highest density intervals. Below the axis, the four 95% confidence intervals that four stopping intentions produce from the same data. Right: frequentist coverage of a 95% credible interval for a Gaussian mean with a Gaussian prior, as a function of the distance between the true value and the prior mean (in units of the standard error), for three ratios γ of data variance to prior variance. A strong prior (large γ) covers only near its own centre.

Deciding with an interval: the ROPE. A test asks whether θ equals a null value exactly, a question whose answer is almost always "no" with enough data. Kruschke replaces it by a **region of practical equivalence** (ROPE) [a small range of values that are, for the purpose at hand, as good as the null value], and a rule [Kruschke §12.1.1]: reject the null value if the 95% HDI lies entirely outside the ROPE; accept it for practical purposes if the HDI lies entirely inside; otherwise, withhold judgement. For a coin with ROPE $[0.45, 0.55]$ and a uniform prior,

325 heads in 500 flips give the HDI $[0.608, 0.691]$, outside the ROPE: not fair. 490 heads in 1000 give $[0.459, 0.521]$, inside: fair for practical purposes (the posterior probability of the ROPE is 0.994). Unlike a p-value, this rule can *accept* a null value, and only when the measurement is precise enough to deserve it. The physicist's analogue is familiar: a calibration is declared good when its uncertainty band lies within the tolerance, not when a test fails to reject perfection.

4.17.4 What each interval says

The two kinds of interval differ in what they claim, what they treat as random, what they need as input, what they guarantee, and how they typically go wrong.

Taxonomy: Confidence interval or credible interval

	confidence interval	credible interval
in words	"a recipe that, applied to data from any θ, contains that θ 95% of the time"	"given these data and this prior, θ is in here with probability 95%"
random	the interval (through the data)	θ (through the posterior)
fixed	the true θ	the observed data
needs	$p(\boldsymbol{x} \mid \theta)$ for all data that could have occurred, including the stopping and search plan	the likelihood of the observed data and a prior
guarantee	coverage for every θ	correct posterior probability, if the prior is right
typical failure	empty or unphysical intervals, flip-flopping, dependence on intentions	poor coverage where the prior is wrong
coincide when	location parameters with flat priors: Gaussian mean, Poisson with no background	

In the pipeline of the book, both are ways of turning the sampling distribution $P(S \mid \theta)$ of a summary into a statement about the physics. The confidence interval stays in the prediction direction and asks which θ would make the observed summary typical. The credible interval crosses into the inference direction and asks how probable each θ is. Cosmologists quote credible intervals from MCMC posteriors (Chapter 9); particle physicists quote confidence intervals and limits; the Fisher forecasts of Chapter 10 are, for linear-Gaussian models, both at once. Chapter 5 builds the Bayesian side properly: priors, the evidence, and the reallocation of credibility that a posterior is.

Recap

- A $1 - \alpha$ confidence interval is a recipe with coverage at least $1 - \alpha$ for every parameter value. The interval is random, the parameter fixed.
- Inverting a family of level-α tests gives a $1 - \alpha$ confidence set: keep the null values that are not rejected. The p-value function displays all the tests at once.
- The Neyman construction draws the acceptance regions of all tests as a belt in the (data, parameter) plane; the interval is its vertical cut. Gaussian estimates give $\hat{\theta} \pm \sigma$, Poisson counts give Garwood's χ^2 intervals, and $\Delta \ln \mathcal{L} = \frac{1}{2}$ (or $\Delta\chi^2 = 1$) is the large-sample shortcut. With m parameters the 68.3% region is $\Delta\chi^2 = 2.30$ for $m = 2$.

- Coverage is checked by simulation. Normal quantiles with an estimated σ undercover at small n; $n \pm \sqrt{n}$ fails at small counts; discrete data force over-coverage with a saw-tooth.
- Near a physical boundary the classical limit can be negative or the interval empty. Choosing between limit and interval after seeing the data undercovers (85% for a nominal 90%). Feldman–Cousins ordering by a likelihood ratio fixes both; CL_s (equal to the flat-prior Bayesian limit for counting) protects against excluding what the experiment cannot see.
- Many tests: Bonferroni controls the chance of any false rejection, BH the fraction of false discoveries. In a scan, the global p-value exceeds the local one by a trials factor that grows linearly with the significance; the Gross–Vitells formula gets it from a few toys.
- A credible interval is a probability statement about θ given the data and a prior. It does not guarantee coverage; a confidence interval does not hold its nominal posterior probability; they agree for flat priors on location parameters.

Reading the literature

- **Main line.** [Wasserman Ch. 10] for hypothesis testing: the null and the alternative, power and size, the Wald test, p-values and their uniformity, the χ^2 and permutation tests, likelihood-ratio tests, multiple testing (Bonferroni and Benjamini–Hochberg) and goodness of fit; [Wasserman §6.3.2] for confidence sets, their interpretation and the Berger–Wolpert example.
- **The physicist's version.** [Cowan Ch. 4] for test statistics, particle selection, the Neyman–Pearson lemma, goodness of fit and the significance of a signal; [Cowan Ch. 9] for confidence belts, Gaussian and Poisson intervals, Fisher's z, likelihood intervals, confidence regions and limits near a boundary, with and without background.
- **Careful derivations.** [CasellaBerger §8.3] for the Neyman–Pearson lemma and valid p-values; [CasellaBerger §10.3] for the asymptotic distribution of the likelihood ratio; [CasellaBerger §§9.1–9.2] for interval estimators, coverage and confidence coefficients, test inversion, pivots, Garwood's Poisson intervals and the coverage of credible sets.
- **The Bayesian critique.** [Kruschke Ch. 11] for how p-values and confidence intervals depend on stopping and testing intentions, and for multiple comparisons; [Kruschke Ch. 12] for the HDI and the region of practical equivalence. [Lambert §2.8] for the frequentist's imaginary samples.
- **Original papers.** Feldman and Cousins' unified approach to small signals [28]; Gross and Vitells on trials factors for the look-elsewhere effect [15]; Cowan, Cranmer, Gross and Vitells on asymptotic formulae for likelihood-based tests of new physics, including the profile likelihood ratio, the Asimov data set and the CL_s remark [27]; Wilks on the asymptotic χ^2 distribution of likelihood ratios [22].

Chapter summary

Before a measurement is published it must answer three questions. Could this be chance? Which parameter values are compatible with what we saw? How often do our recipes mislead us? All three start from the sampling distribution of a summary. Tests answer the first: they are built around the one hypothesis whose predictions we know, the null, and measure surprise with p-values. Intervals and limits answer the second: they are families of tests, inverted. Coverage checks and corrections for the many places we looked answer the third. In the pipeline of the book,

$$\text{model} \to \text{data} \to \text{summary } S \to P(S \mid \theta) \to \boxed{\text{infer physics}},$$

tests and intervals are the frequentist form of the last arrow, and Chapter 5 gives the Bayesian one.

Testing the null (first half)

concept	operational meaning	used later in
null hypothesis	the hypothesis whose sampling distribution we know and can simulate: "it happened by chance"	every search, Chapters 8 and 12
test statistic, rejection region	one number, large under the alternative; reject when it exceeds a critical value fixed in advance	Chapters 8 and 12
p-value	probability under H_0 of a statistic at least as extreme, in the alternative's direction; simulate if no formula	Chapters 6, 8 and 12
one- and two-sided, stopping rule	the direction and the experimental plan are part of the question; they change the p-value	Chapter 5
size, power	false-alarm rate α; rejection probability under the truth; plan the experiment by the power	Chapter 10
uniformity of p	under a continuous null $p \sim \mathrm{U}(0,1)$; for discrete data $\mathrm{P}(p \le \alpha) \le \alpha$; a histogram of null p-values checks a pipeline	Chapters 6 and 8
$p \neq \mathrm{P}(H_0 \mid \text{data})$	the posterior of the null needs a prior and the power (Bayes)	Chapter 5
Wald, t, permutation tests	estimate over standard error; Student t for small Gaussian samples; relabel the data for a distribution-free null	Chapters 6 and 12
Neyman–Pearson, likelihood ratio	for simple hypotheses the likelihood ratio is the most powerful statistic	Chapters 8 and 10
Wilks, q_0, $Z = \sqrt{q_0}$	$-2\ln\lambda \sim \chi^2$ with as many degrees of freedom as the null fixes; half-χ^2 on a boundary	Chapters 9 and 10
Pearson χ^2, KS	goodness of fit of a whole model; fitted parameters reduce the degrees of freedom; simulate sparse cases	Chapters 7 and 8
$p \to Z\sigma, 5\sigma$	$Z = \Phi^{-1}(1-p)$; 3σ evidence, 5σ observation, because of many searches, backgrounds and rarity	Chapter 12

Intervals, coverage and many tests (second half)

concept	operational meaning	used later in
confidence interval	a recipe with coverage $\geq 1-\alpha$ for every θ; the interval is random, θ fixed	Chapters 8 and 10
test inversion, p-value function	keep the null values a level-α test does not reject	Chapter 8
Neyman belt	acceptance intervals at every θ, cut vertically at the observation; build on a grid from simulations when there is no formula	Chapters 6 and 8
Gaussian and Poisson intervals	$\hat{\theta} \pm \sigma$ is the 68.3% central interval; Garwood's χ^2 quantiles for counts; $-\ln 0.05 = 3$ for zero events	Chapters 7 and 8
likelihood interval, $\Delta\chi^2$ regions	$\Delta \ln \mathcal{L} = \frac{1}{2}$, $\Delta\chi^2 = 1$ in one dimension; 2.30 (68.3%) and 5.99 (95%) in two	Chapters 9 and 10
coverage check	simulate at fixed θ, count hits; exact sums for counts; t not z for estimated σ	Chapters 6, 8 and 9
limits near a boundary	report $\hat{\theta} \pm \sigma$ always; classical, shifted, Bayesian, Feldman–Cousins, CL_s; fix the recipe in advance	Chapter 10
Bonferroni, BH	$P_i < \alpha/m$ controls any false rejection; the BH line $i\alpha/m$ controls the false discovery rate	Chapter 12
look-elsewhere effect	global p-value from toys scanned like the data; $p_{\mathrm{glob}} \approx p_{\mathrm{loc}} + \mathbb{E}\, N(c_0) e^{-(c-c_0)/2}$; trials factor $\propto Z$	Chapters 8 and 12
search likelihood, nuisance parameters	Poisson bins of signal and background templates times one constraint term per calibration; profile to get $\lambda(\mu)$; $\sigma^2_{\mathrm{syst}} = \sigma^2_{\mathrm{tot}} - \sigma^2_{\mathrm{stat}}$	Chapters 8 and 10
q_0, $\tilde{q}_\mu$, CL_s, Brazil band	$Z = \sqrt{q_0}$ with profiled nuisances (Wald); exclude when $\mathrm{CL}_{s+b}/\mathrm{CL}_b < 0.05$; expected limit and $\pm 1, 2\sigma$ bands from the Asimov data set; toys for small counts	Chapters 10 and 12
credible interval, HDI, ROPE	posterior probability of a set given data and prior; shortest set; accept or reject a value for practical purposes	Chapters 5 and 9

Chapter 5

Bayes' Theorem and Bayesian Inference

Where this chapter is going

The same observation can admit several explanations. A positive test can occur with or without disease, and a run of heads can come from a fair or a biased coin. In cosmology, several values of Ω_m can be consistent with the sky we observe. In each case our model tells us how likely the data are under each hypothesis. Bayes' theorem uses those likelihoods together with the prior probabilities to tell us how much probability to assign to each hypothesis after the observation. Picture the sample space with area proportional to probability. Conditioning restricts our attention to the region compatible with the observation, and the posterior probability of a hypothesis is its share of the area that remains.

With a handful of discrete hypotheses the picture can be drawn and computed by hand. Credibility is reallocated: hypotheses that predicted the observation well gain, the others lose, and the total stays one. The ratio of two hypotheses changes by a single factor, the likelihood ratio, so evidence multiplies odds and adds in logarithms. Observations arriving one by one are absorbed one by one, the posterior of each step becoming the prior of the next, and the order does not matter. The denominator of the theorem is the sum over every path that could have led to the observation. When the hypotheses become the values of a continuous parameter, the list becomes a density, the likelihood comes from a model of the noise, and we need ways to summarise a posterior and to integrate out parameters we do not care about. Finally the hypotheses can be whole models with their own parameters. Then the denominator, the evidence, is what compares them, and computing it in many dimensions is so hard that it motivates the sampling methods of Chapter 9.

Bayes' theorem is the last arrow of the pipeline of Section 0.3: it takes the sampling distribution $P(S \mid \theta)$ of a summary S and turns it into a statement about the physics. Chapters 2 to 4 built the forward object $P(S \mid \theta)$; here it is read backwards.

5.1 One joint table, read in two directions

A collider experiment writes an event to disk because its first-level trigger fired. Three independent trigger paths can fire it: the muon system, the calorimeter, and a fast track trigger. From the trigger rates we know what share of the recorded events each path is responsible for: 25% muon, 60% calorimeter, 15% track. The event record of one particular event has lost the flag that says which path fired. Offline, the reconstruction finds an isolated muon with large transverse momentum in it. Which path most probably fired the trigger?

The physics of each trigger tells us something in the *forward* direction. If the muon path fired, the event contains an isolated high-p_T muon 70% of the time (the rest are fakes and

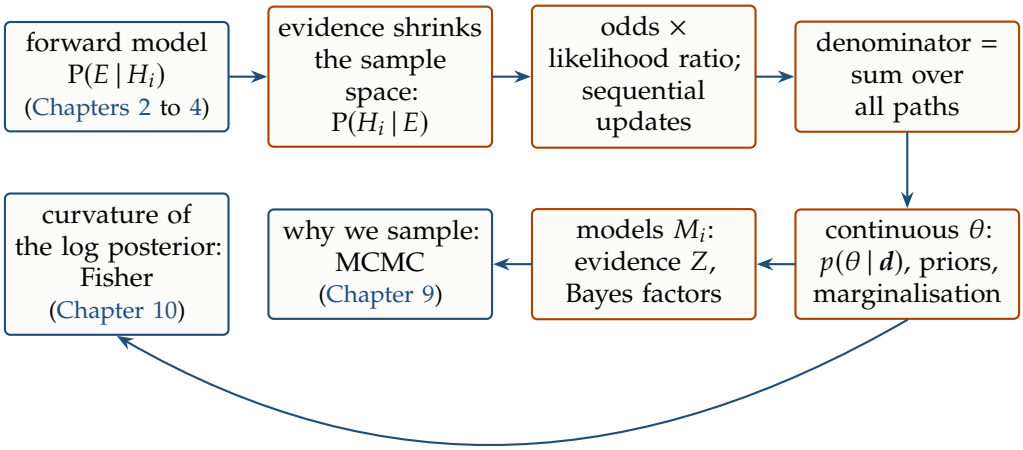


Figure 5.1: How the ideas of Bayesian inference connect. The top row works with a list of discrete hypotheses: a forward model gives the probability of the observation under each, the observation shrinks the sample space, odds change by the likelihood ratio, and the denominator collects every path to the observation. The bottom row replaces the list by a continuous parameter and then by competing models; the cost of the denominator in many dimensions leads to Markov chain Monte Carlo, and the curvature of the log posterior near its peak leads to the Fisher matrix. The boxes with blue borders belong to other chapters.

muons outside the offline acceptance). If the calorimeter path fired, only 5% of such events contain one. If the track path fired, 20% do. These numbers answer "given the path, what do we see?". We want the reverse: "given what we see, which path?".

The two questions use probability in two opposite directions. In the first, we fix one possible situation (here, the path that fired) and ask what it can lead to: which outcomes, and how probable each one is. The outcome is still uncertain, and probability forecasts it. This is **prediction**. In the second, the outcome has already happened: the muon is on disk. Several situations could have led to it, and we want to know which one did. So we start from the outcome and work back to its possible sources. We ask which hypotheses could have produced what we saw, and how probable each one is now that we have seen it. This is **inference**. We met the two directions in Chapter 0 and read a first tree backwards in Chapter 1. All of Bayesian inference rests on this one step: turning the forward question into the backward one.

The question in words. Every backward question can be put into four short questions before a single symbol is written:

1. What did we actually see?
2. Which hypotheses could have given rise to it?
3. Under each of those hypotheses, how probable was what we saw?
4. In the light of that evidence, how must the probabilities we gave the hypotheses beforehand be revised?

For the trigger: (1) we observed an isolated high-p_T muon, call it E; (2) the three paths H_μ, H_{cal}, H_{trk} could have produced it, and exactly one of them fired this event (we ignore the rare events fired by two paths at once); (3) the observation had probability 0.70, 0.05 and 0.20 under them; (4) the original probabilities 0.25, 0.60, 0.15 must be updated in the light of E. Steps 1 to 3 are prediction, done once for each hypothesis. Only step 4 is new.

5.1.1 Deduction runs forward, inference runs backward

The same two directions appear in the difference between deduction and inference [Sivia §1.1, Fig. 1.1]. In *deductive* reasoning the cause is given and we work out its consequences. If a coin is fair, the probability of seven heads in ten tosses follows by counting. Pure mathematics, and every game of chance, runs this way. Most of science runs the other way. We see the effects and ask for the cause: ten tosses gave seven heads; is the coin fair? Several causes could have produced the same effect, so no amount of logic makes the answer certain. Sivia calls this *plausible reasoning*. The question is whether plausible reasoning has rules we can compute with.

It does, and they are the rules of probability we already know. Cox asked a plain question: if we give numbers to degrees of belief, what rules must those numbers obey for the reasoning to be consistent? Consistent means, among other things, that two different ways of using the same information must give the same answer. His answer was that any such numbers can be relabelled so that they obey exactly two rules [Sivia §1.2, (1.1)–(1.2)]:

$$\mathrm{P}(X \mid I) + \mathrm{P}(\bar{X} \mid I) = 1 \qquad \text{(sum rule)}, \tag{5.1}$$

$$\mathrm{P}(X, Y \mid I) = \mathrm{P}(X \mid Y, I)\ \mathrm{P}(Y \mid I) \qquad \text{(product rule)}. \tag{5.2}$$

Here $\bar{X}$ ["X is false"] is the complement, the comma means "and", and I is the **background information** [everything we take as given, such as the trigger rates and the detector model]. Every probability in inference is conditional on some I, even when we stop writing it. Many heated arguments about data analysis are really arguments about an I that nobody wrote down [Sivia §1.2].

The product rule already contains Bayes' theorem. The joint statement "X and Y" is the same statement as "Y and X", so the product rule can be applied in either order and must give the same number:

$$\mathrm{P}(X \mid Y, I)\ \mathrm{P}(Y \mid I) \overset{(1)}{=} \mathrm{P}(X, Y \mid I) \overset{(2)}{=} \mathrm{P}(Y, X \mid I) \overset{(3)}{=} \mathrm{P}(Y \mid X, I)\ \mathrm{P}(X \mid I).$$

Dividing by $\mathrm{P}(Y \mid I) > 0$,

$$\mathrm{P}(X \mid Y, I) \overset{(4)}{=} \frac{\mathrm{P}(Y \mid X, I)\ \mathrm{P}(X \mid I)}{\mathrm{P}(Y \mid I)}. \tag{5.3}$$

(1) Product rule (5.2).
(2) "X and Y" is logically the same proposition as "Y and X".
(3) Product rule again, with the roles of X and Y exchanged.
(4) Divide both sides of the first line by $\mathrm{P}(Y \mid I)$.

Put $X = H$, a hypothesis, and $Y = E$, the observed evidence. The theorem turns the quantity physics hands us, $\mathrm{P}(E \mid H, I)$, into the quantity we want, $\mathrm{P}(H \mid E, I)$ [Sivia §1.3], [Kruschke §5.1.1], [Lambert §3.6].

Bayes' theorem is the first consequence of the two rules. The second is **marginalisation** [Sivia (1.5)–(1.7)]: it answers "how probable is E when we do not know which hypothesis is true?". If $H_1, \ldots, H_k$ are **mutually exclusive and exhaustive** [exactly one of them is true], then

$$\mathrm{P}(E \mid I) \overset{(1)}{=} \sum_{i=1}^{k} \mathrm{P}(E, H_i \mid I) \overset{(2)}{=} \sum_{i=1}^{k} \mathrm{P}(E \mid H_i, I)\ \mathrm{P}(H_i \mid I). \tag{5.4}$$

(1) E happens together with exactly one of the H_i, so its probability is the sum of the probabilities of the k disjoint pieces $E \cap H_i$. This is the law of total probability of Chapter 1; Sivia derives it from the sum and product rules alone.

(2) Product rule on each term.

It is the denominator of Bayes' theorem, and it will get a section of its own (Section 5.6).

Put $X = H_i$ in (5.3) and replace the denominator by the marginalisation (5.4). For mutually exclusive and exhaustive hypotheses $H_1, \dots, H_k$ and an observation E with $\mathrm{P}(E) > 0$ this gives the working form of Bayes' theorem:

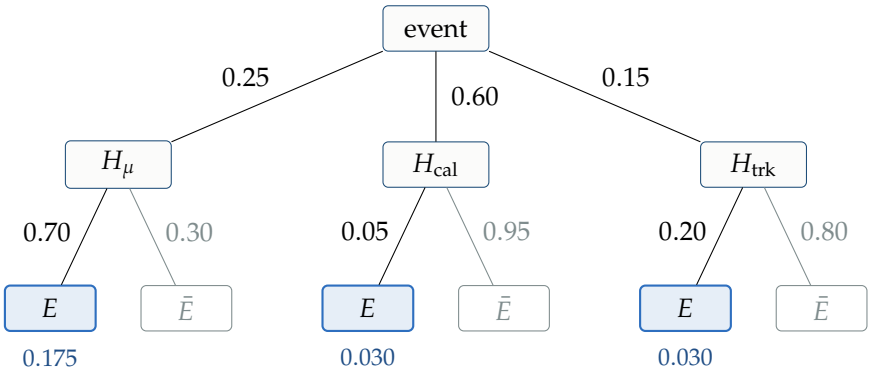
$$\underbrace{\mathrm{P}(H_i \mid E)}_{\text{posterior}} = \frac{\overbrace{\mathrm{P}(E \mid H_i)}^{\text{likelihood}}\ \overbrace{\mathrm{P}(H_i)}^{\text{prior}}}{\underbrace{\sum_j \mathrm{P}(E \mid H_j)\ \mathrm{P}(H_j)}_{\text{evidence } \mathrm{P}(E)}}. \tag{5.5}$$

Key idea: Bayes' theorem in words

Posterior = likelihood × prior / evidence. It is the product rule applied in both orders. The numerator is one path through the tree; the denominator is the sum over every path that ends in E.

5.1.2 The tree and the trigger

Draw the trigger problem as a tree (Figure 5.2). The first split is nature's choice of path, with the priors on the branches. The second split is what the offline reconstruction sees, with the forward probabilities on the branches. Multiplying along a path gives the probability of the leaf, the joint probability $\mathrm{P}(H_i \cap E)$.



leaf probabilities of the observed branch E (products along each path)

Figure 5.2: The trigger problem as a tree. Nature first chooses the path that fires, with the priors on the upper branches; the lower branches give the probability that each path's events contain an isolated high-p_T muon (E) or not ($\bar{E}$). The observed branch is outlined in blue, and the number under each blue leaf is the product along its path. Inference keeps only the blue leaves and asks what share of their total each one carries.

Prediction reads the tree top-down: pick a path, follow it, get the probability of each outcome. Inference starts at the bottom. We saw the muon, so we know we are on a blue leaf, but not on which one. So we collect the three blue leaves and ask what share of their total each one holds:

$$\mathrm{P}(E) \overset{(1)}{=} 0.25 \times 0.70 + 0.60 \times 0.05 + 0.15 \times 0.20 = 0.175 + 0.030 + 0.030 = 0.235,$$

$$\mathrm{P}(H_\mu \mid E) \overset{(2)}{=} \frac{0.175}{0.235} \approx 0.745, \qquad \mathrm{P}(H_{\text{cal}} \mid E) = \mathrm{P}(H_{\text{trk}} \mid E) = \frac{0.030}{0.235} \approx 0.128.$$

(1) Marginalisation (5.4): the sum of the three blue leaves.
(2) Bayes' theorem (5.5): one blue leaf over all of them.

The muon path most probably fired the trigger. The calorimeter path was the most probable before we looked (0.60), and it is now one of the two least probable. It lost because it rarely produces what we saw. The track path was the least probable before, and it ends level with the calorimeter path. Only the product of prior and likelihood enters, so a small prior times a larger likelihood (0.15×0.20) can match a large prior times a small likelihood (0.60×0.05). The prior still matters: the posteriors are *not* the likelihoods rescaled to sum to one. Those would be 0.70, 0.05, 0.20 divided by their sum 0.95, that is 0.74, 0.05, 0.21.

5.1.3 The joint table: a column for prediction, a row for inference

A tree is one way to store a joint distribution. A table is another, and it makes the two directions of reading especially plain [Kruschke §5.1.2, Tables 5.1 and 5.5]. Let the columns be the hypotheses and the rows be the possible observations. Each cell holds the joint probability $\mathrm{P}(E_r, H_c) = \mathrm{P}(E_r \mid H_c)\,\mathrm{P}(H_c)$. Summing a column gives the prior $\mathrm{P}(H_c)$ in the bottom margin; summing a row gives $\mathrm{P}(E_r)$ in the right margin. Figure 5.3 shows it for the trigger.

	H_μ	H_{cal}	H_{trk}	row sum
E	0.175	0.030	0.030	0.235
$\bar{E}$	0.075	0.570	0.120	0.765
column sum	0.25	0.60	0.15	1

prediction: one column, divided by its sum
inference: the observed row, divided by its sum

Figure 5.3: The joint probabilities of trigger path (columns) and offline observation (rows). A prediction fixes a hypothesis and looks down its column: $0.175/0.25 = 0.70$ is the probability of seeing the muon if the muon path fired. An inference fixes the observation and looks along its row: $0.175/0.235 = 0.745$ is the probability that the muon path fired given that we see the muon. Both divide the same cell, by different sums.

Now the two directions are two ways of dividing the same cell. Prediction fixes a column and divides by the column sum: $\mathrm{P}(E \mid H_\mu) = 0.175/0.25$. Inference fixes the row we observed and divides by the row sum: $\mathrm{P}(H_\mu \mid E) = 0.175/0.235$. Observing E means *restricting attention to one row of the table* [Kruschke §5.2]. The other row, $\bar{E}$, is still a perfectly good row. It is simply not what happened, and it drops out of every calculation. That is the shrinking of the sample space in its most compact form. Section 5.2 draws it to scale.

The table also shows why the forward numbers are called a **likelihood** rather than a probability once we read them the inference way [Lambert §4.4]. Along a column, the forward probabilities of all possible observations sum to one: $0.70 + 0.30 = 1$. Along the observed row, the forward probabilities $\mathrm{P}(E \mid H_c)$ of that *one* observation under the different hypotheses are 0.70, 0.05, 0.20. They sum to 0.95, and nothing forces them to sum to one. Read along a row they are a function of the hypothesis, not a distribution over it.

Example: a coin table. Lambert makes the same point with a coin allowed only six values of its probability of heads, $\theta \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$, tossed twice [Lambert §4.4, Table 4.1]. The number of heads X is binomial, so the table entries are $\mathrm{P}(X \mid \theta) = \binom{2}{X}\theta^X(1-\theta)^{2-X}$:

θ	$X=0$	$X=1$	$X=2$	sum over X
0.0	1.00	0.00	0.00	1
0.2	0.64	0.32	0.04	1
0.4	0.36	0.48	0.16	1
0.6	0.16	0.48	0.36	1
0.8	0.04	0.32	0.64	1
1.0	0.00	0.00	1.00	1
sum over θ	2.20	1.60	2.20	

Each row (fixed θ) is a probability distribution for the data: it sums to one. Each column (fixed data) is a likelihood: for $X=1$ it sums to 1.60. The same happens for a continuous $\theta \in [0,1]$. With one head in two tosses,

$$\int_0^1 \mathrm{P}(X=1 \mid \theta)\,\mathrm{d}\theta \overset{(1)}{=} \int_0^1 2\theta(1-\theta)\,\mathrm{d}\theta \overset{(2)}{=} 2\Big(\frac{1}{2}-\frac{1}{3}\Big) = \frac{1}{3} \neq 1.$$

(1) The binomial probability of one head in two tosses.

(2) $\int_0^1 \theta\,\mathrm{d}\theta = \frac{1}{2}$ and $\int_0^1 \theta^2\,\mathrm{d}\theta = \frac{1}{3}$.

A likelihood has no reason to integrate to one over the parameter. That is why Bayes' theorem needs both a prior, to turn it into a joint probability, and a denominator, to normalise the row.

Watch out: A likelihood is not a probability of the hypothesis

$\mathrm{P}(E \mid H)$ read as a function of H is a likelihood. It can be large for two hypotheses at once, and its sum over hypotheses means nothing. "The muon path explains the data with probability 0.70" is a misreading: 0.70 is the probability of the data if the muon path fired. The probability of the path is 0.745 here, and only because of the particular priors.

5.1.4 Logical, not causal, connections

Inference runs backwards in logic, not in time. A later event cannot cause an earlier one, but it can still tell us about it. Jaynes' example shows this [Sivia §1.4].

Example: the second ball tells us about the first. A bag holds five red and seven green balls. We draw two, one after the other, without putting the first back.

The event. We are not told the colour of the first ball. We are told that the second is green. Call this G_2.

The question. Is the probability that the first ball was red still 5/12, as it was before we heard about the second? The first draw happened before the second, so the second cannot have caused it. But the second can be evidence about it.

The candidates. There are two: the first ball was red (R_1) or green (G_1). Under each, the second draw is a draw from a different bag. After a red first ball the bag holds 4 red and 7 green, so a green second ball has probability 7/11. After a green first ball it holds 5 red and 6 green, so a green second ball has probability 6/11. A green second ball is a little more likely if the first was red, because the first draw then left all seven greens in the bag.

Only then the numbers. The priors are $\mathrm{P}(R_1) = 5/12$ and $\mathrm{P}(G_1) = 7/12$:

$$\mathrm{P}(G_2) \overset{(1)}{=} \mathrm{P}(G_2 \mid R_1)\,\mathrm{P}(R_1) + \mathrm{P}(G_2 \mid G_1)\,\mathrm{P}(G_1) = \frac{7}{11}\cdot\frac{5}{12} + \frac{6}{11}\cdot\frac{7}{12} = \frac{35+42}{132} = \frac{7}{12},$$

$$\mathrm{P}(R_1 \mid G_2) \overset{(2)}{=} \frac{\mathrm{P}(G_2 \mid R_1)\,\mathrm{P}(R_1)}{\mathrm{P}(G_2)} = \frac{35/132}{77/132} = \frac{5}{11}.$$

(1) Marginalisation over the first draw, with $P(G_2 \mid R_1) = 7/11$ and $P(G_2 \mid G_1) = 6/11$ found above.
(2) Bayes' theorem.

The answer moved from $5/12 \approx 0.417$ to $5/11 \approx 0.455$. The extreme case makes it obvious: with one red and one green ball in the bag, a green second ball means the first was red with certainty. A conditional probability records what one statement *tells us* about another, not what causes what. Inference from the leaves of a tree back to its first split works the same way: the observation comes later in the physics, but it tells us about what happened first.

5.1.5 What the probabilities mean

There are two readings of what a probability is, and Wasserman sets them side by side [Wasserman pp. 175–176]. The frequentist reading rests on three postulates. Probability is a long-run relative frequency. Parameters are fixed constants, so no probability statements are made about them. A procedure is judged by how often it gives the right answer over many repetitions. The Bayesian reading rests on three others. Probability describes a degree of belief. We may therefore make probability statements about parameters. Inference about a parameter means producing a probability distribution for it. The second reading is the one that lets us give a probability to a one-off event. The weather forecast's "90% chance of rain tomorrow" is such a probability [SeeingTheory p. 49]. So is Laplace's posterior for the mass of Saturn. He summarised it as a bet of 11 000 to 1 that his estimate was not wrong by more than 1%. A further 150 years of data changed the estimate by only 0.63% [Sivia §1.4].

For the trigger path both readings work. Which path fires is the outcome of a process that repeats with every event, and 0.745 is the fraction of recorded events with an isolated muon that were fired by the muon path. A physical constant is different: nature chose its value once, so only the degree-of-belief reading is available. The arithmetic of (5.5) is the same either way. A degree of belief is not arbitrary: two people with the same background information I must assign the same probabilities [Sivia §1.4]. When they disagree, they disagree about I, and Bayes' theorem forces each of them to write I down.

5.2 Information shrinks the sample space

A disease affects one person in a hundred. A test for it comes out positive for 95% of the people who have the disease and, falsely, for 10% of the people who do not. Suppose a person picked at random tests positive. The 95% sensitivity tells us how often the test detects the disease in someone who has it. To find the probability that this person has the disease, we must also count the false positives among the much larger group of healthy people. This gave about 9% in Chapter 1. Here we represent the same calculation as an area so that we can see how each group contributes.

The picture rests on one idea. Before the evidence, every possibility in the sample space is still in play, and the probability of a hypothesis is the fraction of the whole space it occupies. The evidence tells us that we are somewhere in the region where the evidence is true. Everything outside that region is ruled out. The sample space has shrunk, and each hypothesis is now measured by the fraction of the *surviving* region that it occupies. Bayes' theorem is nothing more than this re-measurement. What makes it a useful calculating tool is that the surviving region can be built out of numbers we know: priors and forward probabilities.

5.2.1 The unit square

Take a square of side one. Its area, one, is the total probability. Cut it by a vertical line into a strip of width $\mathrm{P}(H)$ on the left, the hypothesis, and a strip of width $\mathrm{P}(\neg H) = 1 - \mathrm{P}(H)$ on the right, its negation. Inside the left strip, shade from the bottom a bar of height $\mathrm{P}(E \mid H)$: the fraction of the H-world in which the evidence E occurs. Inside the right strip, shade a bar of height $\mathrm{P}(E \mid \neg H)$. That is the whole construction (Figure 5.4, top left).

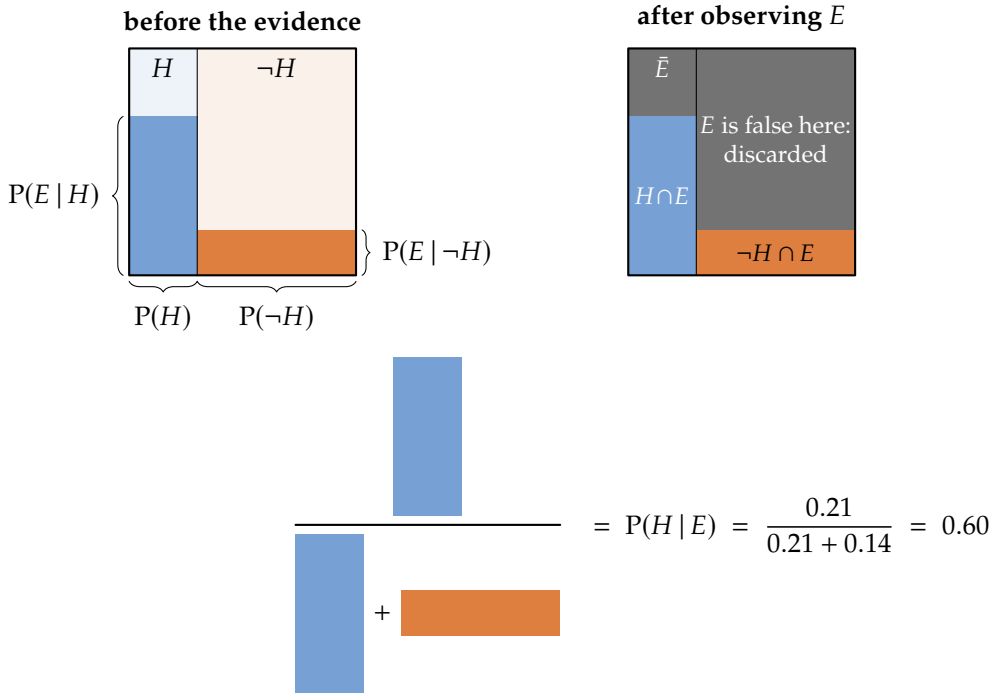


Figure 5.4: Bayes' theorem as information shrinking the sample space, for $\mathrm{P}(H) = 0.3$, $\mathrm{P}(E \mid H) = 0.7$, $\mathrm{P}(E \mid \neg H) = 0.2$. Top left: the square is split at the prior, and each strip carries a bar whose height is the probability that its hypothesis produces the evidence. Top right: once E is observed, everything above the bars is greyed out; only the two bars survive. Bottom: the posterior is the blue rectangle divided by the sum of the two surviving rectangles, drawn at the same scale as the square.

Every number in the problem now has a shape. A prior is a *width*. A forward probability, the likelihood of the evidence under one hypothesis, is a *height* measured inside that hypothesis's strip. The joint probability of "H and E" is the *area* of the blue bar, because the product rule says

$$\text{area of the blue bar} \overset{(1)}{=} \mathrm{P}(H) \times \mathrm{P}(E \mid H) \overset{(2)}{=} \mathrm{P}(H \cap E). \tag{5.6}$$

(1) A rectangle's area is its width times its height.
(2) Product rule (5.2).

The orange bar, likewise, has area $\mathrm{P}(\neg H \cap E)$. The two bars are disjoint and together they are all of E, so their total area is $\mathrm{P}(E)$.

Now observe E. Every point of the square above the bars is a possibility in which E did *not* happen, and we grey it out (Figure 5.4, top right). What survives is the two bars. The probability of H is now the fraction of the surviving area that lies in the H strip:

$$\begin{aligned} \mathrm{P}(H \mid E) &\overset{(1)}{=} \frac{\text{area}(H \cap E)}{\text{area}(E)} \overset{(2)}{=} \frac{\text{area}(H \cap E)}{\text{area}(H \cap E) + \text{area}(\neg H \cap E)} \\ &\overset{(3)}{=} \frac{\mathrm{P}(E \mid H)\ \mathrm{P}(H)}{\mathrm{P}(E \mid H)\ \mathrm{P}(H) + \mathrm{P}(E \mid \neg H)\ \mathrm{P}(\neg H)}. \end{aligned} \tag{5.7}$$

(1) Definition of conditional probability: E is the new, shrunk sample space, and $\mathrm{P}(\cdot \mid E)$ measures areas relative to it.
(2) The surviving region E is the union of the two disjoint bars.
(3) Each bar's area is width times height, (5.6).

The last line is Bayes' theorem (5.5) for the two hypotheses H and $\neg H$. The picture has derived it with no algebra beyond "area is width times height".

The same overlap also gives the forward probability. Measured against the H strip instead of against E,

$$\mathrm{P}(E \mid H) = \frac{\text{area}(H \cap E)}{\text{area}(H)}.$$

The numerator is the same blue bar in both conditional probabilities. What changes is the region we divide by: the whole H strip for prediction, the surviving region E for inference.

Once the square is drawn, Bayes' theorem is hard to avoid. There is a single overlap, the blue bar $H \cap E$, and two places to stand when we measure it. Standing in H, we ask how much of the H strip the evidence also covers, and the answer is $\mathrm{P}(E \mid H)$. Standing in E, we ask how much of the surviving region belongs to H, and the answer is $\mathrm{P}(H \mid E)$. Bayes' theorem is the algebra that turns the first measurement of this one overlap into the second.

Working definition 5.1 (prior, likelihood, posterior, evidence)

For hypotheses H_i that partition the sample space and an observation E:

- the **prior** $\mathrm{P}(H_i)$ [how plausible H_i was before E] is the width of strip i;
- the **likelihood** $\mathrm{P}(E \mid H_i)$ [how probable the observed E would be if H_i were true, read as a function of i] is the height of the bar in strip i (in the box picture of Figure 1.11, the share of strip i that E covers);
- the **evidence** $\mathrm{P}(E) = \sum_j \mathrm{P}(E \mid H_j)\,\mathrm{P}(H_j)$ [the total probability of the observation] is the total surviving area;
- the **posterior** $\mathrm{P}(H_i \mid E)$ [how plausible H_i is after E] is the area of E inside strip i divided by the total surviving area.

Operationally: multiply each prior by its likelihood, add the products, divide each product by the sum.

5.2.2 The medical test, to scale

Now draw the medical test with its true proportions. The disease strip has width $\mathrm{P}(D) = 0.01$; on a rectangle 5 cm wide that is half a millimetre. On the square of Figure 5.4 its bar would be a needle of that width, too thin to see. So we use the box picture of Figure 1.11 instead. It shows the same areas, but draws the evidence as one region, the box (Figure 5.5). A positive test covers 0.95 of the thin disease strip, because the test catches the disease 95% of the time, and 0.10 of the healthy strip, which is 0.99 wide. The two pieces sit side by side and make one box, the event +. The reason for the surprising answer is visible at once: the box is almost all healthy piece. That piece covers only a tenth of its strip, but its strip is ninety-nine times wider.

The event. What we observed is a positive test, written +.

What could have produced it? The one thing we do not know about the person is whether they are ill, so there are two candidate explanations: D, they have the disease, and D^c, they do not. Each is a route by which a positive result can arise.

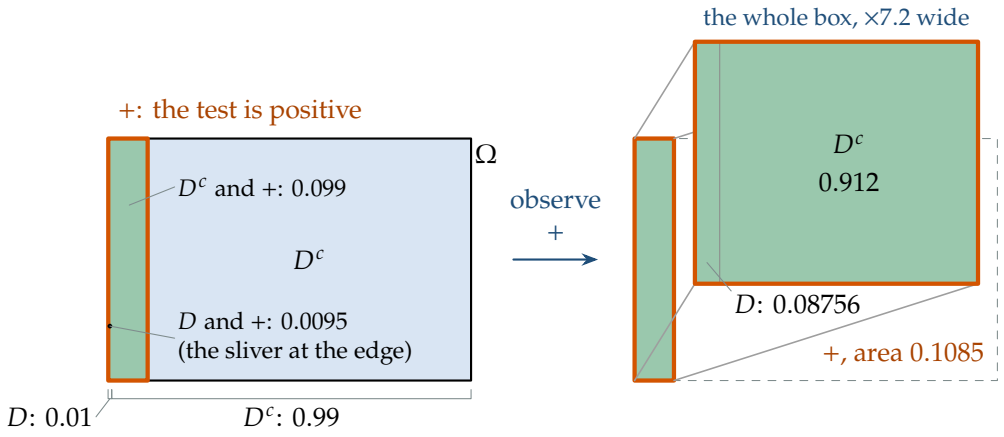


Figure 5.5: The medical test drawn to scale, in the picture of Figure 1.11. Left: Ω has area 1, cut into the disease strip D, of width 0.01, and the healthy strip D^c, of width 0.99. The event + (orange outline) covers 0.95 of the D strip, because the test catches the disease 95% of the time, and 0.10 of the D^c strip, because it gives a false alarm to one healthy person in ten. Its two pieces have areas $0.01 \times 0.95 = 0.0095$ and $0.99 \times 0.10 = 0.099$; at this scale the D piece hides under the orange outline. Right: a positive result discards everything outside the box. The raised inset magnifies the whole box about seven times horizontally, with the height unchanged (the grey lines join each corner of the box to the matching corner of the inset): the D piece is the narrow piece at the start of the box, and it fills only 0.08756 of the box's area.

How probable was the event along each route? If the person is ill, the test catches the disease 95% of the time, so $\mathrm{P}(+ \mid D) = 0.95$. If they are healthy, a positive result is a false alarm, which the test gives 10% of the time, so $\mathrm{P}(+ \mid D^c) = 0.10$. These are the shares of the two strips that the box covers.

Only then the priors, and the normalisation. Before the test, the person was ill with probability $\mathrm{P}(D) = 0.01$ and healthy with $\mathrm{P}(D^c) = 0.99$: the widths of the strips. The numbers are the areas:

$$\begin{aligned} \mathrm{P}(+) &\overset{(1)}{=} \mathrm{P}(+ \mid D)\,\mathrm{P}(D) + \mathrm{P}(+ \mid D^c)\,\mathrm{P}(D^c) = 0.95 \times 0.01 + 0.10 \times 0.99 \\ &= 0.0095 + 0.099 = 0.1085, \\ \mathrm{P}(D \mid +) &\overset{(2)}{=} \frac{0.0095}{0.1085} = 0.08756. \end{aligned}$$

(1) The area of the box: the sum of its two pieces.
(2) The D piece's share of it, (5.7).

The arithmetic is not the lesson. A positive result can come by two routes: through the disease, or through a false alarm on a healthy person. The disease route begins with little weight, 0.01, and the test favours it over the healthy route by a factor of only about 9.5 (0.95 against 0.10). A factor of 9.5 applied to so small a start cannot lift the posterior anywhere near 95%. In the box, the positive result covers 9.5 times more of the disease strip than of the healthy strip, but the healthy strip is 99 times wider.

Other versions of the same problem teach the same lesson: prevalence 0.001 with sensitivity 0.99 and false-alarm rate 0.05 gives 0.019 [Kruschke §5.1.2, Table 5.4]; prevalence 0.001 with 95% accuracy either way also gives 0.019, a twentyfold increase on the prior that is still far from certainty [SeeingTheory pp. 50–51]; breast-cancer screening with prevalence 1%, sensitivity 80% and false-positive rate 10% gives 0.075 [Lambert §3.6.2]. In every case the healthy piece covers a small share of a very wide strip.

Watch out: A share of the strip is not a share of the box

The 95% that people quote is the share of the disease strip that a positive result covers, $P(+ \mid D)$. The answer is the disease piece's share of the box, $P(D \mid +)$. The first tells us how the test behaves on diseased people; only the widths of both strips turn it into a statement about this person.

5.2.3 More than two hypotheses: Monty Hall

Nothing in the construction needs two strips. With k hypotheses the sample space is cut into k strips of widths $P(H_i)$, the evidence covers the share $P(E \mid H_i)$ of strip i, and the posterior of H_i is the share of the evidence that lies in strip i.

Example: Monty Hall, to scale. Three doors hide one car and two goats. We choose door A. Monty, who knows where the car is, always opens one of the two other doors to show a goat, choosing at random when he has a choice. Then he opens door C.

The event. What we observed is "Monty opened door C". Call it M_C.

What made him open that door? Before any number, we ask which situations could have produced this event. The only thing we do not know is where the car is, so the possible explanations are the three possible locations of the car. We call them H_A, H_B and H_C, with the car behind A, B or C respectively. The question is which of these locations could have led to the event we saw, Monty opening C.

How probable was the event under each hypothesis? We go through the three locations one by one and ask how likely Monty was to open C if that location were the truth. Monty knows where the car is, he never opens the door we chose, and he never reveals the car.

- Under H_A the car is behind our door. Both B and C hide goats, so he may open either, and he picks at random: $P(M_C \mid H_A) = \frac{1}{2}$.
- Under H_B the car is behind B. He cannot open A (ours) or B (the car), so he is forced to open C: $P(M_C \mid H_B) = 1$.
- Under H_C the car is behind C. He never reveals the car, so he would not open C: $P(M_C \mid H_C) = 0$.

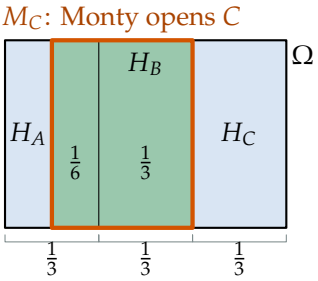
These three numbers are the shares of the three strips covered by the box in Figure 5.6 (the picture of Figure 1.22, here with doors A, B, C). They are also the three branches ending in "Monty opens C" of the tree in Chapter 1.

Only then the priors, and the normalisation. Before Monty acted, each location was equally probable, $P(H_A) = P(H_B) = P(H_C) = \frac{1}{3}$: these are the widths of the three strips. Multiplying each width by the share the box covers gives the area of each piece of the box: half of the A strip, all of the B strip and none of the C strip. Dividing each piece by the whole box,

$$P(H_B \mid M_C) \overset{(1)}{=} \frac{P(M_C \mid H_B)\,P(H_B)}{P(M_C \mid H_A)\,P(H_A) + P(M_C \mid H_B)\,P(H_B) + P(M_C \mid H_C)\,P(H_C)}$$
$$\overset{(2)}{=} \frac{1 \cdot \frac{1}{3}}{\frac{1}{2} \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3}} = \frac{1/3}{1/2} = \frac{2}{3}, \qquad P(H_A \mid M_C) = \frac{1/6}{1/2} = \frac{1}{3}, \qquad P(H_C \mid M_C) = 0.$$

(1) Bayes' theorem (5.5): the B piece over the whole box, the denominator being the sum over the three paths that can end in "Monty opens C".
(2) The shares found above and the widths $\frac{1}{3}$. The area of the box is 0.5.

Figure 5.6 draws the last line. Bayes' theorem is a division of areas: the piece of the box that lies in the B strip, divided by the whole box, which is the A piece and the B piece side by side.



priors are the strip widths; the box covers half of A, all of B, none of C

$$P(H_B \mid M_C) = \frac{B}{A + B} = \frac{1/3}{1/6 + 1/3} = \frac{2}{3}$$

Bayes: the B piece over the whole box

Figure 5.6: Monty Hall drawn to scale, and Bayes' theorem as a division of areas. Left: the sample space Ω has area 1. The three places the car can be fill it in strips of width $\frac{1}{3}$, the priors. The box M_C, "Monty opens door C", covers half of the A strip (with the car behind our door he picks B or C at random), all of the B strip (he is forced to open C) and none of the C strip (he never reveals the car); each green piece has area prior × likelihood. Right: the posterior of H_B is the B piece divided by the whole box, the A piece and the B piece together. All three pieces are shrunk by the same factor, so the ratio of areas you see is the posterior, $\frac{2}{3}$. The posterior of H_A is the A piece over the same box, $\frac{1}{3}$.

Why not fifty–fifty? The tempting argument says that Monty has taken one door away, so the two that remain must be equally likely. It ignores what Monty's choice tells us. Opening C was not equally probable under the two hypotheses still standing: if the car is behind B he had to open C, while if it is behind A he opened C only half the time. The evidence is twice as probable under H_B as under H_A, and with equal priors that factor of two passes straight into the posteriors. In the box the factor of two is visible directly: the B piece is twice as wide as the A piece, cut from strips of the same width.

The two-strip picture still works if we lump hypotheses together. Take $H = H_B$ and $\neg H =$ "H_A or H_C" (the car is behind A or behind C). Then $P(H) = \frac{1}{3}$, $P(M_C \mid H) = 1$, and the share of the $\neg H$ strip that the evidence covers is the average of the shares it replaces, weighted by their widths:

$$P(M_C \mid \neg H) \overset{(1)}{=} \frac{P(M_C \cap H_A) + P(M_C \cap H_C)}{P(H_A) + P(H_C)} = \frac{\frac{1}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot 0}{\frac{2}{3}} = \frac{1}{4},$$

(1) Definition of conditional probability, with the event $\neg H = H_A \cup H_C$ split into its disjoint parts.

and (5.7) gives $\frac{1/3}{1/3+(2/3)(1/4)} = \frac{2}{3}$ again. Merging strips merges their pieces into one piece of the same total area.

Example: two urns. The urn problem is the simplest picture of all. Urn A holds 8 red and 2 blue balls, urn B holds 2 red and 8 blue, and one urn is picked with probability $\frac{1}{2}$ each. A red ball is drawn.

The event is the red ball, R. What could have produced it? Either urn: the two candidate sources are A and B. How probable was a red ball from each? Urn A is eight-tenths red, so $P(R \mid A) = 0.8$; urn B is two-tenths red, so $P(R \mid B) = 0.2$. These are the shares of the two strips that the evidence covers. Only then the priors: each urn was picked with probability $\frac{1}{2}$, so the

strips are equal, and normalising the surviving areas gives $\mathrm{P}(A \mid R) = 0.4/(0.4 + 0.1) = 0.8$ and $\mathrm{P}(B \mid R) = 0.2$.

Drawing the red ball did nothing to the urns: each still holds what it held. What changed is our belief about which urn the ball came from, and that belief is fixed by two things together, the probability with which each urn was picked and the probability that each urn gives a red ball.

5.2.4 A physics example: signal or background after the trigger

Before the trigger, 5% of the collisions entering an analysis preselection are signal (S) and 95% are background (B). The trigger keeps 90% of signal events and 2% of background events.

The event. An event has been triggered.

The question. Is it signal? There are two candidate explanations, S and B. Under S the trigger fires with probability 0.90; under B with probability 0.02.

The numbers. Drawn to scale (Figure 5.7), the S strip has width 0.05 and the trigger keeps 0.90 of it; the B strip has width 0.95 and the trigger keeps only 0.02 of it. The two pieces of the triggered box have areas 0.045 for signal and 0.019 for background, so the triggered sample has signal fraction

$$\mathrm{P}(S \mid \mathrm{trig}) \overset{(1)}{=} \frac{0.05 \times 0.90}{0.05 \times 0.90 + 0.95 \times 0.02} = \frac{0.045}{0.064} = 0.7031.$$

(1) Bayes' theorem (5.7); 0.064 is the fraction of all preselected events that the trigger keeps.

The trigger has raised the signal fraction from 5% to 70%. In detector language, $\mathrm{P}(\mathrm{trig} \mid S)$ is the *efficiency* and $\mathrm{P}(S \mid \mathrm{trig})$ is the *purity* of the triggered sample. The picture shows the difference. The efficiency is a share of the S strip, so it does not depend on the strip widths. The purity is a share of the box, so it does.

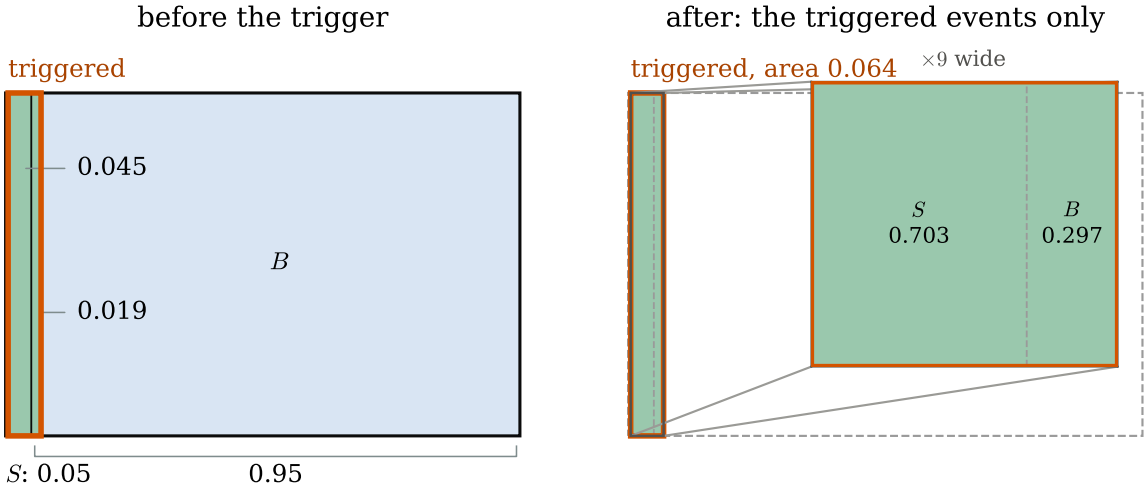


Figure 5.7: The trigger drawn to scale, in the picture of Figure 1.11. Left: the signal strip S has width 0.05 and the background strip B width 0.95; the triggered events (orange outline) cover 0.90 of the signal strip but only 0.02 of the background strip, so the two pieces have areas 0.045 and 0.019. The background piece covers a small share of its strip, but its strip is nineteen times wider. Right: once we keep only triggered events, the box is the new sample space; the inset magnifies it horizontally. Signal fills 0.7031 of it.

5.2.5 Drawing the boxes in Python

The trigger figure was drawn by one short function, `box_diagram`, which takes any list of priors and likelihoods and draws the box picture to scale. Its geometry is the loop over strips:

Listing 5.1: The strips and the pieces of the evidence (`code/ch05/lib_area.py`)

```
143     priors, likes = np.asarray(priors, float), np.asarray(likes, float)
144     k = len(priors)
145     edges = W_OMEGA * np.concatenate([[0.0], np.cumsum(priors)])
146     pieces = []
147     for i in range(k):
148         x0, x1 = edges[i], edges[i + 1]
149         w = (x1 - x0) * likes[i]
150         left_has = i > 0 and likes[i - 1] > 0
151         right_has = i < k - 1 and likes[i + 1] > 0
152         if right_has and not left_has:     # lean right, towards the next piece
153             pieces.append((x1 - w, x1))
154         else:                              # lean left (or fill the strip if likes[i] = 1)
155             pieces.append((x0, x0 + w))
156     return edges, pieces
```

Line 145 turns the priors into the positions of the strip edges, so strip i runs from the sum of the earlier priors to that sum plus its own prior. Line 149 is the whole of the likelihood's role: the piece of E in strip i is as wide as the strip times $\mathrm{P}(E \mid H_i)$, and it runs the full height, so its area is prior times likelihood. Lines 150–155 only decide where the piece sits inside its strip: it is pushed against a neighbouring strip that also has a piece, so that the pieces join into one box, as in Figure 1.22. The companion function `posterior` in the same file computes the numbers: the joint areas are `priors * likes`, their sum is $\mathrm{P}(E)$, and the posterior is their ratio. The same file keeps `area_diagram`, which draws the square of Figure 5.4 with bars. The script `code/ch05/01_area_diagrams.py` computes the medical test, Monty Hall and the trigger, draws the trigger, and writes the posteriors quoted above.

5.3 Reallocating credibility across possibilities

We step outside in the morning and the pavement is wet. Rain? A garden sprinkler? A burst pipe? A passer-by who spilled a drink? Each possibility has some credibility before we look further, rain more than a spilled drink. Then we look around. If the trees and the parked cars are wet as far as we can see, credibility flows to rain. If the wet patch is small and there is an empty cup beside it, credibility flows to the spilled drink, although it started low. Kruschke puts the whole of Bayesian inference into one sentence: it is the **reallocation of credibility across possibilities** [Kruschke §2.1].

The phrase says two things, and the square of Section 5.2 makes both exact. First, the total credibility is fixed. It is always one, shared among possibilities that exclude each other and together cover everything we consider. Second, credibility moves towards the possibilities that account well for what was seen. Which ones gain? The posterior of a possibility is its prior times its likelihood, divided by $\mathrm{P}(E)$. And $\mathrm{P}(E)$ is the average of all the likelihoods, each weighted by its prior. So a possibility gains exactly when its likelihood, the height of its bar in the square, is above that average. This means any inference problem needs only three things: a list of possibilities, a prior over the list, and, for each item, the probability of what we saw.

Key idea: Posterior is proportional to likelihood times prior

$$\mathrm{P}(H_i \mid E) \;\propto\; \mathrm{P}(E \mid H_i)\,\mathrm{P}(H_i),$$

with the constant fixed by $\sum_i \mathrm{P}(H_i \mid E) = 1$ [Wasserman p. 177, (11.2)]. Equivalently, the posterior of H_i is its prior multiplied by the factor $\mathrm{P}(E \mid H_i)/\mathrm{P}(E)$, which exceeds one exactly when H_i predicted E better than the prior-weighted average of all hypotheses.

5.3.1 The Bayes box

The bookkeeping fits in a table with one row per possibility, which Lambert calls a **Bayes box** [Lambert §5.5.1]. Columns: the possibility, its prior, its likelihood, their product, and the product divided by the column sum.

Example: the fish game. A covered bowl holds five fish, each red or white. We want the number α of red fish. We pull out one fish at random and it is red: that is the event. The candidates are the six possible bowls, $\alpha \in \{0, 1, \ldots, 5\}$. We have no idea beforehand, so each gets prior $\frac{1}{6}$. How probable is a red fish under each candidate? A bowl with α red fish out of five gives a red fish with probability $\alpha/5$.

α	prior	likelihood $\alpha/5$	prior × likelihood	posterior
0	1/6	0	0	0
1	1/6	1/5	1/30	1/15
2	1/6	2/5	2/30	2/15
3	1/6	3/5	3/30	3/15
4	1/6	4/5	4/30	4/15
5	1/6	1	5/30	5/15
sum	1	3	15/30 = P(red)	1

The posterior column is the product column divided by its sum, $\frac{1}{2}$. The table shows three things. First, the likelihood column sums to 3, not to 1: as in Section 5.1, a likelihood is not a distribution over the possibilities. Second, $\alpha = 0$ is eliminated, because a bowl with no red fish cannot produce a red fish. Third, with a flat prior the posterior has the same shape as the likelihood, here a straight line rising to $\alpha = 5$. Suppose instead we believed that the game-maker likes mixed bowls. A prior peaked at $\alpha = 2, 3$ would then pull the posterior back towards the middle. The posterior is a compromise between prior and likelihood [Lambert §5.5.1], [Kruschke §5.3].

5.3.2 Elimination and exoneration

Sherlock Holmes said: "when you have eliminated the impossible, whatever remains, however improbable, must be the truth". In the Bayes box this is a likelihood of zero. If the observation is impossible under H_j, then $\mathrm{P}(E \mid H_j) = 0$, and that row's posterior is zero. If the other possibilities all have the same likelihood, their posteriors are just their priors rescaled to sum to one. Kruschke draws four suspects at 0.25 each: rule out A, and B, C, D get $\frac{1}{3}$ each; rule out B as well, and C, D get $\frac{1}{2}$; rule out C, and D gets 1 [Kruschke §2.1, Fig. 2.1]. The reverse, *exoneration*, is an observation that only one possibility can produce: then that possibility gets credibility one and all the others are cleared [Kruschke Fig. 2.2].

Example: the trigger path, continued. Recall the event of Section 5.1 that had lost its trigger flag. One of three paths fired it: muon (H_μ), calorimeter (H_{cal}) or track (H_{trk}). The isolated

muon E moved their credibility from $(0.25, 0.60, 0.15)$ to $(0.745, 0.128, 0.128)$. Now we read the run log and find that during this run the muon trigger was masked: it could not fire. How probable is that log entry under each path? If the muon path fired the event, a mask is impossible, so the probability is 0. If either other path fired it, the mask is certain to be in the log, so the probability is 1. The new likelihoods multiply the old posteriors:

$$\mathrm{P}(H_\mu \mid E, \text{mask}) \overset{(1)}{=} \frac{0 \times 0.745}{0 \times 0.745 + 1 \times 0.128 + 1 \times 0.128} = 0,$$
$$\mathrm{P}(H_{\text{cal}} \mid E, \text{mask}) = \mathrm{P}(H_{\text{trk}} \mid E, \text{mask}) \overset{(2)}{=} \frac{1}{2}.$$

(1) Bayes' theorem with the previous posterior as the new prior (justified in Section 5.5); the mask entry has likelihoods $(0, 1, 1)$.

(2) The two survivors have equal likelihood, so they keep the ratio of their previous credibilities, which was $1 : 1$.

The muon path, the favourite a moment ago, is gone. Its 0.745 has been shared between the two survivors in proportion to what they had. Figure 5.8 shows the whole sequence as bars.

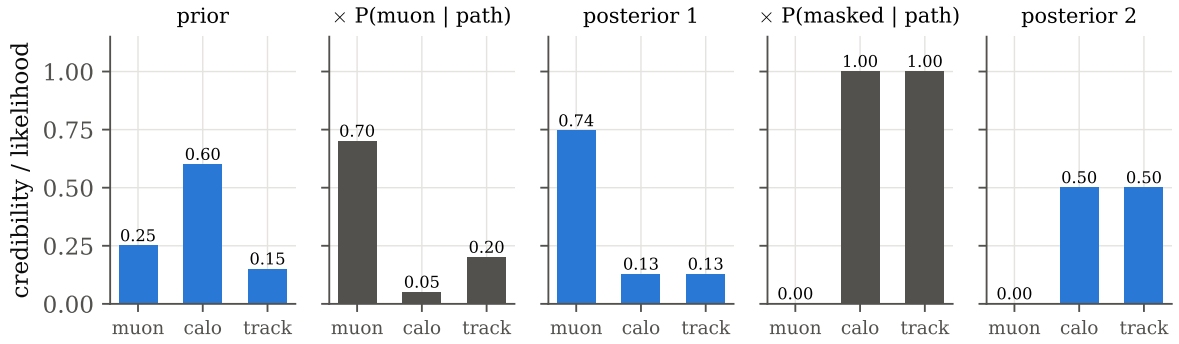


Figure 5.8: How two observations change the probabilities of the three trigger paths. Read from left to right, starting with the prior share of recorded events for each path. The isolated high-p_T muon supplies the first likelihood and the run-log entry "muon trigger masked" supplies the second. Each update multiplies the preceding blue probabilities by the corresponding grey likelihoods, then normalises the products so that the new blue bars sum to one.

Watch out: A zero prior can never be revived

If a possibility has prior zero, its posterior is zero whatever is observed, because $0 \times \mathrm{P}(E \mid H) = 0$. Lambert's example is someone so certain that the Earth is flat that the round Earth gets prior zero; no amount of astronomical data can then move it [Lambert §5.7.1]. Truncating a prior ("the prevalence is surely below 25%") does the same to every value beyond the cut. Give every possibility you cannot rule out on physical grounds a nonzero prior.

5.3.3 Noisy data only nudge

Holmes' clues were decisive: each one set some likelihoods exactly to zero. Real measurements are noisy. A noisy measurement is only somewhat more probable under one possibility than under another, so credibility moves gradually and nothing is ruled out.

Example: which ball size? A factory makes balls in four sizes, with nominal diameters $1, 2, 3, 4$ (in some unit). Inflation varies, so a ball of size s has a diameter spread around s; take

it Gaussian with standard deviation $\sigma = 1.17$.[1] We ordered three balls of size 2.

The event. We measure the three diameters $d = (1.77, 2.23, 2.70)$.

The question. Did we get what we ordered [Kruschke §2.1.1, Fig. 2.3]? Which size could have produced these three diameters?

The candidates. The four sizes $s = 1, 2, 3, 4$. The factory may have lost our order, so we do not trust the label and give each size prior $\frac{1}{4}$.

How probable are the data under each size? The three diameters are independent once the size is fixed, so the likelihood is a product of three Gaussian densities:

$$p(\boldsymbol{d} \mid s) \overset{(1)}{=} \prod_{i=1}^{3} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\!\left[-\frac{(d_i - s)^2}{2\sigma^2}\right] \overset{(2)}{=} \frac{1}{(2\pi\sigma^2)^{3/2}} \exp\!\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{3}(d_i - s)^2\right]. \tag{5.8}$$

(1) Independence given s: the joint density is the product of the three single-ball densities $\mathcal{N}(d_i; s, \sigma^2)$.

(2) Collect the three exponentials into one.

The prefactor is the same for every s, so it cancels in the posterior. Only the sum of squared distances from the data to s matters, and the smaller it is, the larger the likelihood. The sums are 5.00, 0.60, 2.20 and 9.80 for $s = 1, 2, 3, 4$: size 2 is closest to the three measurements.

Only then the numbers. With equal priors, the posterior (Figure 5.9) is 0.1118, 0.5578, 0.311, 0.01937 for sizes 1 to 4. Size 2 is the most credible, but far from certain. Its likelihood is only 1.794 times that of size 3, because the spread of a single ball, $\sigma = 1.17$, is larger than the distance between sizes. No size is ruled out, since a Gaussian is never exactly zero. Size 4 is merely very unlikely.

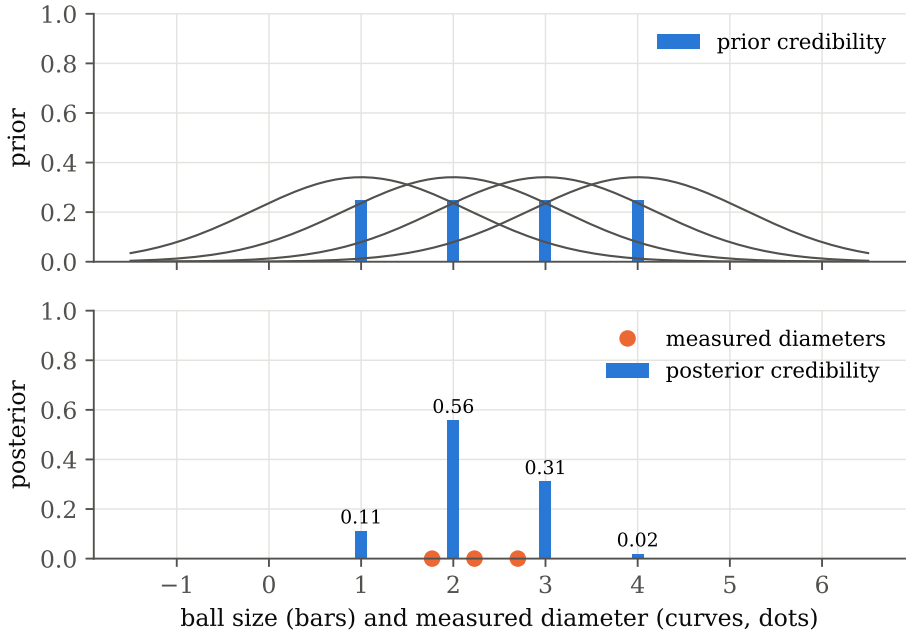


Figure 5.9: Top: four candidate ball sizes with equal prior credibility (bars), and the spread of diameters each size produces (curves). Bottom: the three measured diameters (dots on the axis) and the posterior credibility of each size. The data sit between sizes 2 and 3, closer to 2, and the credibility moves accordingly, without ruling anything out.

Data analysis is exactly this, with the balls replaced by physical parameters. We list the

[1] This value reproduces Kruschke's percentages; he does not state the spread he used.

possible sources of the data, write down how each would produce data, and let the data reallocate the credibility [Kruschke §2.1.1]. The recipe has the same steps in every textbook. Identify the data. Write a model that gives the probability of the data for each possibility. Put a prior on the possibilities. Apply Bayes' theorem. Check that the most credible possibilities actually reproduce the data [Kruschke §2.3], [Wasserman p. 176]. Repeat as new data arrive [SeeingTheory p. 51]. The possibilities are usually the values of the parameters of a model [Kruschke §2.2]. When a parameter is continuous, the row of bars becomes a density over it. Each value of the parameter, here the ball size s, is then one candidate explanation of the measured data, and its likelihood says how easily that value could have produced them; this is parameter estimation, taken up for continuous parameters in Section 5.7.

5.3.4 Which source produced this photon?

A gamma-ray burst goes off, and in the 100 s after it a gamma-ray telescope records photons from a $10^\circ \times 10^\circ$ patch of sky around it. Each photon comes with a reconstructed direction $\boldsymbol{r}$ and energy E. Three sources can produce a photon in this patch: the burst itself (G), a steady blazar 1.3° away (Bz), and the diffuse gamma-ray background spread over the whole patch (D). For each photon we ask: which source produced it? This is the trigger-path problem again. The candidates are the three sources, and the evidence is the photon's direction and energy, which are continuous.

Priors from rates. Before we look at a photon's direction or energy, the probability that it came from source k is that source's share of the expected photon count in the window. We take 30 expected burst photons, 8 from the blazar and 40 from the diffuse background, so the priors are 0.3846, 0.1026 and 0.5128. The diffuse background is the most probable source of a random photon.

Likelihoods from the instrument. The telescope does not report a direction exactly. A photon from a point source at $\boldsymbol{r}_k$ is reconstructed at $\boldsymbol{r}$ with a probability density given by the **point-spread function** [the density of reconstructed directions for a point source, here a 2-D Gaussian of width $\sigma(E)$], which is wide at low energy and narrow at high energy; we take $\sigma(E) = 0.8^\circ\,(E/1\,\mathrm{GeV})^{-0.8}$. The energy of a photon from source k is distributed according to that source's spectrum, a power law $p(E \mid k) \propto E^{-\gamma_k}$ with $\gamma = 2.0$ for the burst, 2.4 for the blazar and 2.6 for the diffuse emission. Direction and energy are independent given the source, apart from the energy dependence of the PSF width, so

$$p(\boldsymbol{r}, E \mid k) \overset{(1)}{=} p(E \mid k)\, p(\boldsymbol{r} \mid E, k) \overset{(2)}{=} \begin{cases} p(E \mid k)\, \dfrac{1}{2\pi\sigma(E)^2} \exp\!\left[-\dfrac{|\boldsymbol{r} - \boldsymbol{r}_k|^2}{2\sigma(E)^2}\right], & k = \mathrm{G}, \mathrm{Bz}, \\ p(E \mid \mathrm{D})\, \dfrac{1}{A}, & k = \mathrm{D}. \end{cases} \tag{5.9}$$

(1) Product rule for densities.

(2) For a point source the direction density is the PSF centred on the source. For the diffuse background every direction in the patch of area $A = 100\,\mathrm{deg}^2$ is equally likely.

Densities as likelihoods. The observation "the photon arrived exactly at $\boldsymbol{r}$ with exactly energy E" has probability zero under every source. What the telescope really reports is "in a small cell $\mathrm{d}^2\boldsymbol{r}\,\mathrm{d}E$ around $(\boldsymbol{r}, E)$", and that has probability $p(\boldsymbol{r}, E \mid k)\,\mathrm{d}^2\boldsymbol{r}\,\mathrm{d}E$. The cell size is the same factor for every source and cancels between the numerator and the denominator of Bayes' theorem. So densities can stand in for probabilities as likelihoods, as long as all of them are densities in the same variables.

Example: one photon. A photon arrives at $\boldsymbol{r} = (0.7°, 0.3°)$ with $E = 2\,\text{GeV}$, between the burst and the blazar and a little closer to the blazar. At this energy the PSF width is 0.4595°. The three likelihoods (5.9) are 0.004776, 0.004006 and 6.629×10^{-5} per deg^2 per GeV. Then

$$\mathrm{P}(\mathrm{G} \mid \boldsymbol{r}, E) \overset{(1)}{=} \frac{0.3846 \times 0.004776}{0.3846 \times 0.004776 + 0.1026 \times 0.004006 + 0.5128 \times 6.629 \times 10^{-5}} \overset{(2)}{=} 0.8051,$$

(1) Bayes' theorem with the three sources as the partition.
(2) Numerically, $1.84 \times 10^{-3} / 2.28 \times 10^{-3}$.

and likewise $\mathrm{P}(\mathrm{Bz} \mid \boldsymbol{r}, E) = 0.18$, $\mathrm{P}(\mathrm{D} \mid \boldsymbol{r}, E) = 0.0149$. The burst wins, and each source's result has a plain reason. The photon is closer to the blazar, so the direction alone slightly favours the blazar. But the burst is nearly four times brighter in this window, which gives it the larger prior, and its harder spectrum makes a 2 GeV photon more probable. The product of prior and likelihood is therefore largest for the burst. The diffuse background was the most probable source before we looked at the photon. It loses because its likelihood is spread evenly over the whole patch: it predicts *this* direction no better than any other.

Figure 5.10 does this for every photon of one simulated window. Photons near the burst are dark (almost surely burst photons), photons far from both point sources are pale (mostly background), and photons between the two point sources are shared. A soft photon far out, at $(-3°, 2°)$ with $E = 0.3\,\text{GeV}$, still has $\mathrm{P}(\mathrm{G} \mid \boldsymbol{r}, E) = 0.4088$: at 0.3 GeV the PSF is about 2° wide, so the burst can put a photon there.

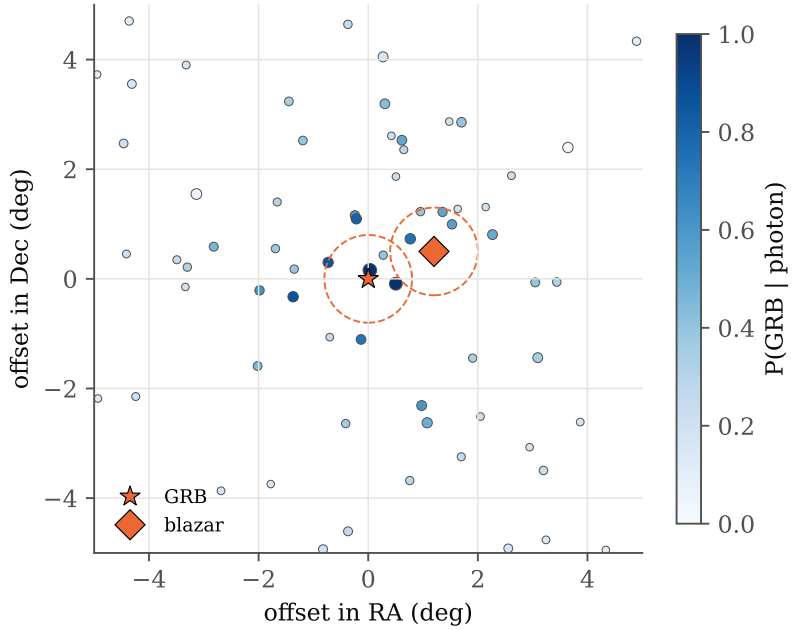


Figure 5.10: One simulated 100 s window of photons around a gamma-ray burst (star), with a blazar nearby (diamond) and a diffuse background. Each dot is a photon, larger for higher energy, shaded by its posterior probability of having come from the burst. The dashed circles have the radius of the point-spread function at 1 GeV. Dark dots cluster on the burst; photons far from both point sources are pale.

Listing 5.2: Posterior source probabilities for a photon (`code/ch05/03_grb_source.py`)

```
def likelihoods(r, E):
    """p(r, E | k) for each source k (columns in SRC order)."""
    return np.stack([p_energy(E, GAMMA["GRB"]) * psf(r, POS["GRB"], E),
                     p_energy(E, GAMMA["blazar"]) * psf(r, POS["blazar"], E),
                     p_energy(E, GAMMA["diffuse"]) / AREA], axis=-1)
```

```
61
62
63 def membership(r, E, use=(0, 1, 2)):
64     prior = np.array([MU[s] for s in SRC])[list(use)]
65     L = likelihoods(r, E)[..., list(use)]
66     w = prior * L
67     return w / w.sum(axis=-1, keepdims=True)
```

Lines 58–60 are the three cases of (5.9), one column per source. Line 66 multiplies each column by its prior (the expected counts; their common normalisation cancels), and line 67 divides by the sum over sources. It works on a whole array of photons at once. Adding the burst column over all photons of the window gives 23.65, an estimate of the number of burst photons; the simulation actually contained 25 of the 68. Such per-photon source probabilities are used in gamma-ray astronomy to weight photons when searching for pulsations or transients, so that background photons dilute the signal less.

5.4 Odds and the likelihood ratio: evidence as a multiplier

A search for a rare process uses a boosted decision tree, a classifier that gives every event a score x between 0 and 1. Signal events tend to score high, background events low, and both distributions are known from simulation. Before we look at its score, an event is signal with probability π, the signal fraction of the sample. Then one event scores $x = 0.9$. How probable is it that this event is signal? And how would the answer change in another analysis with the same classifier but a hundred times less signal?

The answer splits into two factors. One depends only on the classifier and the score. The other depends only on the sample. The second question then needs no new work: we keep the first factor and change the second. This split is the odds form of Bayes' theorem, and it gives the simplest working rule for the question "which hypothesis produced this?".

5.4.1 Odds

The **odds** on a hypothesis are its probability divided by the probability of its negation, $O(H) = \mathrm{P}(H)/\mathrm{P}(\neg H)$. Odds and probability carry the same information:

$$O = \frac{p}{1-p} \qquad \Longleftrightarrow \qquad p = \frac{O}{1+O}. \tag{5.10}$$

A probability of $\frac{1}{2}$ is odds 1 ("evens"), 0.9 is odds 9, 0.01 is odds 1/99. More generally, for any two hypotheses H_1, H_2 in a partition, the ratio $\mathrm{P}(H_1)/\mathrm{P}(H_2)$ is the odds of H_1 against H_2.

How do the odds of H_1 against H_2 change when we observe E? Write Bayes' theorem (5.5) for H_1 and for H_2 and divide:

$$\frac{\mathrm{P}(H_1 \mid E)}{\mathrm{P}(H_2 \mid E)} \overset{(1)}{=} \frac{\mathrm{P}(E \mid H_1)\ \mathrm{P}(H_1)/\mathrm{P}(E)}{\mathrm{P}(E \mid H_2)\ \mathrm{P}(H_2)/\mathrm{P}(E)} \overset{(2)}{=} \underbrace{\frac{\mathrm{P}(E \mid H_1)}{\mathrm{P}(E \mid H_2)}}_{\text{likelihood ratio}} \times \underbrace{\frac{\mathrm{P}(H_1)}{\mathrm{P}(H_2)}}_{\text{prior odds}}. \tag{5.11}$$

(1) Bayes' theorem in numerator and denominator.
(2) The evidence $\mathrm{P}(E)$ is common to both and cancels.

We write Λ for the likelihood ratio.

Key idea: Evidence multiplies odds

$$\text{posterior odds} = \Lambda \times \text{prior odds}, \qquad \Lambda = \frac{\mathrm{P}(E \mid H_1)}{\mathrm{P}(E \mid H_2)}.$$

The **likelihood ratio** Λ is everything the observation has to say about H_1 versus H_2. It does not involve the priors, it does not involve any other hypothesis, and it does not need $\mathrm{P}(E)$.

In the square of Section 5.2 every factor of (5.11) is a ratio of shapes. The prior odds are the ratio of the two strip *widths*. The likelihood ratio is the ratio of the two bar *heights*. Their product, width ratio times height ratio, is the ratio of the two surviving *areas*, which is the posterior odds. Take the medical test of Figure 5.5. The strip widths are $\mathrm{P}(D) = 0.01$ and $\mathrm{P}(D^c) = 0.99$, a ratio of $1 : 99$. The positive result covers 0.95 of the disease strip and 0.10 of the healthy one, a likelihood ratio of $9.5 : 1$. So the surviving areas are in the ratio $9.5 : 99 \approx 1 : 10.4$, and the posterior probability is $1/11.4 \approx 0.088$ by (5.10). The test is good evidence, a factor of 9.5, but it is applied to odds that started at one to ninety-nine.

5.4.2 Evidence adds in logarithms

Take logarithms of (5.11). Products become sums:

$$\log O(H_1{:}H_2 \mid E) = \log \Lambda + \log O(H_1{:}H_2). \tag{5.12}$$

So each piece of evidence *adds* its log likelihood ratio to the log odds. We call $\log \Lambda$ the *weight of evidence*. A convenient unit is the **decibel**: $10 \log_{10}$ of the odds, the same unit physicists use for the gains along an amplifier chain. Odds of $1 : 99$ are -20 dB, evens are 0 dB, odds of $99 : 1$ are $+20$ dB. A positive medical test adds $10 \log_{10} 9.5 = +9.8$ dB. One positive test takes the patient from -20.0 dB to -10.2 dB, which is odds 0.096 and probability 0.088. Suppose a second positive test is taken, independent of the first *given* the disease status (a condition examined in Section 5.5). It adds another 9.8 dB, to -0.4 dB, probability 0.48. A third takes it to $+9.4$ dB, probability 0.90 (Figure 5.11).

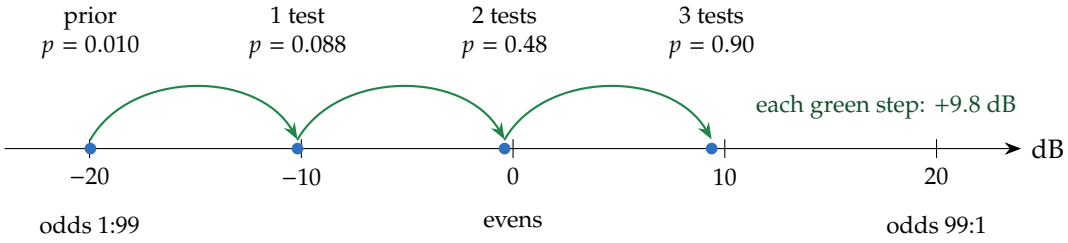


Figure 5.11: The medical test as a walk on a scale of log odds. The prior, one person in a hundred, sits at -20 dB. Each positive result from an independent test adds the same weight of evidence, $10 \log_{10} 9.5 \approx 9.8$ dB. The probability above each dot follows from the odds by $p = O/(1 + O)$: three positive tests are needed before the disease becomes the likely explanation.

The additive form explains two habits of practising physicists. First, analyses add log likelihoods rather than multiply likelihoods, because independent pieces of evidence add in logarithms. Second, "how strong is this evidence?" has an answer that does not depend on the prior: $\log \Lambda$. Whether it is strong *enough* depends on where the prior odds started. A $+10$ dB piece of evidence decides a question that started at evens, and barely dents one that started at -40 dB.

5.4.3 Signal or background, given a classifier score

Back to the boosted decision tree of the opening, which gives each event a score x between 0 and 1, and to the question: given one event's score, is it signal (S) or background (B)? Suppose the simulation shows that signal scores follow the density $f_S(x) = 30\,x^4(1-x)$ and background scores the density $f_B(x) = 30\,x(1-x)^4$ on $[0,1]$ (Figure 5.12, left). Both are Beta densities, Beta(5, 2) and Beta(2, 5); all we need is that each integrates to one, which the factor 30 ensures. The evidence is the score of one event, a continuous quantity, so the likelihoods are densities, and the cell $\mathrm{d}x$ cancels in the ratio exactly as the cell $\mathrm{d}^2\boldsymbol{r}\,\mathrm{d}E$ did for the photon in Section 5.3. The likelihood ratio is

$$\Lambda(x) \overset{(1)}{=} \frac{f_S(x)}{f_B(x)} \overset{(2)}{=} \frac{30\,x^4(1-x)}{30\,x(1-x)^4} \overset{(3)}{=} \left(\frac{x}{1-x}\right)^3. \tag{5.13}$$

(1) Likelihood ratio of signal to background for the observed score.
(2) The two densities.
(3) Cancel 30, one power of x and one power of $(1-x)$.

With prior odds $\pi/(1-\pi)$, the posterior odds are $\Lambda(x)\,\pi/(1-\pi)$. In logarithms, with the **logit** $\operatorname{logit} p = \ln[p/(1-p)]$ [the natural log of the odds],

$$\operatorname{logit} \mathrm{P}(S \mid x) = \operatorname{logit} \pi + 3 \operatorname{logit} x. \tag{5.14}$$

This is the split promised at the start. The score contributes a term, $3 \operatorname{logit} x$, that depends only on the classifier. The sample contributes a term, $\operatorname{logit} \pi$, that depends only on the signal fraction π.

Example: the same score in two analyses. For $x = 0.9$ the likelihood ratio is $9^3 = 729$. With $\pi = 0.01$ the prior odds are $1/99$, the posterior odds $729/99 = 7.36$, and

$$\mathrm{P}(S \mid x = 0.9) \overset{(1)}{=} \frac{7.36}{1 + 7.36} = 0.8804.$$

(1) Posterior odds converted to a probability by (5.10).

In the analysis with a hundred times less signal, $\pi = 10^{-4}$, the likelihood ratio is unchanged and only the prior odds change. The same score gives posterior odds $729/9999 = 0.073$, so $\mathrm{P}(S \mid x = 0.9) = 0.06795$. Same event, same classifier, and now it is probably background. A score of 0.8 is weaker evidence, $\Lambda = 4^3 = 64$, and gives only 0.3926 at $\pi = 0.01$. The score at which an event becomes more likely signal than not, $\mathrm{P}(S \mid x) = \frac{1}{2}$, is where $3 \operatorname{logit} x = -\operatorname{logit} \pi$: $x_{1/2} = 0.8223$ for $\pi = 0.01$ and 0.9556 for $\pi = 10^{-4}$ (Figure 5.12, right).

Checking it by selection. We can check the formula without algebra, because conditioning on an observation is the same as selecting the events that show it (Chapter 1). The script `code/ch05/04_bdt_posterior.py` generates 10^7 events with $\pi = 0.01$, draws each event's score from f_S or f_B, keeps the 2238 events with $|x - 0.9| < 0.005$, and counts the signal among them: the fraction is 0.882 ± 0.006818, against 0.8804 from (5.14). Bayes' theorem is the statement of what this selection returns.

5.4.4 Be exact about what you observed

A score and a cut are different observations, and they give different answers. The dotted curve in Figure 5.12 shows the cut. Suppose we do not look at each event's score but select all events

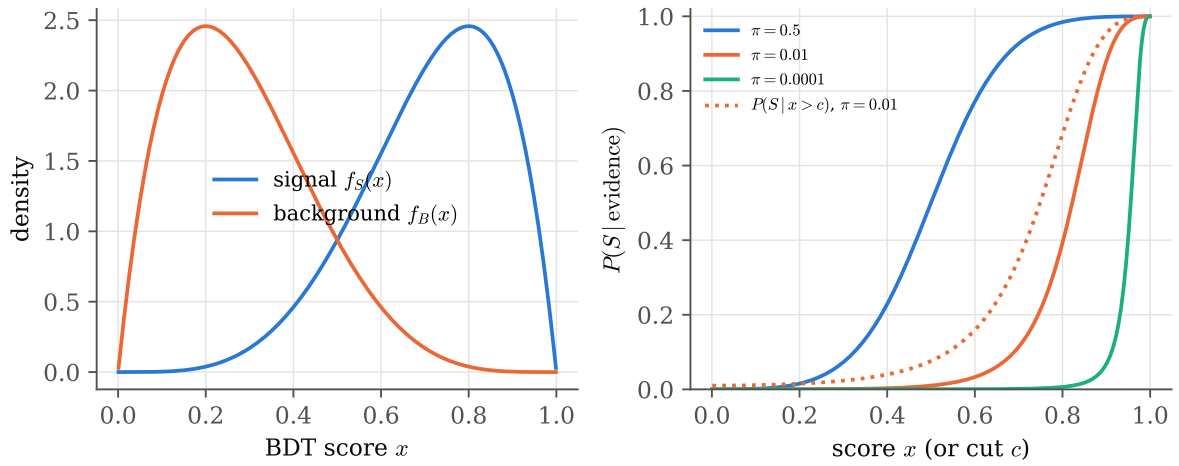


Figure 5.12: Left: the score densities of signal and background. Right: the probability that an event is signal given its score x, for three signal fractions π (solid). The curves are the same likelihood ratio combined with different prior odds, so they are shifted copies of one another on the logit scale. The dotted curve is a different question: the signal fraction of the whole sample with $x > c$, plotted against the cut c, for $\pi = 0.01$.

with $x > c$ and ask for the signal fraction of the selected sample, its **purity**. The observed event is now "$x > c$", whose probability under each hypothesis is an efficiency, $\varepsilon_S(c) = \int_c^1 f_S \, \mathrm{d}x$ and $\varepsilon_B(c) = \int_c^1 f_B \, \mathrm{d}x$. For $c = 0.8$ these are 0.3446 and 0.0016, so

$$\Lambda(x > 0.8) \overset{(1)}{=} \frac{\varepsilon_S}{\varepsilon_B} = 215.4, \qquad \mathrm{P}(S \mid x > 0.8) \overset{(2)}{=} 0.6851.$$

(1) The likelihood ratio of the event "$x > 0.8$" is the ratio of the probabilities of that event, the two efficiencies.

(2) Posterior odds 215.4/99, converted with (5.10).

This exceeds $\mathrm{P}(S \mid x = 0.8) = 0.3926$ because applying the cut selects the entire range of scores between 0.8 and 1. Most of them score well above 0.8, and their scores are much stronger evidence. An event that sits exactly at the cut is among the *least* signal-like in the sample. Bayes' theorem conditions on exactly the statement we say we know, nothing more. "The score is 0.8" and "the score is above 0.8" are different observations with different likelihood ratios, and mixing them is a common error in quoting per-event probabilities.

Why the likelihood ratio is the right score. For two hypotheses, (5.11) says that the posterior depends on the data only through Λ. Any two events with the same Λ are equally signal-like, whatever else differs between them. This is the Bayesian face of the Neyman–Pearson lemma of Chapter 4: the most powerful selection at any background rejection is a cut on Λ. A classifier whose score is a monotonic function of Λ is already optimal, and our score is one, since $(x/(1-x))^3$ increases with x.

How this is used in real research: The likelihood-ratio trick

A classifier trained with balanced samples of simulated signal and background, and a loss such as the cross-entropy, learns in the limit of enough data and flexibility to output $s(x) = \mathrm{P}(S \mid x)$ at $\pi = \frac{1}{2}$, that is $s = \Lambda/(1+\Lambda)$. Then $\Lambda(x) = s/(1-s)$ is recovered from the output, and (5.11) converts it to the posterior for the actual signal fraction of any sample. The same trick turns classifiers into estimators of likelihood ratios between two parameter

points, which is how simulation-based inference compares hypotheses whose likelihoods cannot be written down.

5.5 Sequential updating: yesterday's posterior is today's prior

We find a coin on the pavement and want to know how biased it is. We toss it once and it lands heads. It is tempting to say that one toss tells us nothing. It tells us something: the coin is not two-tailed, a possibility that was open a moment ago, and coins that favour heads are now a little more credible than before [SeeingTheory p. 53]. We toss again, and again. Each toss should move our belief a little more, starting from wherever the previous tosses left it. Is that legitimate? Do we get the same answer as if we had waited for all the tosses and analysed them at once? And does it matter in which order the tosses came?

5.5.1 The update rule

Let D_1 be the first piece of data and D_2 the second. Everything we know after both is $\mathrm{P}(H \mid D_1, D_2)$. The question is whether we can reach it in two steps: first update on D_1, then update the result on D_2. To see, apply Bayes' theorem to D_2 alone, with D_1 added to the background information of every probability:

$$\mathrm{P}(H \mid D_1, D_2) \overset{(1)}{=} \frac{\mathrm{P}(D_2 \mid H, D_1)\ \mathrm{P}(H \mid D_1)}{\mathrm{P}(D_2 \mid D_1)} \tag{5.15}$$

$$\overset{(2)}{=} \frac{\mathrm{P}(D_2 \mid H)\ \mathrm{P}(H \mid D_1)}{\sum_j \mathrm{P}(D_2 \mid H_j)\ \mathrm{P}(H_j \mid D_1)}. \tag{5.16}$$

(1) Bayes' theorem (5.3) with $X = H$, $Y = D_2$ and background information I replaced by "I and D_1". Every rule of probability holds with extra information in the conditioning, since $\mathrm{P}(\cdot \mid D_1)$ is itself a probability [Sivia §1.2].

(2) *If* D_2 is independent of D_1 given each hypothesis, $\mathrm{P}(D_2 \mid H_j, D_1) = \mathrm{P}(D_2 \mid H_j)$. The denominator is then the marginalisation (5.4) over the H_j, done inside $\mathrm{P}(\cdot \mid D_1)$.

Line (5.16) is Bayes' theorem for D_2 alone, with the posterior after D_1 in the place of the prior. So the answer is yes: the posterior of one step is the prior of the next. This needs one condition, that the data are **conditionally independent** [independent once the hypothesis is fixed, although not independent when the hypothesis is unknown]. Line (5.15) needs no assumption at all.

The order does not matter. Doing D_2 first and D_1 second must give the same answer. Both routes compute the probability of H given the *single* statement "D_1 and D_2", and "D_1 and D_2" is the same statement as "D_2 and D_1". The formula shows the same thing. Under conditional independence, doing the two updates in one go gives [Kruschke §5.2.1]

$$\mathrm{P}(H \mid D_1, D_2) \overset{(1)}{=} \frac{\mathrm{P}(D_1, D_2 \mid H)\ \mathrm{P}(H)}{\sum_j \mathrm{P}(D_1, D_2 \mid H_j)\ \mathrm{P}(H_j)} \overset{(2)}{=} \frac{\mathrm{P}(D_1 \mid H)\ \mathrm{P}(D_2 \mid H)\ \mathrm{P}(H)}{\sum_j \mathrm{P}(D_1 \mid H_j)\ \mathrm{P}(D_2 \mid H_j)\ \mathrm{P}(H_j)}$$

$$\overset{(3)}{=} \frac{\mathrm{P}(D_2 \mid H)\ \mathrm{P}(D_1 \mid H)\ \mathrm{P}(H)}{\sum_j \mathrm{P}(D_2 \mid H_j)\ \mathrm{P}(D_1 \mid H_j)\ \mathrm{P}(H_j)} \overset{(4)}{=} \mathrm{P}(H \mid D_2, D_1).$$

(1) Bayes' theorem for the combined data.

(2) Conditional independence: the joint likelihood is the product.

(3) Multiplication of numbers commutes.

(4) Bayes' theorem read backwards, with the data in the other order.

Updating one datum at a time computes the same product in stages. Each stage multiplies by one likelihood and renormalises. Renormalising early instead of at the end changes nothing, because a constant factor cancels in the ratio.

Key idea: Sequential updating

For data that are independent given the hypothesis, updating one datum at a time, each posterior serving as the next prior, gives exactly the posterior of updating with all the data at once, in any order. In log-odds form each datum simply adds its log likelihood ratio. Without conditional independence the general rule (5.15) still holds, and the order still does not matter, but each datum's likelihood must then be conditioned on the data before it.

5.5.2 The sample space, toss by toss

A coin is either fair, P(heads) = 0.5, or biased, P(heads) = 0.8, and we give each prior $\frac{1}{2}$. We toss it four times and see H, H, T, H. Figure 5.13 shows two ways of watching the sample space shrink.

The top row keeps the original sample space, drawn as in Figure 1.11: two strips of width $\frac{1}{2}$, the biased coin on the left. After each toss, the evidence "the tosses so far" is a box, and the part of each strip that it covers is the part that survives. The share of the strip it covers is the probability of the whole sequence so far under that coin: 0.8, 0.64, 0.128, 0.1024 for the biased coin and 0.5, 0.25, 0.125, 0.0625 for the fair one. The box only ever shrinks. After all four tosses its area is

$$\mathrm{P}(\mathrm{HHTH}) \overset{(1)}{=} \tfrac{1}{2} \times 0.8^3 \times 0.2 + \tfrac{1}{2} \times 0.5^4 = 0.0512 + 0.03125 = 0.08245,$$

(1) Marginalisation over the two coins, with the product of the four toss probabilities (conditionally independent given the coin) as each likelihood.

the probability, before any toss, of seeing exactly this sequence.

The bottom row blows the surviving box back up to area one after each toss. The dividing line is then the current posterior, which is the prior for the next toss: 0.5, then 0.6154 after H, 0.7191 after HH, 0.5059 after HHT, and 0.621 after HHTH. The box in each rescaled panel is the next toss, covering a share of each strip equal to its likelihood. A head is $0.8/0.5 = 1.6$ times more likely from the biased coin and moves the line right; the tail is $0.2/0.5 = 0.4$ times as likely and moves it back left. Both rows give the same final posterior, $0.0512/0.08245 = 0.621$: renormalising at every step or only at the end is the same calculation.

5.5.3 A grid of biases

Real coins are not limited to two types. The real question is: what is the coin's probability of heads θ, which could be anywhere in $[0, 1]$? Kruschke's answer is to allow a grid of candidate values, say $\theta_j \in \{0, 0.01, \ldots, 1\}$, as if a factory made coins of 101 types, and to treat each θ_j as one hypothesis [Kruschke §5.3]. A single toss y ($y = 1$ for heads, 0 for tails) has the Bernoulli likelihood [Kruschke (5.10)]

$$\mathrm{P}(y \mid \theta_j) = \theta_j^{\,y}(1 - \theta_j)^{1-y}, \tag{5.17}$$

which is θ_j for a head and $1 - \theta_j$ for a tail. Updating is then one line of code:

Listing 5.3: One toss: multiply by the likelihood on the grid and renormalise (`code/ch05/05_coin_grid.py`)

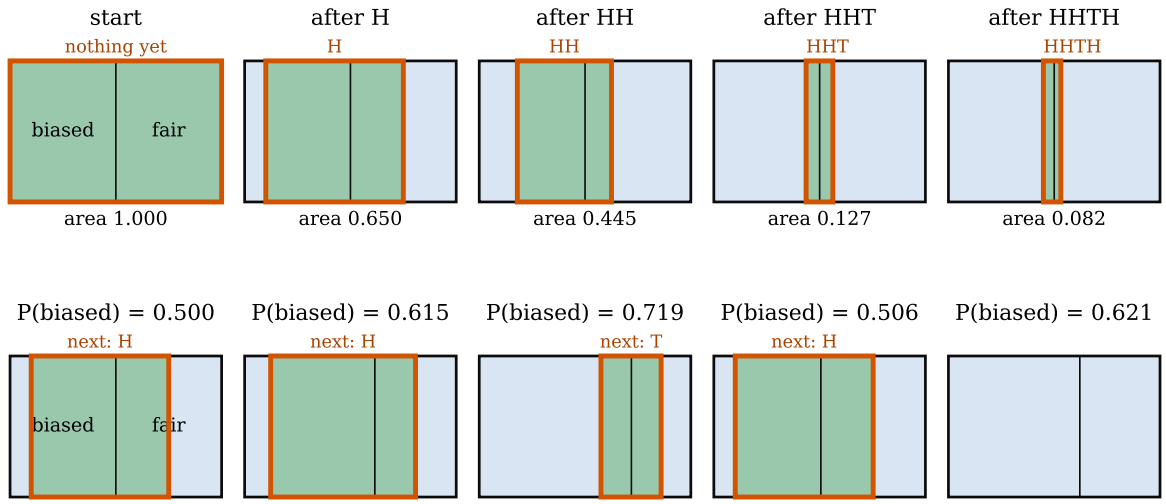


Figure 5.13: The sample space shrinking toss by toss, for a biased coin (left strip) against a fair one (right strip), drawn to scale as in Figure 1.11. Top row: the original sample space, with the tosses so far as a box (orange outline) that covers 0.8 or 0.5 of each strip per head and 0.2 or 0.5 per tail; the number under each panel is the area of the box, the probability of the sequence so far. Bottom row: the surviving box rescaled to area one, so that the dividing line is the current posterior of the biased coin; the box drawn in it is the next toss (the last panel has no next toss).

```
34 def update(prior, y):
35     """One toss y (1 = heads, 0 = tails): multiply by the Bernoulli likelihood,
       renormalise."""
36     post = prior * (theta if y == 1 else 1 - theta)
37     return post / post.sum()
```

Line 36 multiplies the whole array of credibilities by the array of likelihoods, one value per grid point, and line 37 divides by the sum, the discrete evidence. This is the Bayes box of Section 5.3 with 101 rows, recomputed after each toss.

The main lesson of Figure 5.14 is that with few data the prior matters, and with many data it hardly does. The figure follows 50 tosses of a coin with true $\theta = 0.3$, which produced 16 heads. It starts from two priors. One is flat. The other is triangular, with its peak at $\frac{1}{2}$ and zero at 0 and 1; it encodes the belief that most coins are roughly fair.

Each toss acts in a way we can predict. After one tail, the flat-prior posterior is the straight line $1 - \theta$, because the likelihood of a tail is $1 - \theta$. The coin that always lands heads, $\theta = 1$, is ruled out. After the first head, $\theta = 0$ is ruled out too. As tosses accumulate, the posterior narrows around the observed fraction of heads. The two posteriors move towards each other, and after 50 tosses they nearly coincide; the flat-prior posterior has mean 0.3269. So with few data the posterior is a compromise between prior and likelihood, and with many the likelihood dominates [Kruschke §5.3.1–5.3.2]. Kruschke's numbers show the same thing on a grid of 1001 values with the triangular prior [Kruschke Fig. 5.2]. One head in 4 tosses gives a posterior mode of 0.4, pulled far from the observed fraction 0.25 towards the prior's peak. Ten heads in 40 tosses give 0.268, close to 0.25.

The script also checks the order claim numerically. It updates the flat prior with the same 50 tosses in 1000 random orders and compares each result with the batch posterior $\propto \theta^z(1-\theta)^{N-z}$ computed in one step (z heads in N tosses, [Kruschke (5.11)]). The largest difference at any grid point is 1.041×10^{-16}, the size of floating-point rounding.

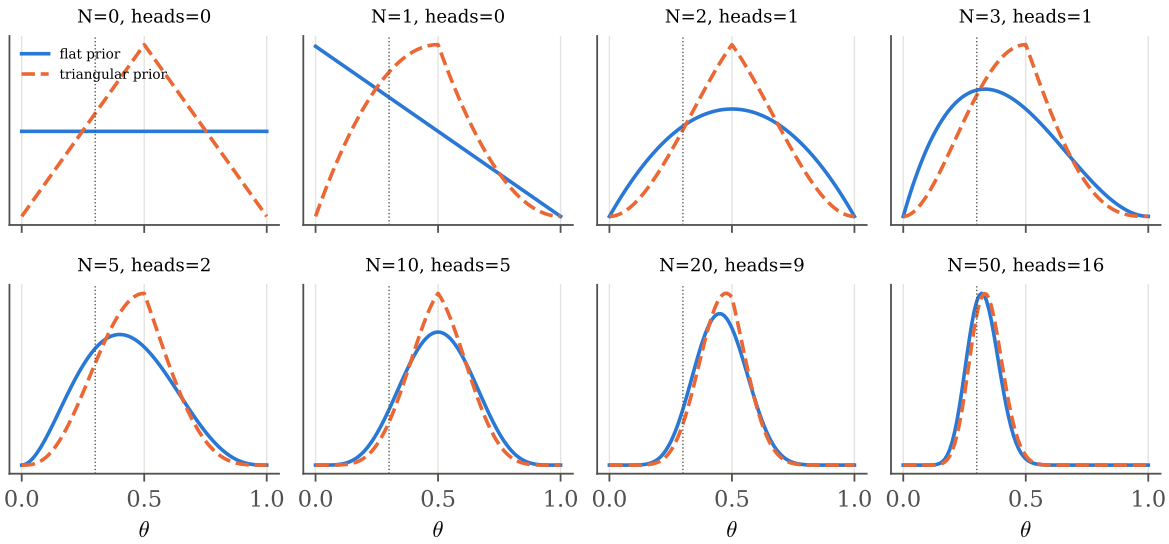


Figure 5.14: Credibility of 101 candidate biases θ after N tosses of a coin whose true bias is 0.3 (dotted line), starting from a flat prior (solid) and from a triangular prior peaked at $\frac{1}{2}$ (dashed). Early on the two posteriors differ, because the data are few; by $N = 50$ both are narrow and nearly identical, centred near the observed fraction of heads.

5.5.4 Evidence accumulates as a random walk

Return to two hypotheses, now a fair coin ($\theta_0 = 0.5$) against a biased one ($\theta_1 = 0.7$), with prior odds 1. By (5.12) the $\log_{10}$ odds of biased against fair add $\log_{10}(0.7/0.5) = 0.1461$ for each head and $\log_{10}(0.3/0.5) = -0.2218$ for each tail. The log odds perform a random walk (Figure 5.15). If the coin really is biased, the average step per toss is

$$\mathbb{E}[\text{step} \mid \theta_1] \overset{(1)}{=} \theta_1 \log \frac{\theta_1}{\theta_0} + (1-\theta_1) \log \frac{1-\theta_1}{1-\theta_0} \equiv D_{\mathrm{KL}}(\theta_1 \| \theta_0), \tag{5.18}$$

(1) A head, with probability θ_1, contributes $\log(\theta_1/\theta_0)$; a tail, with probability $1-\theta_1$, contributes $\log[(1-\theta_1)/(1-\theta_0)]$.

This average step is the **Kullback–Leibler divergence** between the two coins [the average log likelihood ratio in favour of the true distribution; in statistical mechanics it is the relative entropy]. It is never negative, and it is zero only if the two coins are the same. So the walk drifts towards the truth, whichever coin is true. The rate is 0.03574 in $\log_{10}$ per toss when the biased coin is true, and 0.03786 when the fair coin is true. Reaching odds of 100 : 1, which is 2 in $\log_{10}$, therefore takes about $2/0.03574 \approx 56$ tosses on average, with large scatter from one sequence to the next. The more alike the two hypotheses, the smaller the divergence and the longer the wait: telling apart two nearly identical coins, like two nearly identical cosmologies, needs a lot of data.

5.5.5 When the data are not independent given the hypothesis

Conditional independence is an assumption about how the data are produced, and it can fail. When it fails, repeated evidence is worth less than the product of its parts. Take the medical test again (prevalence 0.01, $\mathrm{P}(+ \mid D) = 0.95$, $\mathrm{P}(+ \mid D^c) = 0.10$), and suppose its false positives have two causes. Half of them come from a harmless antibody that 5% of healthy people carry, and it makes the test positive every time. The other half are random, and happen to the other healthy people with probability $r = 0.05/0.95$ per test. The single-test false positive rate is still

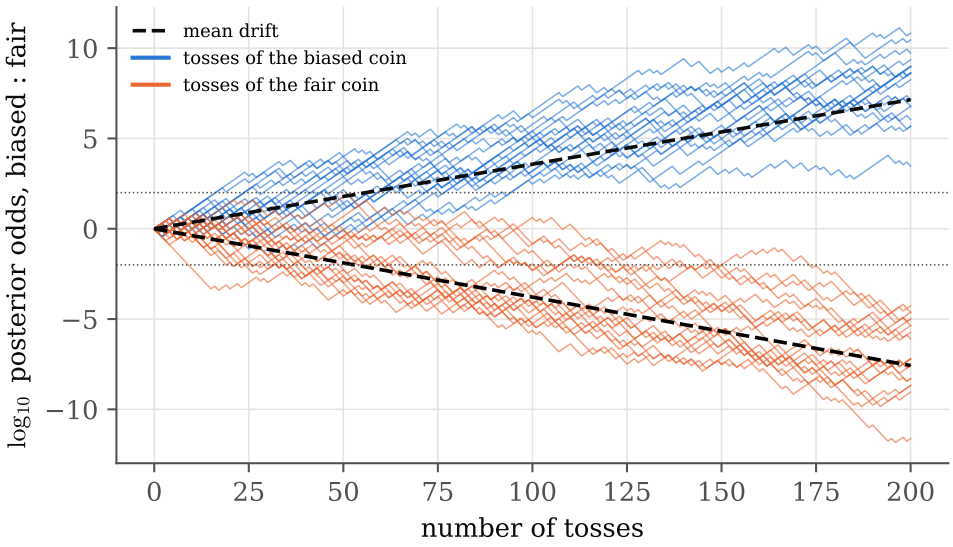


Figure 5.15: Log posterior odds of "biased ($\theta = 0.7$)" against "fair" along 20 simulated sequences of 200 tosses from each coin, starting at odds 1. Every head steps up, every tail steps down. Sequences from the biased coin (blue) drift upward and those from the fair coin (orange) downward, each at the rate given by a Kullback–Leibler divergence (dashed lines). The dotted lines mark odds of 100 : 1 each way.

$0.05 + 0.95\,r = 0.10$. But for a healthy person two positive tests in a row are now much more likely than 0.10^2:

$$\mathrm{P}(++ \mid D^c) \overset{(1)}{=} 0.05 \times 1 + 0.95 \times r^2 = 0.05 + 0.95 \times 0.00277 = 0.0526.$$

(1) Split the healthy population into antibody carriers (both tests positive for sure) and the others (two independent random false positives).

For the diseased we keep independent tests, $\mathrm{P}(++ \mid D) = 0.95^2 = 0.9025$. The likelihood ratio of two positives is $0.9025/0.0526 = 17.1$, not $9.5^2 = 90$. Starting from odds 1/99, the posterior odds are 0.173 and $\mathrm{P}(D \mid ++) = 0.15$, where the naive product would have given 0.48. The second test is weaker evidence because a large part of what made the first test positive in a healthy person repeats itself.

Watch out: Correlated data are not independent evidence

Multiplying likelihoods, or adding log likelihood ratios, assumes that the data are independent *given* each hypothesis. Repeated measurements that share a calibration offset, a common background mismodelling, overlapping sky patches or the same patient-specific cause are correlated even given the hypothesis. Treating them as independent double-counts the shared part and makes the evidence look far stronger than it is. The remedy is to write the joint likelihood with its correlations, as a multivariate Gaussian with a full covariance matrix $\mathbf{C}$ does for Gaussian data.

5.5.6 Screening for breast cancer, test by test

A woman of fifty with no symptoms goes for a routine screening mammogram. The result is positive. She is recalled for an ultrasound, and if that is positive too, for a biopsy. After each result she, and her doctor, want to know one thing: how probable is it now that she has cancer? We use round numbers of the right size for this age group, illustrative rather than

taken from a particular study: a **prevalence** [the fraction of women screened at this age who have breast cancer at the time of screening] of 1%; for the mammogram a sensitivity of 0.87 and a specificity of 0.90; for the ultrasound 0.80 and 0.90; for the biopsy 0.95 and 0.99. Sensitivity is $\mathrm{P}(+ \mid C)$, the chance that the test is positive if she has cancer (C); specificity is $\mathrm{P}(- \mid C^c)$, the chance that it is negative if she has none (C^c).

There are two hypotheses throughout, C and C^c, and each test redistributes the whole unit of probability between them. The posterior after one test is the prior for the next, by (5.16), as long as the tests are conditionally independent given her true state; we come back to that assumption at the end.

Test 1: the mammogram. *The event.* What we observed is "the mammogram is positive". Call it M_+.

What could have produced it? The only thing we do not know is whether she has cancer. So there are two explanations: she has cancer and the mammogram detected it, or she has none and the mammogram raised a false alarm.

How probable was the event under each hypothesis?

- Under C the mammogram is positive when it detects the tumour, which is the sensitivity: $\mathrm{P}(M_+ \mid C) = 0.87$.
- Under C^c it is positive only by a false alarm, from dense tissue, a benign lump or a shadow: $\mathrm{P}(M_+ \mid C^c) = 1 - 0.90 = 0.10$.

Only then the priors, and the update. Before the test her probability of cancer was the prevalence, 0.01, so her prior odds were 1/99. With the odds form (5.11),

$$O(C \mid M_+) \overset{(1)}{=} \frac{\mathrm{P}(M_+ \mid C)}{\mathrm{P}(M_+ \mid C^c)} \cdot \frac{\mathrm{P}(C)}{\mathrm{P}(C^c)} \overset{(2)}{=} \frac{0.87}{0.10} \cdot \frac{0.01}{0.99} = \frac{8.7}{99} = 0.0879, \qquad \mathrm{P}(C \mid M_+) \overset{(3)}{=} \frac{0.0879}{1.0879} = 0.0808.$$

(1) Posterior odds are the likelihood ratio times the prior odds.
(2) The two likelihoods just found, and the prevalence.
(3) Odds back to probability, (5.10).

A positive mammogram leaves her with about an 8% chance of cancer. The positive result is 8.7 times more probable if she has cancer, but cancer was 99 to 1 against, and 8.7 times 1/99 is still about 1 to 11. The low base rate is the whole reason: among a hundred women screened, about one has cancer and is very likely to test positive, while about ten of the ninety-nine healthy women test positive too. This is the picture of Figure 5.5, drawn for the mammogram's numbers in the first panel of Figure 5.17.

A positive and a negative result do not carry the same information. For any test, write the odds update for each result:

$$O(C \mid +) \overset{(1)}{=} \frac{\mathrm{P}(+ \mid C)}{\mathrm{P}(+ \mid C^c)}\, O(C) = \underbrace{\frac{\text{sens}}{1 - \text{spec}}}_{\mathrm{LR}_+}\, O(C), \qquad O(C \mid -) \overset{(2)}{=} \frac{\mathrm{P}(- \mid C)}{\mathrm{P}(- \mid C^c)}\, O(C) = \underbrace{\frac{1 - \text{sens}}{\text{spec}}}_{\mathrm{LR}_-}\, O(C).$$

(1) (5.11) with $E = +$; $\mathrm{P}(+ \mid C^c) = 1 - \mathrm{P}(- \mid C^c)$.
(2) (5.11) with $E = -$; $\mathrm{P}(- \mid C) = 1 - \mathrm{P}(+ \mid C)$.

These two numbers are the **positive and negative likelihood ratios** of the test. For the mammogram $\mathrm{LR}_+ = 8.70$ and $\mathrm{LR}_- = 0.144$: a positive multiplies the odds by 8.7, a negative

divides them by 6.9. They are not reciprocals, and they are controlled by different numbers. LR_+ is large when the false-alarm rate $1 - \text{spec}$ is small, so a highly specific test is good at confirming (the biopsy has $\mathrm{LR}_+ = 95.0$). LR_- is small when the miss rate $1 - \text{sens}$ is small, so a highly sensitive test is good at excluding. On the probability scale the low prior sharpens the asymmetry: a negative mammogram would have moved her from 1% to 0.00146, less than one percentage point, while the positive moved her to 8%.

Test 2: the ultrasound. *The event.* The ultrasound is positive, U_+.

What could have produced it? The same two explanations, but their weights are no longer 0.01 and 0.99. They are the posteriors after the mammogram, 0.0808 and $1 - 0.0808$: the first result is now part of the prior.

How probable was the event under each? Under C, $\mathrm{P}(U_+ \mid C) = 0.80$. Under C^c, $\mathrm{P}(U_+ \mid C^c) = 0.10$, provided the ultrasound's false alarms have nothing to do with the mammogram's once her state is fixed. So $\mathrm{LR}_+ = 8.00$, and

$$O(C \mid M_+, U_+) \overset{(1)}{=} 8.0 \times 0.0879 = 0.703, \qquad \mathrm{P}(C \mid M_+, U_+) = \frac{0.703}{1.703} = 0.413.$$

(1) The update (5.16) in odds form: the posterior odds after the mammogram are the prior odds for the ultrasound.

A second positive multiplies the odds again, and she is now at about 41%. Had the ultrasound been negative, its $\mathrm{LR}_- = 0.222$ would have pulled the odds down by a factor 4.5, to 0.0195 and $\mathrm{P}(C) = 0.0192$. One positive and one negative do not cancel: $8.7 \times 0.222 = 1.93$, so she would still be about twice as likely to have cancer as before screening.

Test 3: the biopsy. *The event.* The biopsy is positive, B_+.

What could have produced it? The same two explanations, now weighted by the posterior after two positive images, odds 0.703.

How probable was the event under each? A pathologist looking at the tissue finds the cancer with probability 0.95 if it is there, and reports cancer in healthy tissue with probability 0.01. So $\mathrm{LR}_+ = 0.95/0.01 = 95$.

The update. The odds become $95 \times 0.703 = 66.8$, and $\mathrm{P}(C \mid M_+, U_+, B_+) = 0.9852$. A negative biopsy instead would give odds $0.703 \times 0.0505 = 0.0355$ and $\mathrm{P}(C) = 0.0343$. That is still above the prevalence, which is why a negative biopsy after two positive images is often repeated.

Figure 5.16 shows all eight sequences of results at once, and Figure 5.17 the sample space re-drawn before each test along the path +, +, +, with the current posterior as the width of the cancer strip. The calculation is one function, called once per test:

Listing 5.4: One test is one Bayes update; three tests are three updates in a row (`code/ch05/30_screening.py`)

```
32 def update(p, sens, spec, result):
33     """One test: multiply by the likelihood of the result under {cancer, no cancer},
       renormalise."""
34     likes = [sens, 1 - spec] if result == "+" else [1 - sens, spec]
35     return posterior([p, 1 - p], likes)[2][0]
36
37
38 def lr(sens, spec):
39     return sens / (1 - spec), (1 - sens) / spec
40
41
42 # ---------- every sequence of results ----------
```

```
paths = {}
for res in itertools.product("+-", repeat=3):
    p, traj = prev, [prev]
    for (name, se, sp), r in zip(tests, res):
        p = update(p, se, sp, r)
        traj.append(p)
    paths["".join(res)] = traj
```

Line 34 picks the likelihoods of the observed result under C and C^c, line 35 is Bayes' theorem through `posterior` of `lib_area.py` (Section 5.2.5), and lines 44–49 feed each posterior into the next test for every one of the 2^3 sequences. The script also simulates 4×10^6 women, gives each a true state and three test results, and keeps those with the observed results, conditioning by filtering as in Chapter 1. The fraction with cancer among them is 0.0805 after M_+, 0.413 after M_+U_+ and 0.9847 after all three, agreeing with the exact values within the Monte Carlo error.

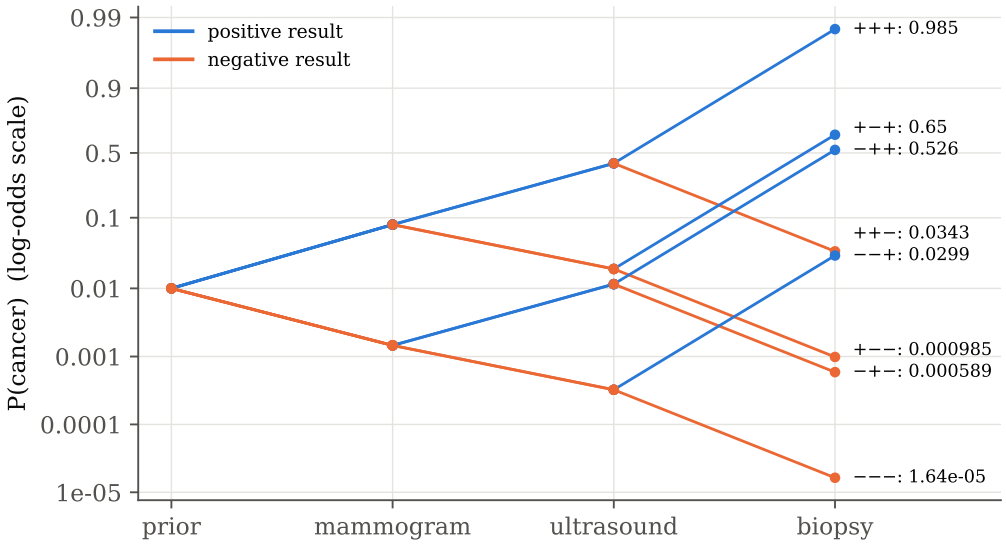


Figure 5.16: The probability of cancer after each test, for every sequence of results, on a log-odds axis, where each test moves the curve by the same step whatever came before. A positive (blue) steps up by $\log_{10} \mathrm{LR}_+$, a negative (orange) down by $\log_{10} \mathrm{LR}_-$; the end labels give the final probability for each sequence. The biopsy's steps are much longer than the imaging tests', because its likelihood ratios are much further from one.

The assumption behind the multiplication. Multiplying the odds by each test's likelihood ratio assumes that the tests are independent given her true state. For a healthy woman this says that a false alarm on the mammogram makes a false alarm on the ultrasound neither more nor less likely. That can fail: two imaging tests can be fooled by the same thing, since a benign lump or very dense tissue may look suspicious to both. Model it as for the antibody in the previous subsection. Say 5% of healthy women carry such a feature and test positive on both. The rest have independent false alarms at the rate $r = 0.05/0.95$, chosen so that each test alone still has false-positive rate 0.10. Then three numbers change: $\mathrm{P}(M_+U_+ \mid C^c) = 0.05 + 0.95\,r^2 = 0.0526$ instead of 0.01; the pair of positives has likelihood ratio $0.87 \times 0.80/0.0526 = 13.2$ instead of $8.7 \times 8.0 = 70$; and her probability after two positive images is 0.118 (simulation 0.119), not 0.413. The second positive is worth less, because much of what made the first one positive in a healthy woman repeats itself. The biopsy examines the tissue itself, a different failure mode altogether, which is why it is the test that settles the question.

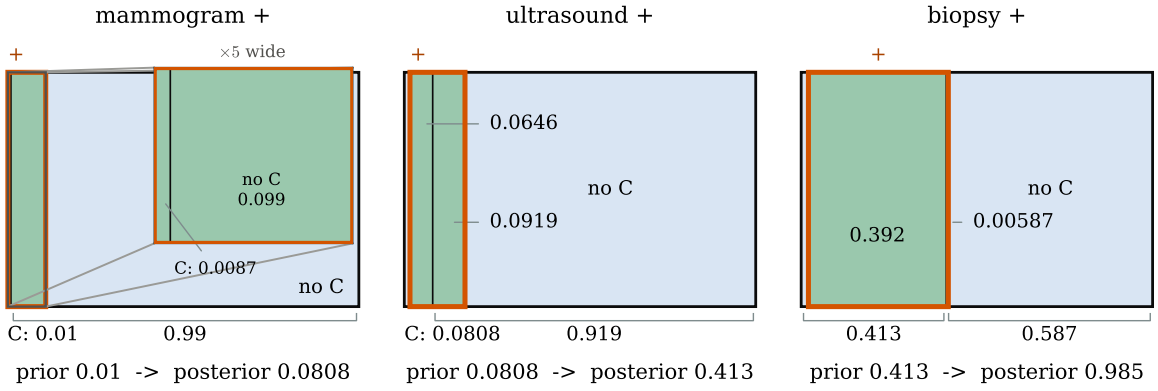


Figure 5.17: The sample space re-drawn before each test along the path +, +, +, in the box picture of Figure 1.11, with the function of Section 5.2.5. In each panel the width of the cancer strip (C) is the current prior, and the positive result is the box (orange outline): it covers the sensitivity's share of the C strip and the false-alarm rate's share of the other, and the numbers are the areas of its two pieces. Before the mammogram the cancer strip is 1% of the width (magnified in the inset); after it, the posterior 0.081 becomes the width of the strip for the ultrasound, and after the ultrasound, 0.41 becomes the width for the biopsy.

Problem. In a younger screening population the prevalence is 0.5%. Test A has sensitivity 0.90 and specificity 0.92; test B has sensitivity 0.85 and specificity 0.95. (a) Assuming the tests are conditionally independent, find $\mathrm{P}(C \mid A_+, B_+)$ and $\mathrm{P}(C \mid A_+, B_-)$. (b) How many independent positive results of test A are needed for the probability of cancer to exceed 95%? (c) Now suppose that 3% of healthy women carry a benign feature that makes both A and B positive, while the other healthy women have independent false positives at rates chosen so that each test alone keeps its specificity. Redo (a), and say why the answers moved.

Solution. (a) The prior odds are $O_0 = 0.005/0.995 = 0.005025$. The likelihood ratios are $\mathrm{LR}^A_+ = 0.90/0.08 = 11.25$, $\mathrm{LR}^B_+ = 0.85/0.05 = 17.0$ and $\mathrm{LR}^B_- = 0.15/0.95 = 0.1579$. Then

$$O(C \mid A_+, B_+) \overset{(1)}{=} \mathrm{LR}^B_+ \, \mathrm{LR}^A_+ \, O_0 = 17.0 \times 11.25 \times 0.005025 = 0.9611, \quad \mathrm{P}(C \mid A_+, B_+) \overset{(2)}{=} 0.490,$$

$$O(C \mid A_+, B_-) \overset{(3)}{=} \mathrm{LR}^B_- \, \mathrm{LR}^A_+ \, O_0 = 0.008926, \qquad \mathrm{P}(C \mid A_+, B_-) = 0.00885.$$

(1) Two sequential updates in odds form; conditional independence lets each test enter through its own likelihood ratio.
(2) $p = O/(1+O)$.
(3) The same with the negative ratio of B.

After a positive A alone the probability is 0.0535; a positive B brings it to about one half, while a negative B brings it back to under 1%, still above the prevalence of 0.5%.

(b) After n independent positives of A the odds are $(\mathrm{LR}^A_+)^n O_0$, and 95% means odds of 19:

$$(\mathrm{LR}^A_+)^n \, O_0 \geq 19 \overset{(1)}{\iff} n \geq \frac{\ln(19/O_0)}{\ln \mathrm{LR}^A_+} = \frac{\ln 3781}{\ln 11.25} = 3.40.$$

(1) Take logarithms; $\ln \mathrm{LR}^A_+ > 0$, so the inequality keeps its direction.

So $n = 4$: three positives give 0.877, four give 0.9877. This assumes that repeating test A on the same woman gives independent errors, which for a repeated image of the same breast is exactly the assumption that part (c) questions.

(c) With a fraction $g = 0.03$ of healthy women carrying the feature, the other healthy women must have false-positive rates $q_A = (0.08 - g)/(1 - g) = 0.05155$ and $q_B = (0.05 - g)/(1 - g) = 0.02062$, so that each test alone is unchanged. For a healthy woman,

$$\mathrm{P}(A_+B_+ \mid C^c) \overset{(1)}{=} g + (1-g)\, q_A q_B = 0.03103, \qquad \mathrm{P}(A_+B_- \mid C^c) \overset{(2)}{=} (1-g)\, q_A (1-q_B) = 0.04897,$$

(1) Carriers are positive on both; the others are positive on both by two independent false alarms.
(2) A carrier is never negative on B, so only non-carriers contribute.

while for a woman with cancer nothing changes, $\mathrm{P}(A_+B_+ \mid C) = 0.90 \times 0.85$ and $\mathrm{P}(A_+B_- \mid C) = 0.90 \times 0.15$. The joint likelihood ratios replace the products of (a):

$$\mathrm{P}(C \mid A_+, B_+) \overset{(1)}{=} \frac{O_0 \cdot 0.765/0.03103}{1 + O_0 \cdot 0.765/0.03103} = 0.110, \qquad \mathrm{P}(C \mid A_+, B_-) \overset{(2)}{=} 0.0137.$$

(1) Odds form with the likelihood ratio of the *pair* of results, $0.765/0.03103 = 24.7$, instead of $11.25 \times 17 = 191$.
(2) The same with $0.135/0.04897$.

Both answers moved towards the prior. Two positives are now worth a factor 24.7 rather than 191, because a healthy woman who is positive on A is quite likely to be a carrier (0.03/0.08 of them), and a carrier is positive on B for certain. The negative B is less reassuring for the same reason: among healthy women positive on A, a negative B is now rarer, so it says less about health. The script checks all four numbers by simulation, 0.489, 0.00898, 0.110 and 0.0139.

5.6 The denominator: the sum over all paths

The denominator $\mathrm{P}(E)$ of Bayes' theorem is the model's prediction for the data. So far it has looked like bookkeeping: the number we divide by so that the posteriors add up to one. In the odds form it cancelled altogether. But it is a probability in its own right, and a meaningful one. Recall the trigger problem of Section 5.1, where E was an isolated high-p_T muon and three trigger paths could have fired the event. There $\mathrm{P}(E) = 0.235$. This says that, *before* we looked at the event, the model of the three paths predicted that 23.5% of recorded events would contain such a muon. Suppose a large sample of events showed 60%. No reallocation of credibility among the three paths could explain that. The trouble would lie in the list of paths or in the forward probabilities, not in the priors. The denominator is also the one place in Bayes' theorem that needs every hypothesis at once.

5.6.1 Every path, and only paths

The denominator is the marginalisation (5.4): the sum, over every hypothesis, of the probability of reaching the observation along that hypothesis's path (Figure 5.18),

$$\mathrm{P}(E) = \sum_i \mathrm{P}(E \mid H_i)\ \mathrm{P}(H_i) \quad \text{(discrete)}, \qquad p(\boldsymbol{d}) = \int p(\boldsymbol{d} \mid \theta)\, p(\theta)\, \mathrm{d}\theta \quad \text{(continuous)}. \tag{5.19}$$

In the square of Section 5.2 it is the total surviving area. It does not depend on which hypothesis we are asking about. It depends only on the observation and on the whole list of hypotheses with their priors. Two names are in use: the **evidence**, and the **marginal likelihood**, because it is the likelihood averaged over the hypotheses with the prior as the weight [Kruschke §5.2, (5.8)–(5.9)]. Wasserman calls its continuous version $c_n = \int \mathcal{L}_n(\theta) f(\theta)\, \mathrm{d}\theta$ the **normalizing constant** and notes that it can be dropped and recovered later [Wasserman p. 177, (11.3)].

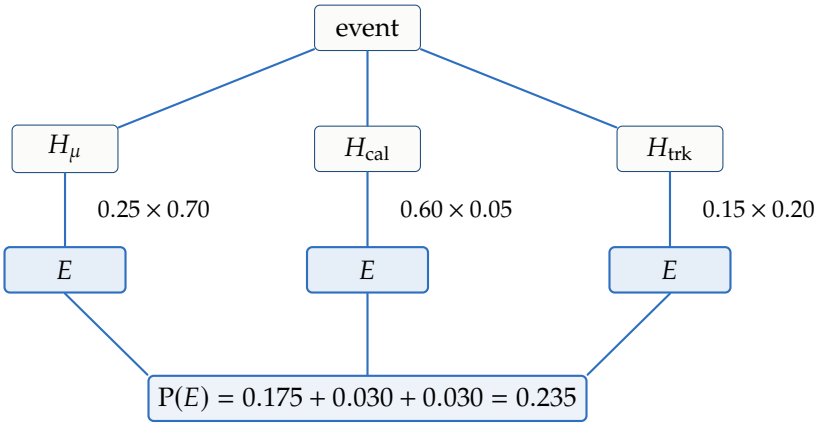


Figure 5.18: The denominator of Bayes' theorem as the sum over paths. Each hypothesis contributes one path from the root to the observed leaf E, with probability equal to the product of its two branch probabilities (written beside the path). The evidence collects all of them. A hypothesis missing from the tree is a path missing from the sum.

5.6.2 The denominator is a prediction

Now read (5.19) as a function of the data rather than as one number. Before the data are taken, ask: how probable is each possible observation E? The answer, $\mathrm{P}(E)$ for every E, is the model's forecast of what we are going to see, averaged over our uncertainty about which hypothesis is true. It is called the **prior predictive distribution** [Lambert §6.3.4]. Unlike the likelihood, it is a genuine probability distribution over the data: summed over all possible observations it gives one, because $\sum_E \mathrm{P}(E \mid H_i) = 1$ for each i and the priors sum to one.

Example: what did we expect from ten tosses? Take the grid of 101 coin biases of Section 5.5 with a flat prior, and ask for the probability of z heads in $N = 10$ tosses before any toss is made. In the continuum limit of the grid, the flat prior density is $p(\theta) = 1$ on $[0, 1]$, and

$$\mathrm{P}(z) \overset{(1)}{=} \int_0^1 \binom{N}{z} \theta^z (1-\theta)^{N-z} \, \mathrm{d}\theta \overset{(2)}{=} \binom{N}{z} \frac{z!\,(N-z)!}{(N+1)!} \overset{(3)}{=} \frac{N!}{z!\,(N-z)!} \cdot \frac{z!\,(N-z)!}{(N+1)!} = \frac{1}{N+1}. \quad (5.20)$$

(1) The evidence (5.19) with the binomial likelihood and the flat prior.

(2) The Beta integral $\int_0^1 \theta^a (1-\theta)^b \, \mathrm{d}\theta = a!\,b!/(a+b+1)!$ for integers $a, b \geq 0$; it follows by integrating by parts b times.

(3) Write out the binomial coefficient; the factorials cancel.

Every number of heads from 0 to 10 was equally expected, with probability $\frac{1}{11} = 0.091$. The 101-point grid gives 0.08993 for $z = 9$, the small difference being the grid's approximation of the integral. A prior concentrated on fair coins predicts something quite different (Figure 5.19). If we are certain that the coin is fair, $\mathrm{P}(z = 9) = \binom{10}{9} 2^{-10} = 0.009766$. The triangular prior sits in between, 0.0604. Each prior predictive sums to one over $z = 0, \ldots, 10$, as the script confirms.

The whole computation is one matrix product:

Listing 5.5: The prior predictive as a sum over paths (`code/ch05/07_evidence.py`)

```
z = np.arange(N + 1)
like = stats.binom.pmf(z[:, None], N, theta[None, :])      # P(z | theta_j), rows z, columns j
pred = {k: like @ p for k, p in priors.items()}            # sum over paths theta_j
```

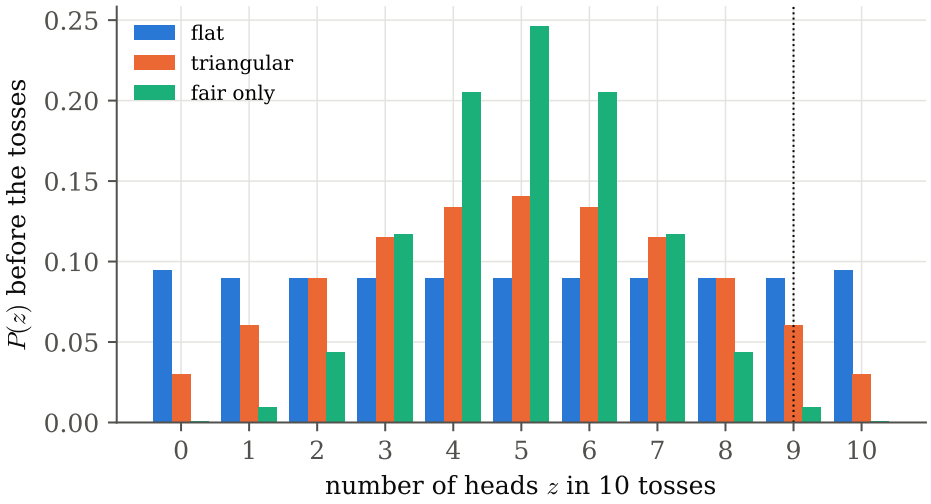


Figure 5.19: The denominator of Bayes' theorem as a forecast. For each of three beliefs about the coin, the bars give the probability of each number of heads in ten tosses, computed before any toss as a sum over all candidate biases. The flat prior spreads its forecast evenly; "the coin is fair" concentrates it near five heads. The dotted line marks the observed nine heads.

Line 34 builds the table of Section 5.1: one row per possible observation z, one column per hypothesis θ_j, entries $\mathrm{P}(z \mid \theta_j)$. Line 35 multiplies it by the vector of priors, which for each row is the sum over the paths θ_j of likelihood times prior.

Now nine heads have been observed. Which model predicted that better? The flat-prior model had given this outcome probability 0.08993, the fair-coin model only 0.009766. So the flat model predicted the observation 9.208 times better. A ratio of two evidences is a likelihood ratio in which each "hypothesis" is a whole model together with its prior. By (5.11) it multiplies the prior odds of one model against the other. This ratio, the *Bayes factor*, is the tool for comparing models, and comparing models is the one job for which the denominator cannot be skipped.

5.6.3 A missing path distorts every posterior

Because the denominator sums over the hypotheses we listed, it silently assumes that the list is complete. If the truth is not on the list, Bayes' theorem still returns posteriors that sum to one, now shared among the wrong candidates. Kruschke's example is the wet pavement caused by something nobody listed [Kruschke §2.2].

Example: forgetting the background. In the gamma-ray example of Section 5.3, the soft photon at $(-3°, 2°)$ had source probabilities 0.4088 (burst), 0.04486 (blazar) and 0.5463 (diffuse background). Suppose we forget the diffuse background and keep only the two point sources. The denominator loses the largest of its three paths, and the same photon now gets 0.9011 for the burst and 0.09887 for the blazar. Nothing in the output warns us: the posteriors still sum to one. The warning is in the denominator itself. Without the background, the model predicts that almost no photons arrive far from the two point sources, so the evidence of a map full of such photons is tiny compared with that of the model with the background. Comparing evidences, or checking whether the most credible hypotheses actually reproduce the data [Kruschke §2.2], is how a missing path is found.

Watch out: A small evidence is not by itself a refutation

Every particular data set has a small probability: any one sequence of ten tosses has probability $2^{-10} \approx 0.001$ under a fair coin. The value of $\mathrm{P}(E)$ alone says little. It becomes informative when compared, either with the evidence of another model for the same data, or with the spread of $\mathrm{P}(E)$ over the data sets the model itself would typically produce.

5.6.4 Why the denominator is the hard part

The denominator is hard because it needs every hypothesis at once, and in many dimensions there are far too many. For a handful of hypotheses the sum (5.19) is trivial. For a continuous parameter it is an integral, which a grid approximates by a sum, as in the 101-point coin grid. For D parameters a grid with n points per axis has n^D points. With $n = 1000$ and $D = 6$, the number of parameters of the standard cosmological model, that is 10^{18} evaluations of the likelihood [Kruschke §5.4], [Lambert §6.4].

There are two ways out. The first is to choose priors that match the form of the likelihood, so that the integral can be done by hand; these *conjugate* priors give posteriors of a known shape. The second is to notice that estimating parameters never needs the denominator at all. It cancels in every ratio of posteriors, exactly as it cancelled in the odds (5.11). A method that only ever compares the posterior at two points, $p(\theta' \mid \boldsymbol{d})/p(\theta \mid \boldsymbol{d}) = p(\boldsymbol{d} \mid \theta')p(\theta')/p(\boldsymbol{d} \mid \theta)p(\theta)$, can explore the posterior without knowing its normalisation. That is the idea behind Markov chain Monte Carlo (Chapter 9) [Lambert §6.5]. Only when models themselves are compared must the evidence be computed, and that needs its own methods.

The trigger paths, the medical test, Monty's door, the urns and the ball sizes were all solved in the same way, and the way is worth making a habit: put the question into words before touching a symbol. First ask: which hypotheses could have produced what we observed? With that list settled, ask: how probable would the observation have been if each of them were true? These two answers sort the whole problem. The first answer says what the posterior is a distribution over, and so fixes the list of paths that the denominator sums over. The second answer is the likelihood: it fixes the height of each bar in the square.

Recap

- Prediction reads $\mathrm{P}(E \mid H)$, inference wants $\mathrm{P}(H \mid E)$. Both come from one joint distribution: a column of the joint table for prediction, the observed row for inference.
- Bayes' theorem is the product rule applied in both orders. Observing E shrinks the sample space to E; each hypothesis is re-measured by its share of the surviving region.
- In the unit square: priors are widths, likelihoods are heights, joint probabilities are areas, the evidence is the surviving area, and the posterior is one surviving rectangle over all of them.
- Base rates matter: evidence covering most of a thin strip (medical test, trigger purity, rare signal) can lose to evidence covering a small share of a wide strip.
- Inference is reallocation of credibility: a zero likelihood eliminates, a zero prior can never be revived, noisy data only nudge.
- Posterior odds = likelihood ratio × prior odds. The likelihood ratio is all the evidence says about two hypotheses; in logarithms, independent evidence adds (decibels, log-odds random walk with drift equal to a Kullback–Leibler divergence).

- Condition on exactly what was observed: "score = 0.8" and "score > 0.8" are different evidence.
- Sequential updating: the posterior becomes the prior; the order of the data does not matter; multiplying likelihoods requires independence given the hypothesis.
- The denominator is the sum over all paths: the normaliser, the prior predictive forecast of the data, and the quantity that compares models. It needs the complete list of hypotheses and is hard to compute in many dimensions, which is why parameter estimation works with ratios of posteriors.

5.7 From a list of hypotheses to a continuous parameter

A theorist hands us a formula. If a parameter λ of nature has a certain value, an instrument should read $\mu(\lambda)$. The parameter could be the mass of a particle, a coupling, a lifetime or the matter density of the universe. The instrument is not perfect, so what it actually reads is the prediction plus some noise. We take one reading, d. Which values of λ could have produced it, and how much credibility should each value get?

We answer it in the same order as the Monty Hall problem (Figure 5.6). *The event.* What we observed is "the instrument showed d". *The question.* What made it show that value? The possible explanations are the possible values of λ, each one a possible state of nature. *Each candidate in turn.* For each value of λ we ask how probable the reading d was if that value were the truth. The prediction would then have been $\mu(\lambda)$, so the question is how likely the noise was to push $\mu(\lambda)$ to exactly the d we saw. *Only then the priors, and the normalisation.* We multiply by the prior weight of each value and divide by the total.

This is the question "which urn did the ball come from?" or "is this event signal or background?" with one change. There the candidate explanations formed a short list, $H_1, H_2, \ldots$. Here every real number in some range is a candidate. Bayes' theorem still holds, but three things need care. A probability spread over a continuum is a density. The sum over paths in the denominator becomes an integral. And a computer can evaluate the density only at finitely many points.

5.7.1 The tree, continuous in spirit

Write the reading as prediction plus noise,

$$d = \mu(\lambda) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2), \tag{5.21}$$

where σ is the known size of the noise. The forward question, prediction, asks: if nature chose λ, what readings would we see, and how often? Its answer is a Gaussian centred on the prediction,

$$p(d \mid \lambda) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\big(d - \mu(\lambda)\big)^2}{2\sigma^2}\right]. \tag{5.22}$$

Held at the reading we have and read as a function of λ, it is the **likelihood** $\mathcal{L}(\lambda)$, which scores each candidate by the probability of the noise, $d - \mu(\lambda)$, it would have needed (Section 3.8).

The inference question runs the other way, and its picture is a probability tree with one branch per candidate value (Figure 5.20). The first split chooses λ with prior weight $p(\lambda)$. The second step produces the reading, with probability $p(d \mid \lambda)$ along each branch. Every branch can reach the same leaf d, because Gaussian noise can turn any prediction into any reading,

only with very different probabilities. With infinitely many branches the tree cannot be drawn literally, but it can be drawn in spirit: pick three candidates $\lambda_1, \lambda_2, \lambda_3$ and remember that there is one branch for every value in between.

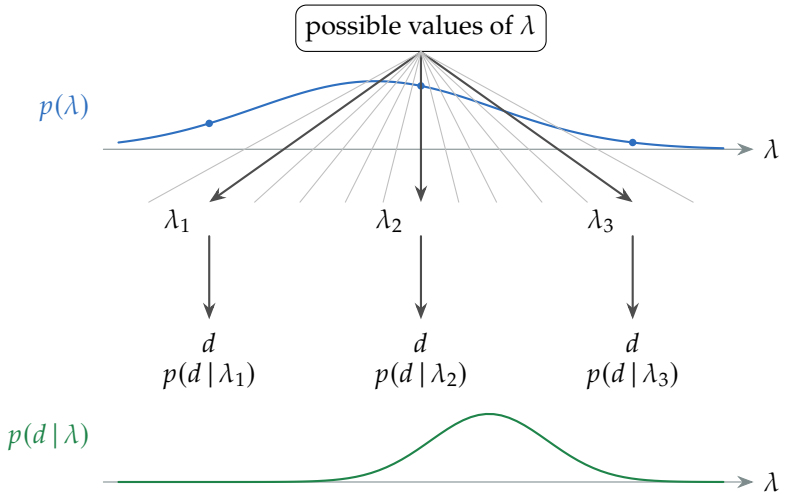


Figure 5.20: The probability tree for a continuous parameter, drawn in spirit. The top curve is the prior density, the weight with which each branch is chosen. Three branches are drawn in full and the faint fan stands for the continuum of branches between them. Every branch can end at the same observed reading d, each with its own probability. The bottom curve collects those probabilities as a function of λ: the likelihood.

Key idea: Parameter estimation is inference with a continuum of explanations

Every candidate value of the parameter λ is one possible explanation of the data. The reading d is the event that actually happened. The likelihood $\mathcal{L}(\lambda) = p(d \mid \lambda)$ measures, value by value, how easily that λ could have produced the event.

5.7.2 From a sum over branches to a density

What does Bayes' theorem become when the hypotheses form a continuum? We reduce the question to the case we already know, a finite list. Chop the λ axis into cells of small width $\Delta\lambda$, centred on the points λ_j, and call "λ lies in cell j" hypothesis H_j. Exactly one cell contains the true value, so these hypotheses are a partition and the discrete theorem applies to them. The prior probability of cell j is $\mathrm{P}(H_j) \approx p(\lambda_j)\,\Delta\lambda$, because a density times a small width is a probability. Then

$$\begin{aligned}\frac{\mathrm{P}(H_j \mid d)}{\Delta\lambda} &\overset{(1)}{=} \frac{1}{\Delta\lambda}\,\frac{p(d \mid \lambda_j)\,p(\lambda_j)\,\Delta\lambda}{\sum_k p(d \mid \lambda_k)\,p(\lambda_k)\,\Delta\lambda} \\ &\overset{(2)}{=} \frac{p(d \mid \lambda_j)\,p(\lambda_j)}{\sum_k p(d \mid \lambda_k)\,p(\lambda_k)\,\Delta\lambda} \\ &\xrightarrow{\Delta\lambda \to 0} \frac{p(d \mid \lambda_j)\,p(\lambda_j)}{\int p(d \mid \lambda)\,p(\lambda)\,\mathrm{d}\lambda}.\end{aligned} \tag{5.23}$$

(1) Discrete Bayes' theorem for the cells, with the prior probability of each cell written as density times width.

(2) The $\Delta\lambda$ in the numerator cancels the $1/\Delta\lambda$ in front. The one in the denominator stays inside the sum.

Now let the cells shrink. In the denominator each term is multiplied by its width, so the sum is a Riemann sum and becomes an integral. On the left, the probability of a cell divided by its width becomes a probability per unit λ: the **posterior density** $p(\lambda \mid d)$. So the continuous theorem is [Wasserman p. 177, (11.1)]

$$p(\lambda \mid d) = \frac{p(d \mid \lambda)\, p(\lambda)}{p(d)}, \qquad p(d) = \int p(d \mid \lambda)\, p(\lambda)\, \mathrm{d}\lambda. \tag{5.24}$$

The denominator $p(d)$ is the old sum over every path to the observation, now an integral over every branch of the tree in Figure 5.20. It does not depend on λ, so the shape of the posterior is set by the numerator alone [Wasserman p. 177, (11.2)–(11.3)]:

$$p(\lambda \mid d) \;\propto\; \mathcal{L}(\lambda)\, p(\lambda), \tag{5.25}$$

"posterior is proportional to likelihood times prior". The constant is recovered at the end by demanding that $\int p(\lambda \mid d)\, \mathrm{d}\lambda = 1$.

Working definition 5.2 (posterior density)

The **posterior density** of a continuous parameter λ given data d is $p(\lambda \mid d) = \mathcal{L}(\lambda)\, p(\lambda)/p(d)$, where $\mathcal{L}(\lambda) = p(d \mid \lambda)$ is the likelihood [the probability density of the data actually observed, read as a function of λ], $p(\lambda)$ is the prior density [credibility per unit λ before the data], and $p(d)$ is the evidence [the same product integrated over all λ. It makes the posterior integrate to one]. *Operationally:* evaluate $\mathcal{L}(\lambda)p(\lambda)$ at every λ you care about, and divide by its integral. The probability that λ lies between a and b is $\int_a^b p(\lambda \mid d)\, \mathrm{d}\lambda$.

A density is not itself a probability. Its value at a point, say $p(\lambda \mid d) = 3$, means "probability $3\,\Delta\lambda$ for a small cell of width $\Delta\lambda$ around this point", and it carries units of one over the units of λ. Only areas under the curve are probabilities. This matters as soon as we change variables (Section 5.12): the height of a density changes when we stretch the axis, while the probabilities of corresponding intervals do not.

Example: a linear prediction and a flat prior. Let $\mu(\lambda) = \lambda$, so the instrument reads the parameter itself plus noise, and take a prior that is constant, $p(\lambda) = 1/(\lambda_{\max} - \lambda_{\min})$, over a range much wider than σ and containing d. Then

$$p(\lambda \mid d) \overset{(1)}{=} \frac{\mathcal{L}(\lambda)\, p(\lambda)}{\int \mathcal{L}(\lambda')\, p(\lambda')\, \mathrm{d}\lambda'} \overset{(2)}{=} \frac{e^{-(d-\lambda)^2/2\sigma^2}}{\int_{\lambda_{\min}}^{\lambda_{\max}} e^{-(d-\lambda')^2/2\sigma^2}\, \mathrm{d}\lambda'} \overset{(3)}{=} \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(\lambda-d)^2/2\sigma^2}.$$

(1) Bayes' theorem (5.24).

(2) The flat prior is the same constant in numerator and denominator and cancels, and so does the factor $1/\sqrt{2\pi\sigma^2}$ of the likelihood.

(3) The range is many σ wide around d, so the integral is, to excellent accuracy, the full Gaussian integral $\int_{-\infty}^{\infty} e^{-u^2/2\sigma^2}\mathrm{d}u = \sqrt{2\pi\sigma^2}$.

The posterior is a Gaussian in λ, centred on the reading d, with width equal to the noise size σ. Values of λ whose prediction lies close to the reading get the most credibility. The formula looks like the likelihood with d and λ swapped, and here it is exactly that. But this is a coincidence of the linear prediction $\mu(\lambda) = \lambda$, as the next example shows.

Example: a rate that depends on the square of a coupling. Many rates in particle physics are proportional to the square of a coupling g: a decay width, a cross-section, an oscillation

probability. Take $\mu(g) = g^2$ in suitable units, with noise $\sigma = 0.1$, and suppose the true coupling is $g = 0.8$. *The event.* One simulated reading came out as $d = 0.6041$. *The question.* Which couplings could have produced it? A coupling g predicts the rate g^2, so the candidates that explain the reading well are those with g^2 close to d. *The candidates.* There are two groups of them, not one: g near $+\sqrt{d}$ and g near $-\sqrt{d}$. The reason is that g and $-g$ predict the same rate, so no reading can tell them apart. *Only then the numbers.* With a flat prior on $g \in [-1.5, 1.5]$,

$$p(g \mid d) \propto \exp\left[-\frac{(d - g^2)^2}{2\sigma^2}\right].$$

This is largest where $g^2 = d$, that is at the *two* points $g = \pm\sqrt{d} = \pm 0.7772$ (Figure 5.21). The function is even in g, so exactly half the posterior probability lies on each side: the computed mass at $g > 0$ is 0.5. *What it means.* The data say "$|g|$ is about 0.7772" and nothing about the sign. The posterior reports this honestly with two peaks; a single "best value with an error bar" could not. If theory tells us that $g > 0$, that is prior information. Put it in the prior, $p(g) = 0$ for $g < 0$, and the posterior becomes the right-hand peak alone, renormalised.

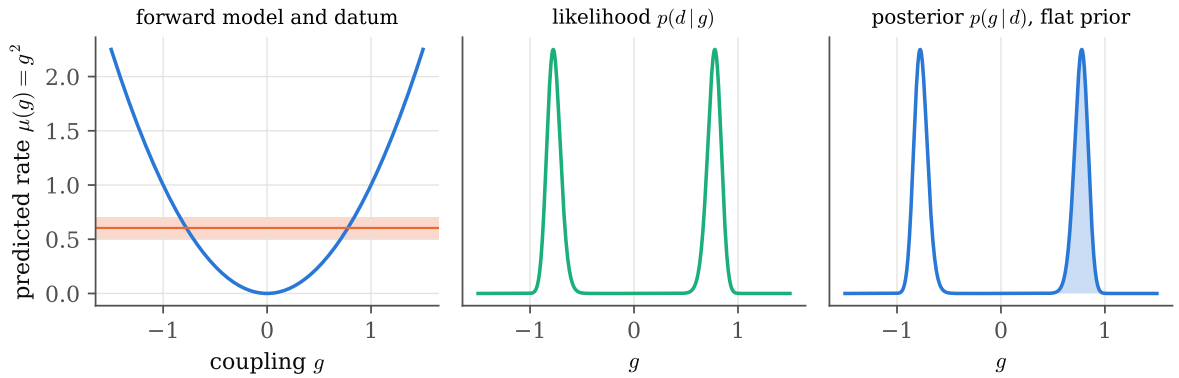


Figure 5.21: Inferring a coupling from a rate proportional to its square. Left: the forward model g^2 and the measured rate with its noise band. Two ranges of g, one on each side of zero, predict rates inside the band. Middle: the likelihood as a function of g has a peak at each of them. Right: with a flat prior the posterior has the same two peaks, and the shaded half, $g > 0$, holds exactly half the probability.

5.7.3 Many data points

A real experiment gives many readings. Let the model predict a whole data vector $\boldsymbol{m}(\lambda) = (m_1(\lambda), \ldots, m_N(\lambda))$ and let the measured values be $\boldsymbol{d} = (d_1, \ldots, d_N)$ with independent Gaussian noise of sizes σ_i. Independence means the joint density of all readings is the product of the single-reading densities [Wasserman p. 177], so

$$\ln \mathcal{L}(\lambda) \overset{(1)}{=} \ln \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{(d_i - m_i(\lambda))^2}{2\sigma_i^2}\right]$$

$$\overset{(2)}{=} -\frac{1}{2}\sum_{i=1}^{N} \frac{(d_i - m_i(\lambda))^2}{\sigma_i^2} - \frac{1}{2}\sum_{i=1}^{N} \ln(2\pi\sigma_i^2) \;=\; -\frac{1}{2}\chi^2(\lambda) + \text{const.} \tag{5.26}$$

(1) The likelihood of independent readings is the product of (5.22) over i.
(2) The log of a product is the sum of the logs, and $\ln e^x = x$.

Here $\chi^2(\lambda) = \sum_i (d_i - m_i(\lambda))^2/\sigma_i^2$ is the sum of squared residuals, each measured in units of its error bar, and the constant does not depend on λ. With correlated noise of covariance matrix $\mathbf{C}$ the product is replaced by one multivariate Gaussian and $\chi^2 = (\boldsymbol{d}-\boldsymbol{m})^{\mathsf{T}}\mathbf{C}^{-1}(\boldsymbol{d}-\boldsymbol{m})$ (Chapter 3). Either way the posterior is

$$p(\lambda \mid \boldsymbol{d}) \propto e^{-\chi^2(\lambda)/2}\, p(\lambda), \tag{5.27}$$

and nothing in the logic has changed. The observed event is now the whole vector $\boldsymbol{d}$. Each hypothesis λ is judged by how plausible the whole vector of fluctuations $\boldsymbol{d} - \boldsymbol{m}(\lambda)$ that it requires would be. Only the model for the probability of the data changes from problem to problem. With a flat prior, $p(\lambda)$ is a constant, so the peak of the posterior is where χ^2 is smallest: the least-squares estimate. Least squares is therefore the special case "flat prior, Gaussian noise" of Bayesian estimation [Sivia §3.5].

The same reasoning works at the scale of a whole survey; only the space of hypotheses grows. In cosmology $\boldsymbol{d}$ may be thousands of band powers of the CMB and λ the handful of parameters of the standard cosmological model (see the cosmology notes). What we observed is the whole data vector $\boldsymbol{d}$, every band power at once. The candidate explanations are the possible values of λ, each one a different cosmological model. For each of them the likelihood asks: had nature picked this model, how plausible would it have been for exactly the data we hold to come out? Bayes' theorem combines these likelihoods with the prior distribution and returns a posterior distribution over the models.

5.7.4 The posterior on a grid

For one or two parameters the safest way to compute a posterior is to evaluate it on a grid of values and look [Sivia §3.4, "brute force and ignorance"]. The grid works whether the posterior is skewed, has several peaks, or has a sharp edge where the prior stops. Its one trap is numerical: a product of many likelihoods is too large or too small for a computer to store.

Example: a lifetime from 2000 readings. A source decays with lifetime τ. We record the signal at $N = 2000$ times, $d_i = e^{-t_i/\tau} + \epsilon_i$ with $\sigma = 0.05$, and put a flat prior on τ. The obvious code multiplies the N Gaussian densities for each grid value of τ. Each density near the curve is about $1/(\sqrt{2\pi}\,\sigma) \approx 8$, and 8^{2000} is far beyond the largest floating-point number, so the product comes out as ∞. Now quote the same data in per cent. Every density is a hundred times smaller, about 0.08, and the product underflows to 0. The posterior cannot depend on the units the data are quoted in, so neither answer means anything. The cure is to work with logarithms [Sivia p. 34]:

1. compute $L_j = \ln \mathcal{L}(\tau_j) + \ln p(\tau_j)$ on the grid, as a *sum* of logs;
2. subtract the largest value, $L_j \to L_j - L_{\max}$, so that the largest is 0;
3. exponentiate, $w_j = e^{L_j - L_{\max}}$, which lies between 0 and 1;
4. normalise, $p(\tau_j \mid \boldsymbol{d}) = w_j / \sum_k w_k \Delta\tau$.

Step 2 is legitimate because it multiplies every w_j by the same constant $e^{-L_{\max}}$, and a common constant cancels in the normalisation of step 4. For our data $L_{\max} = 3149$: the likelihood at the peak is e^{3149}, a number no computer stores, but the *ratios* of likelihoods between grid points are perfectly ordinary numbers.

Listing 5.6: The lifetime posterior on a grid: the raw product fails, the log-sum with the maximum subtracted works (`code/ch05/08_grid_posterior.py`)

```
A, TAU_TRUE, N, SIGD = 1.0, 2.0, 2000, 0.05
t = np.sort(rng.uniform(0, 6, N))
data = A * np.exp(-t / TAU_TRUE) + SIGD * rng.standard_normal(N)
tau = np.linspace(1.9, 2.1, 801)
dtau = tau[1] - tau[0]
model = A * np.exp(-t[None, :] / tau[:, None])                  # (n_tau, N)
dens = np.exp(-(data - model) ** 2 / (2 * SIGD**2)) / np.sqrt(2 * np.pi * SIGD**2)
with np.errstate(over="ignore", under="ignore"):
    raw_max = np.prod(dens, axis=1).max()                       # overflows to inf
    # the same data quoted in per cent: every density is 100 times smaller
    raw_max_pct = np.prod(dens / 100.0, axis=1).max()           # underflows to 0
logL = np.sum(-(data - model) ** 2 / (2 * SIGD**2) - 0.5 * np.log(2 * np.pi * SIGD**2),
    axis=1)
logL_max = logL.max()
w = np.exp(logL - logL_max)                                     # flat prior on tau
post_tau = w / (w.sum() * dtau)
mean_tau = np.sum(tau * post_tau) * dtau
sd_tau = np.sqrt(np.sum((tau - mean_tau) ** 2 * post_tau) * dtau)
```

Line 63 builds the model prediction for every grid value of τ at every time at once, as an 801×2000 array, and line 64 the Gaussian density of every reading. Lines 65–68 form the raw products and show them overflow and, in per cent, underflow. Line 69 is the log-likelihood as a sum, line 71 subtracts the maximum and exponentiates, line 72 normalises, and lines 73–74 compute the posterior mean and standard deviation by summing over the grid. The result is $\tau = 1.996 \pm 0.007948$ (Figure 5.22), consistent with the true value 2 used in the simulation.

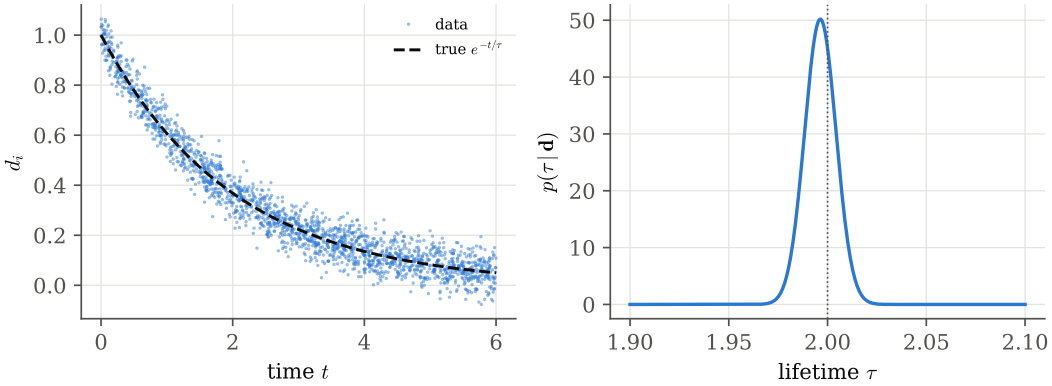


Figure 5.22: A lifetime from 2000 noisy readings. Left: the readings (dots) and the true decay curve (dashed). Right: the posterior for the lifetime computed on a grid of 801 values by summing log-likelihoods and subtracting their maximum. The dotted line marks the value used to simulate the data.

Watch out: Never multiply raw likelihoods

The product of many densities overflows or underflows depending only on the units the data are quoted in. Always sum log-likelihoods and subtract the maximum before exponentiating. The same rule applies to every sampler of Chapter 9: they compare $\ln p$ values, never p values.

A grid fails in many dimensions, because its cost grows as a power of the number of parameters. With n points per axis, a posterior in D parameters needs n^D evaluations: 100 points per axis is 10^4 evaluations in two dimensions and 10^{12} in six, the size of a minimal cosmological model. That cost, and the evidence integral it makes impossible, are what drive us to sampling methods (Chapters 6 and 9).

5.8 Conjugate priors: posteriors on the back of an envelope

For counts of successes, counts of events and Gaussian readings, a suitable choice of prior allows us to work out the posterior on paper. The grid calculation of Section 5.7 represents the posterior as a table of numbers. With a conjugate prior we can instead write it as a member of the same family as the prior and identify exactly which parameters the data have changed. We can then see at a glance how much of the answer comes from the data and how much from the prior. Updating becomes bookkeeping: add the number of heads here, the number of counts there.

The question is: for which pairs of likelihood and prior does the product have the same functional form as the prior, and what are the update rules?

Working definition 5.3 (conjugate prior)

A family of prior densities is **conjugate** to a likelihood if, for every data set, prior from the family × likelihood is again a member of the family [same functional form in the parameter, different values of its constants] [Wasserman p. 179]. Conjugacy is a property of the *pair*: a prior is conjugate only with respect to a particular likelihood [Kruschke §6.2]. *Operationally:* write the likelihood as a function of the parameter, drop every factor that does not contain it, and look for a density with the same θ-dependence.

5.8.1 Coins: the Beta–Binomial pair

A coin, or any process with two outcomes, has unknown probability θ of "success". One toss $y \in \{0, 1\}$ has the Bernoulli probability $\theta^y(1-\theta)^{1-y}$, and N independent tosses with z successes have [Kruschke (6.2)]

$$p(\{y_i\} \mid \theta) \overset{(1)}{=} \prod_{i=1}^{N} \theta^{y_i}(1-\theta)^{1-y_i} \overset{(2)}{=} \theta^{\sum_i y_i}(1-\theta)^{\sum_i (1-y_i)} = \theta^z(1-\theta)^{N-z}.$$

(1) Independent tosses: the joint probability is the product.
(2) Powers of the same base multiply by adding exponents.

As a function of θ this is a power of θ times a power of $1-\theta$. A prior of the same shape is the **Beta distribution** (Chapter 2),

$$p(\theta) = \mathrm{Beta}(\theta \mid a, b) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}, \qquad B(a,b) = \int_0^1 \theta^{a-1}(1-\theta)^{b-1}\mathrm{d}\theta = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}, \tag{5.28}$$

with $a, b > 0$ [Kruschke (6.3)–(6.4)]. The Beta function $B(a,b)$ is only a normalising number. Recall $\Gamma(n) = (n-1)!$ for integers and $\Gamma(x+1) = x\Gamma(x)$ in general. The flat prior is $\mathrm{Beta}(1,1)$. Multiply prior and likelihood:

$$\begin{aligned} p(\theta \mid z, N) &\overset{(1)}{=} \frac{\theta^z(1-\theta)^{N-z}\ \theta^{a-1}(1-\theta)^{b-1}/B(a,b)}{p(z,N)} \overset{(2)}{=} \frac{\theta^{(z+a)-1}(1-\theta)^{(N-z+b)-1}}{B(a,b)\,p(z,N)} \\ &\overset{(3)}{=} \frac{\theta^{(z+a)-1}(1-\theta)^{(N-z+b)-1}}{B(z+a,\ N-z+b)} = \mathrm{Beta}(\theta \mid z+a,\ N-z+b). \end{aligned} \tag{5.29}$$

(1) Bayes' theorem with the Bernoulli likelihood and the Beta prior.
(2) Collect the powers of θ and of $1-\theta$.
(3) The numerator is the θ-dependence of a Beta density with parameters $z+a$ and $N-z+b$. A density must integrate to one, and there is only one constant that makes this one do

so, $B(z+a, N-z+b)$. So the denominator must be that constant, and no integral was needed [Kruschke §6.3].

The update rule is: *add the successes to a and the failures to b*. Adding is done in any order, so the result is the same whether the tosses arrive one at a time or all at once, as sequential updating requires (Section 5.5). The rule also tells us how to read a prior. Beta(a, b) behaves like the posterior we would hold after seeing about a successes and b failures. So $\kappa = a + b$ counts how many tosses' worth of information the prior carries [Kruschke §6.2.1].

The posterior mean is a compromise. The mean of a Beta density is

$$\mathbb{E}[\theta] \overset{(1)}{=} \frac{1}{B(a,b)} \int_0^1 \theta^a (1-\theta)^{b-1} \mathrm{d}\theta \overset{(2)}{=} \frac{B(a+1,b)}{B(a,b)} \overset{(3)}{=} \frac{\Gamma(a+1)\,\Gamma(a+b)}{\Gamma(a)\,\Gamma(a+b+1)} \overset{(4)}{=} \frac{a}{a+b}.$$

(1) Definition of the mean, $\int \theta\, p(\theta) \mathrm{d}\theta$, with one extra power of θ.
(2) The integral is the Beta function with $a \to a+1$.
(3) Write both Beta functions with Gammas. $\Gamma(b)$ cancels.
(4) $\Gamma(a+1) = a\Gamma(a)$ and $\Gamma(a+b+1) = (a+b)\Gamma(a+b)$.

For the posterior (5.29) this gives [Kruschke (6.9)], [Wasserman p. 179]

$$\frac{z+a}{N+a+b} \overset{(1)}{=} \frac{z}{N} \cdot \frac{N}{N+a+b} + \frac{a}{a+b} \cdot \frac{a+b}{N+a+b}. \tag{5.30}$$

(1) Multiply out the right side: the N and the $a+b$ cancel, leaving $(z+a)/(N+a+b)$.

The posterior mean is a weighted average of two numbers: the observed fraction z/N (the maximum likelihood estimate) and the prior mean $a/(a+b)$. Their weights are proportional to the number of tosses N and to the prior's worth in tosses, $\kappa = a+b$. So the data win once $N \gg \kappa$. For a flat prior, $a = b = 1$, it is $(z+1)/(N+2)$, Laplace's rule of succession [Wasserman p. 178, (11.5)–(11.6)]. For example, Kruschke's prior Beta$(5,5)$ with $z = 1$ success in $N = 10$ gives posterior mean $(1+5)/(10+10) = 0.3$, exactly halfway between the prior mean 0.5 and the data fraction 0.1, because $N = \kappa$ [Kruschke Fig. 6.3].

Example: the same data, three priors. *The event.* A two-outcome process is tried 20 times and gives 17 successes, 85%. *The question.* What should we now believe about its success probability θ? The answer depends on what we knew before, that is, on what the process is [Kruschke §6.4.1]. *Three cases in turn.* If it is a freshly minted government coin, we are confident it is fair: prior Beta$(100, 100)$, worth 200 tosses of evidence for 0.5. If a success is a professional basketball player making a free throw, we know such players make about 75% of them: prior Beta$(18.25, 6.75)$, mode 0.75 and $\kappa = 25$. If it is the colour of samples of a substance on a distant planet, we know only that both outcomes are possible: Beta$(1,1)$. *Only then the numbers.* Adding 17 to a and 3 to b, the three posteriors are Beta$(117, 103)$, Beta$(35.25, 9.75)$ and Beta$(18, 4)$, with modes 0.5321, 0.7965 and 0.85 (Figure 5.23). The 17 successes barely move the government coin, pull the basketball player just above his league's typical rate, and fully determine the planetary substance. None of the three analysts is wrong. They started from different knowledge, and twenty tosses are not enough to overwhelm a prior worth two hundred.

Example: predicting the next two tosses. Conjugacy also gives predictions in closed form. A coin gave 10 heads in 14 tosses. What is the probability that the next two tosses are both heads [Padilla §2.3, §3.1]? With a flat prior, Beta$(1,1)$, the posterior is Beta$(1+10, 1+4)$ = Beta$(11, 5)$. If

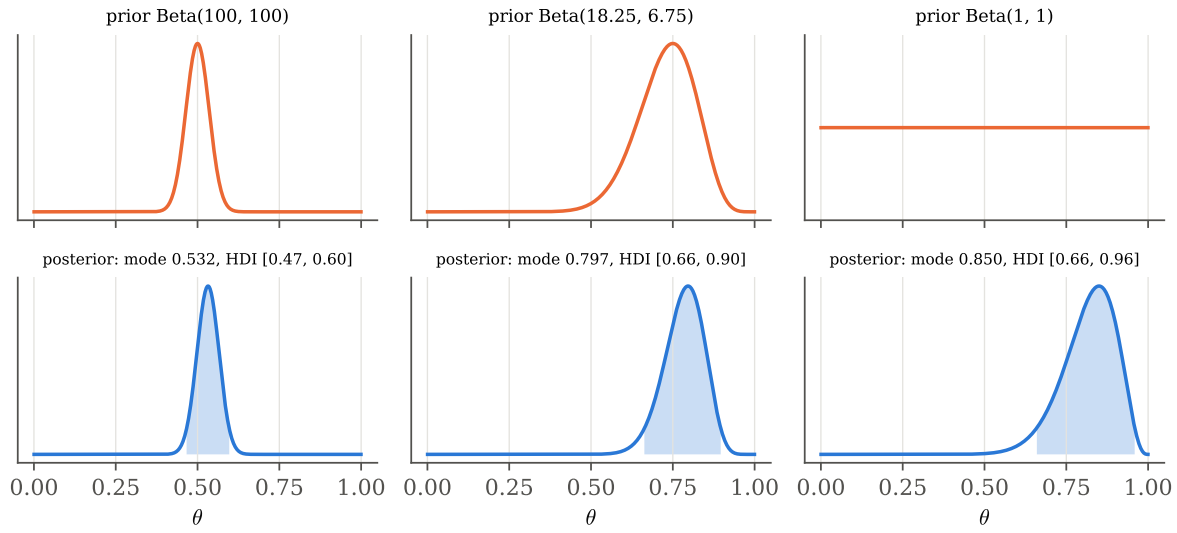


Figure 5.23: Seventeen heads in twenty tosses, analysed with three priors. Top row: a confident fair-coin prior, a basketball league's free-throw rate, and a flat prior. Bottom row: the three posteriors, each a Beta density, with the shaded 95% highest density interval (Section 5.9). The same data move the three beliefs by very different amounts.

we knew θ, two heads would have probability θ^2. We do not know θ, so we average θ^2 over the posterior, that is, we marginalise θ:

$$\mathrm{P}(\mathrm{HH} \mid \boldsymbol{d}) \overset{(1)}{=} \int_0^1 \theta^2 \, \frac{\theta^{10}(1-\theta)^4}{B(11,5)} \, \mathrm{d}\theta \overset{(2)}{=} \frac{B(13,5)}{B(11,5)} \overset{(3)}{=} \frac{12 \cdot 11}{17 \cdot 16} = 0.4853.$$

(1) Sum over the paths "θ, then two heads": the probability of HH given θ times the posterior density of θ.
(2) The integral is a Beta function with the first argument raised by two.
(3) $B(13,5)/B(11,5) = \Gamma(13)\Gamma(16)/[\Gamma(11)\Gamma(18)]$, and $\Gamma(13)/\Gamma(11) = 12 \cdot 11$, $\Gamma(18)/\Gamma(16) = 17 \cdot 16$.

Plugging in the single best value $\hat{\theta} = 10/14$ instead gives $(10/14)^2 = 0.5102$. The posterior predictive is smaller, and it is the honest answer. It remembers that θ might be lower than 10/14; the plug-in pretends that 14 tosses fixed θ exactly.

5.8.2 Counts: the Gamma–Poisson pair

A detector counts cosmic-ray muons in one-minute windows. The counts $n_1, \ldots, n_M$ are Poisson with unknown rate λ per minute, so the likelihood is [Lambert (9.8)]

$$p(\{n_i\} \mid \lambda) \overset{(1)}{=} \prod_{i=1}^{M} \frac{\lambda^{n_i} e^{-\lambda}}{n_i!} \overset{(2)}{=} \frac{\lambda^{\sum_i n_i} \, e^{-M\lambda}}{\prod_i n_i!} \;\propto\; \lambda^{S} e^{-M\lambda}, \qquad S = \sum_i n_i.$$

(1) Independent windows: the product of Poisson probabilities.
(2) Add the exponents of λ, and M copies of $e^{-\lambda}$ make $e^{-M\lambda}$. The factorials do not contain λ.

A power of λ times a decaying exponential in λ is the shape of the **Gamma distribution**, $\mathrm{Gamma}(\lambda \mid \alpha, \beta) = \beta^{\alpha} \lambda^{\alpha-1} e^{-\beta\lambda} / \Gamma(\alpha)$ for $\lambda > 0$, with mean α/β and variance α/β^2 (Chapter 2). Multiplying,

$$p(\lambda \mid \{n_i\}) \overset{(1)}{=} \lambda^{S} e^{-M\lambda} \cdot \lambda^{\alpha-1} e^{-\beta\lambda} \overset{(2)}{=} \lambda^{(\alpha+S)-1} e^{-(\beta+M)\lambda} \;\Rightarrow\; \lambda \mid \{n_i\} \sim \mathrm{Gamma}(\alpha + S,\ \beta + M). \tag{5.31}$$

(1) Posterior $\propto$ likelihood $\times$ prior, dropping constants.
(2) Collect powers and exponents. The normalisation follows as in the Beta case.

The rule is: *add the total count to α and the number of windows to β* [Lambert (9.11)]. So Gamma(α, β) is like having already seen about α counts in β windows, and the posterior mean $(\alpha + S)/(\beta + M)$ is again a weighted average, of the prior mean α/β and the observed mean S/M.

Example: the muon rate. A previous run suggested "about 3 per minute, not very sure": prior Gamma$(3, 1)$, mean 3, standard deviation $\sqrt{3} \approx 1.7$. Five windows gave $3, 5, 7, 4, 9$ counts, total $S = 28$, observed mean 5.6. The posterior is Gamma$(31, 6)$ with mean 5.167 and standard deviation 0.928: pulled a little from the data's 5.6 towards the prior's 3, since the prior is worth one window against the data's five (Figure 5.24, middle).

5.8.3 Gaussian readings: inverse-variance weighting emerges

Several laboratories measure the same quantity μ with different instruments. Reading x_k has Gaussian error σ_k, and we have a Gaussian prior $\mathcal{N}(a, b^2)$ of mean a and width b. We work with the logarithm, because the log of a product of Gaussians is a sum of squares, which is easy to collect. Up to constants that do not contain μ,

$$\begin{aligned} -2\ln p(\mu \mid \{x_k\}) &\overset{(1)}{=} \sum_k \frac{(x_k-\mu)^2}{\sigma_k^2} + \frac{(\mu-a)^2}{b^2} \\ &\overset{(2)}{=} \mu^2\Big(\sum_k w_k + \frac{1}{b^2}\Big) - 2\mu\Big(\sum_k w_k x_k + \frac{a}{b^2}\Big) + \text{const} \\ &\overset{(3)}{=} P\,(\mu - m)^2 + \text{const}, \qquad P = \sum_k w_k + \frac{1}{b^2}, \quad m = \frac{\sum_k w_k x_k + a/b^2}{P}. \end{aligned} \tag{5.32}$$

(1) Log of Gaussian likelihoods times a Gaussian prior, multiplied by -2.
(2) Expand the squares and write $w_k = 1/\sigma_k^2$. Terms without μ go into the constant.
(3) Complete the square: $P\mu^2 - 2\mu P m = P(\mu - m)^2 - Pm^2$, and Pm^2 is a constant.

The posterior is therefore Gaussian, $\mu \mid \{x_k\} \sim \mathcal{N}(m, 1/P)$ [Wasserman p. 179, (11.7)]. Two rules have appeared by themselves.

- *Precisions add.* The precision (inverse variance) of the posterior is the sum of the precisions of the measurements and of the prior.
- *Means are precision-weighted.* Each measurement counts in proportion to $1/\sigma_k^2$, and the prior counts as one more measurement, of value a and error b.

With a very broad prior, $b \to \infty$, we get the familiar inverse-variance weighted mean $\bar{\mu} = \sum_k w_k x_k / \sum_k w_k$ with error $(\sum_k w_k)^{-1/2}$ [Sivia (2.30)–(2.31)]. For N equal errors this is the sample mean with error $\sigma/\sqrt{N}$. Wasserman writes the same result as $m = w\bar{X} + (1 - w)a$ with $w = (1/\mathrm{se}^2)/(1/\mathrm{se}^2 + 1/b^2)$ and $\mathrm{se} = \sigma/\sqrt{n}$ [Wasserman p. 179].

Example: three laboratories. Three simulated measurements are 10.2 ± 0.5, 9.936 ± 1.0 and 10.07 ± 2.0, and a vague earlier estimate gives the prior $\mathcal{N}(8, 3^2)$. The posterior is $\mu = 10.1 \pm 0.4319$. Without the prior, the inverse-variance weighted mean is 10.15 ± 0.4364. The prior counts as one more measurement, of error 3. It shifts the answer by a tenth of its error bar and shrinks the error bar slightly (Figure 5.24, right).

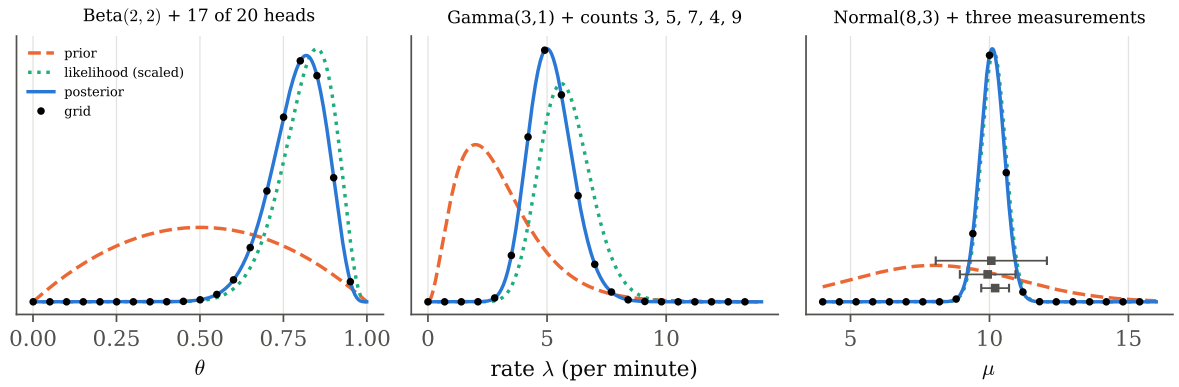


Figure 5.24: The three conjugate pairs. In each panel the dashed curve is the prior, the dotted curve the likelihood (scaled to unit area), the solid curve the closed-form posterior and the dots the posterior computed by brute force on a grid. Left: a Beta prior and 17 heads in 20. Middle: a Gamma prior and five minutes of muon counts. Right: a Gaussian prior and three measurements with different error bars (squares with bars). The dots lie on the curves, to a relative accuracy of 10^{-11}.

The update rules are exact, and the grid check in Figure 5.24 confirms it. The largest relative difference between the brute-force and closed-form posteriors is 1.247×10^{-11} for the Beta case, 1.018×10^{-11} for the Gamma case and 2.260×10^{-14} for the Gaussian case: pure round-off.

Taxonomy: common conjugate pairs

likelihood (data)	conjugate prior	posterior
Bernoulli/binomial (z of N)	Beta(a, b)	Beta$(a + z,\ b + N - z)$
Poisson (total S in M windows)	Gamma(α, β)	Gamma$(\alpha + S,\ \beta + M)$
multinomial (counts $\boldsymbol{n}$)	Dirichlet$(\boldsymbol{\alpha})$	Dirichlet$(\boldsymbol{\alpha} + \boldsymbol{n})$
Gaussian, known σ_k	$\mathcal{N}(a, b^2)$	$\mathcal{N}(m, 1/P)$, (5.32)
Gaussian, known mean, variance v	Inv-Gamma(α, β)	Inv-Gamma$\left(\alpha + \frac{N}{2},\ \beta + \frac{1}{2}\sum(x_i - \mu)^2\right)$

Sources: [Lambert Table 9.1], [Wasserman pp. 178–179]. The last row is the one used for a power-spectrum amplitude below.

The limits of conjugacy. Conjugate pairs exist only for a handful of simple likelihoods, and even then only priors of one particular shape are allowed. A coin factory that makes coins biased either near 0.25 or near 0.75 has a two-humped prior that no Beta density can represent. Its posterior after 14 heads in 27 has three bumps, a compromise between the two humps of the prior and the single peak of the likelihood, and must be computed on a grid [Kruschke §6.4.2]. Likewise, replacing a Gaussian likelihood by a more robust Student-t destroys conjugacy altogether [Lambert §9.7]. The lesson is not to bend the model to fit a conjugate prior. Conjugacy is a source of exact test cases and of intuition. Realistic models are handled by grids in low dimension and by sampling (Chapter 9) in high dimension.

5.8.4 One calculation, three instruments

The same few lines of algebra appear in three very different corners of physics. Only the likelihood changes: counts at a collider are Poisson, a CMB map is a Gaussian random field, and a gravitational-wave strain is a known waveform in Gaussian noise described by a power spectral density (Figure 5.25).

A collider search with no background. A search for a rare process expects s signal events and, after cuts, essentially no background. The number observed is Poisson with mean s. Suppose nothing is seen. How large could s be? The flat prior on $s \geq 0$ is the Gamma$(1,0)$ limit, and the search is a single window, so the Gamma update (5.31) with n events gives $s \mid n \sim \text{Gamma}(n+1,1)$. For $n=0$ the posterior is e^{-s}. The 95% credible upper limit s_{95} is the value below which s lies with probability 95%, so it solves

$$0.95 \overset{(1)}{=} \int_0^{s_{95}} e^{-s}\,\mathrm{d}s \overset{(2)}{=} 1 - e^{-s_{95}} \quad \Rightarrow \quad s_{95} = \ln 20 = 2.996.$$

(1) The posterior probability that s is below the limit is 95%.
(2) Elementary integral.

This is the famous "three events" upper limit for an empty search. With one or three events observed the limits are 4.744 and 7.754.

The CMB: a power-spectrum amplitude from $2\ell+1$ numbers. The temperature map is expanded in spherical harmonics, and at multipole ℓ the coefficients $a_{\ell m}$ amount to $2\ell+1$ independent real Gaussian numbers of zero mean and variance C_ℓ (see the cosmology notes, and Chapter 7). Their likelihood is

$$p(\{a_{\ell m}\} \mid C_\ell) \overset{(1)}{=} \prod_{m=-\ell}^{\ell} \frac{e^{-|a_{\ell m}|^2/2C_\ell}}{\sqrt{2\pi C_\ell}} \;\propto\; C_\ell^{-(2\ell+1)/2} \exp\!\left[-\frac{(2\ell+1)\widehat{C}_\ell}{2C_\ell}\right], \qquad \widehat{C}_\ell = \frac{1}{2\ell+1}\sum_m |a_{\ell m}|^2.$$

(1) $2\ell+1$ independent Gaussians of variance C_ℓ, and the sum of their squares written through the estimator $\widehat{C}_\ell$.

As a function of C_ℓ this is an inverse power times $e^{-\text{const}/C_\ell}$. That is the shape of an inverse-Gamma density, the last row of the table. With the scale-invariant prior $p(C_\ell) \propto 1/C_\ell$ (Section 5.12) the posterior is Inv-Gamma$\big(\frac{2\ell+1}{2}, \frac{(2\ell+1)\widehat{C}_\ell}{2}\big)$. For the quadrupole, $\ell=2$, only five numbers exist in the whole sky. The posterior of $C_\ell/\widehat{C}_\ell$ has its mode at 0.7143, its mean at 1.667, and its central 68% interval from 0.6303 to 2.421. So five numbers cannot pin the quadrupole power to better than a factor of about two, however good the telescope. This is cosmic variance seen as a posterior. At $\ell = 50$ the interval has shrunk to $[0.8776, 1.162]$ and the posterior is nearly Gaussian.

A gravitational-wave amplitude. A detector records $d = A\,h_0 + n$, where h_0 is a known waveform, A an unknown amplitude and n stationary Gaussian noise with power spectral density $S_n(f)$. The likelihood is $\mathcal{L} \propto \exp[-\tfrac{1}{2}(d - Ah_0 \mid d - Ah_0)]$ with the noise-weighted inner product $(a|b) = 4\,\mathrm{Re}\int_0^\infty \tilde{a}\tilde{b}^*/S_n\,\mathrm{d}f$ (Chapter 3). The inner product is linear in each slot, so

$$(d - Ah_0 \mid d - Ah_0) \overset{(1)}{=} (d|d) - 2A\,(d|h_0) + A^2\,(h_0|h_0) \overset{(2)}{=} (h_0|h_0)\,\big(A - \hat{A}\big)^2 + \text{const},$$
$$\hat{A} = \frac{(d|h_0)}{(h_0|h_0)}.$$

(1) Expand, using $(h_0|d) = (d|h_0)$ for the real inner product.
(2) Complete the square in A.

The posterior for A with a flat or Gaussian prior is Gaussian, centred on the matched-filter estimate $\hat{A}$, with standard deviation $1/\sqrt{(h_0|h_0)}$. For our toy chirp in coloured noise, normalised to $(h_0|h_0) = 100$, the posterior is $A = 1.06 \pm 0.1$ for true $A = 1$, and over 4000 independent noise realisations the estimate $\hat{A}$ scatters by 0.1008, the posterior width, as it should.

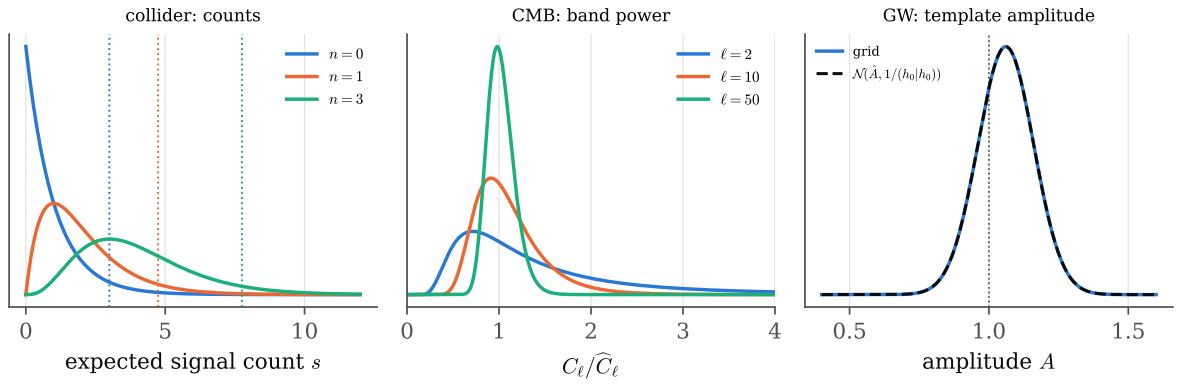


Figure 5.25: One posterior calculation in three instruments. Left: a background-free collider search, posterior for the expected signal after 0, 1 and 3 observed events, with the 95% upper limits dotted. Middle: the posterior of a CMB band power relative to its estimate, for $\ell = 2$, 10 and 50. The quadrupole's posterior is broad and skewed. Right: the amplitude of a known gravitational-wave template in coloured noise, computed on a grid (solid) and from the closed form (dashed).

How this is used in real research: closed forms inside large analyses

Conjugate results survive as building blocks inside big pipelines. Collider limits for counting experiments are routinely quoted as Bayesian upper limits with a flat prior on the signal rate, exactly the Gamma calculation above, and physicists check them against frequentist constructions (Chapter 4). CMB likelihoods at low ℓ use the exact inverse-Gamma shape rather than a Gaussian, because the quadrupole's posterior is far from Gaussian (Chapter 7). Gravitational-wave pipelines marginalise analytically over parameters that enter linearly, such as an amplitude or a phase, before sampling the rest (Chapter 9). Everywhere, the closed forms are also the test cases that a new sampler must reproduce.

5.9 Summarising a posterior: best values and credible intervals

The posterior is the complete answer: every statement we can make about the parameter is an integral of it. But papers quote only a handful of numbers, such as "$\theta = 0.28 \pm 0.08$" or "$s < 3$ events at 95%". A decision, such as building the next detector or claiming a discovery, also needs one number or one interval. So which numbers should we quote, and what exactly does each one mean?

Our running example is the posterior of the CMB quadrupole power C_2 from Section 5.8. There the five numbers a_{2m} ($m = -2, \dots, 2$) measured over the whole sky gave an inverse-Gamma posterior. Write it in terms of the ratio $x = C_2/\widehat{C}_2$ of the power to its estimate $\widehat{C}_2 = \frac{1}{5}\sum_m |a_{2m}|^2$. Then its parameters are $\alpha = \beta = 5/2$, and it is strongly skewed, with a long tail to large x (Figure 5.27, left).

5.9.1 Three best values, three losses

There are three standard point summaries [Lambert §7.6]: the **posterior mean** $\mathbb{E}[\theta \mid \boldsymbol{d}] = \int \theta\, p(\theta \mid \boldsymbol{d}) \mathrm{d}\theta$ [Wasserman p. 177, (11.4)], the **posterior median** (half the probability on each side), and the **posterior mode** or MAP, *maximum a posteriori*, the highest point of the density. For a

symmetric single-peaked posterior they coincide. For the quadrupole they are

$$\text{mode } 0.7143, \qquad \text{median } 1.149, \qquad \text{mean } 1.667.$$

Which one is "right"? Each is the best answer to a different question: what does it cost us to be wrong? Suppose we must announce one number M, and we pay a penalty $\text{loss}(\theta - M)$ if the truth turns out to be θ. We do not know θ, so we average the penalty over the posterior. The expected penalty is $R(M) = \int \text{loss}(\theta - M)\, p(\theta \mid \boldsymbol{d})\, \mathrm{d}\theta$, and the best M is the one that makes it smallest [Kruschke §4.3.3]. We take three penalties in turn.

Squared loss gives the mean. With $\text{loss} = (\theta - M)^2$,

$$\frac{\mathrm{d}R}{\mathrm{d}M} \overset{(1)}{=} \int \frac{\partial}{\partial M}(\theta - M)^2\, p(\theta \mid \boldsymbol{d})\, \mathrm{d}\theta \overset{(2)}{=} -2 \int (\theta - M)\, p(\theta \mid \boldsymbol{d})\, \mathrm{d}\theta \overset{(3)}{=} -2\big(\mathbb{E}[\theta \mid \boldsymbol{d}] - M\big),$$

(1) Differentiate under the integral, since the limits do not depend on M.
(2) $\partial_M(\theta - M)^2 = -2(\theta - M)$.
(3) Split the integral: $\int \theta p\, \mathrm{d}\theta$ is the mean and $\int M p\, \mathrm{d}\theta = M$ because p is normalised.

which vanishes at $M = \mathbb{E}[\theta \mid \boldsymbol{d}]$. A large miss costs the square of its size, so the rare values far out in a long tail weigh heavily, and they drag the answer towards that tail.

Absolute loss gives the median. With $\text{loss} = |\theta - M|$,

$$\frac{\mathrm{d}R}{\mathrm{d}M} \overset{(1)}{=} \int_{-\infty}^{M} p(\theta \mid \boldsymbol{d})\, \mathrm{d}\theta - \int_{M}^{\infty} p(\theta \mid \boldsymbol{d})\, \mathrm{d}\theta \overset{(2)}{=} \mathrm{P}(\theta < M \mid \boldsymbol{d}) - \mathrm{P}(\theta > M \mid \boldsymbol{d}),$$

(1) $\partial_M|\theta - M|$ is $+1$ for $\theta < M$ and -1 for $\theta > M$.
(2) Each integral of the density over a half-line is a probability.

which vanishes when half the probability lies on each side: the median.

"All or nothing" gives the mode. If we win only when M lands within a tiny window $\pm\delta$ of the truth, the expected reward is $\int_{M-\delta}^{M+\delta} p\, \mathrm{d}\theta \approx 2\delta\, p(M \mid \boldsymbol{d})$, largest where the density is highest: the MAP.

The mode is the easiest to compute. The evidence does not depend on θ, so the mode is simply the maximum of likelihood times prior, and no normalisation is needed. With a flat prior it is the maximum-likelihood estimate. But it is the most fragile of the three, for two reasons. First, it depends on the parametrisation: a density changes shape when the axis is stretched (Section 5.12). Second, for a skewed posterior it can sit far from the bulk of the probability. For the quadrupole, 78% of the posterior probability lies *above* the mode.

The median does not have the first weakness. If $y = g(\theta)$ is increasing, the median of y is g of the median of θ. The reason is that "half the probability lies below" is a statement about the probabilities of corresponding intervals, and a change of variable does not alter those. Lambert ranks the mean first, the median second and the MAP last [Lambert §7.6]. Whichever is quoted, say which one it is, and give an interval with it.

5.9.2 Credible intervals

A **credible interval** of level $1 - \alpha$ is any region holding posterior probability $1 - \alpha$. Many regions do so, and two conventions pick one of them.

The **equal-tailed interval** (ETI) puts probability $\alpha/2$ in each tail: find a, b with $\int_{-\infty}^{a} p\,\mathrm{d}\theta = \int_{b}^{\infty} p\,\mathrm{d}\theta = \alpha/2$ [Wasserman p. 178]. It is the interval between two quantiles, easy to compute from a cumulative distribution or from samples (Section 5.14), and, like the median, it maps onto itself under any increasing change of variable.

The **highest density interval** (HDI) instead takes the most credible values first [Kruschke §4.3.4]. Picture the posterior curve as a landscape, with a water level W that is lowered from above. The HDI is the set of θ that stick out above the water, $p(\theta \mid \boldsymbol{d}) > W$, with W lowered until that set holds probability $1 - \alpha$ (Figure 5.26).

Working definition 5.4 (highest density interval)

The $(1-\alpha)$ **HDI** is the set $\{\theta : p(\theta \mid \boldsymbol{d}) > W\}$, where the level W is chosen so that $\int_{p>W} p(\theta \mid \boldsymbol{d})\,\mathrm{d}\theta = 1 - \alpha$. Every point inside has higher density than every point outside [so it contains the most credible values]. For a single-peaked posterior it is the *shortest* interval with probability $1 - \alpha$ [Sivia (2.21)]. For a posterior with several peaks it can be several disjoint intervals. *Operationally:* on a grid, sort the grid values by density and accumulate them from the top until the probability reaches $1 - \alpha$. From samples, take the shortest interval that contains a fraction $1 - \alpha$ of the sorted samples.

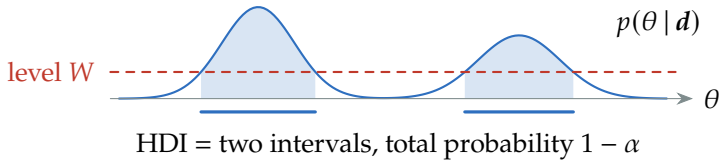


Figure 5.26: The highest density interval as a water level. The level W is lowered until the parts of the posterior above it (shaded) hold probability $1 - \alpha$. The threshold selects points by posterior density, so even the lowest density inside the shaded region is at least as high as any density outside it. For the two peaks shown, the same threshold selects two disjoint intervals.

Why the HDI is the shortest interval. For a single-peaked density, look for the shortest interval $[a, b]$ with fixed probability $F(b) - F(a) = 1 - \alpha$, where F is the cumulative distribution. Minimise $b - a$ with a Lagrange multiplier η for the constraint:

$$0 \overset{(1)}{=} \frac{\partial}{\partial a}\Big[(b-a) - \eta\big(F(b) - F(a) - (1-\alpha)\big)\Big] \overset{(2)}{=} -1 + \eta\, p(a), \qquad 0 \overset{(3)}{=} \frac{\partial}{\partial b}\Big[\cdots\Big] = 1 - \eta\, p(b),$$

(1) Stationarity of the Lagrangian with respect to the lower end.
(2) $\partial_a F(a) = p(a)$, the density at the lower end.
(3) The same with respect to the upper end.

So $p(a) = p(b) = 1/\eta$: the shortest interval has the same density at both ends. That is exactly the water level picture, with the level at height $1/\eta$. For a skewed posterior the two conventions therefore differ. The HDI is shifted towards the peak and has unequal tails. The ETI has equal tails and unequal densities at its ends.

Example: the quadrupole. Recall that $x = C_2/\widehat{C}_2$ is the quadrupole power relative to its estimate. Its 95% ETI is $[0.3896, 6.015]$ and its 95% HDI is $[0.2294, 4.389]$, at water level $W = 0.02375$ (Figure 5.27, left). The HDI starts lower and is shorter because it includes values of x between 0.2294 and 0.3896 that have higher density than values near 6.015, yet the ETI leaves out the former and keeps the latter [Kruschke §12.1.2.2]. From 10^5 posterior samples, the shortest interval containing 95% of them is $[0.2275, 4.356]$, the same up to sampling noise.

The HDI has a price: it depends on the variable it is computed in. Compute the HDI of the same posterior in the variable $y = \ln x$ and map its ends back to x. It becomes $[0.348, 5.054]$, a different interval. The reason is that the HDI is built from density heights, and heights change when the axis is stretched. The ETI is built from probabilities of intervals, so it is the same whichever variable it is computed in. Kruschke still prefers the HDI whenever the parameter has a natural scale, because on that scale it is the interval of the most credible values.

Example: two peaks. The coupling posterior of Section 5.7 has two peaks at $g = \pm 0.7772$. Its 68% HDI is two intervals, $[-0.8394, -0.7096] \cup [0.7096, 0.8394]$ (Figure 5.27, right). Its posterior mean is 0.000, in the gap between the peaks, where the posterior density is essentially zero. So the mean is the one value of g that the data rule out most firmly [Sivia §2.2, Fig. 2.6]. When a posterior has several peaks, no single number is an honest summary. Quote each peak with its own interval, or show the curve. Lambert's pirate makes the same point with money: if treasure is buried along a beach with a two-humped posterior, digging up the split HDI costs less than digging the whole central interval for the same chance of finding the gold [Lambert §7.7.2].

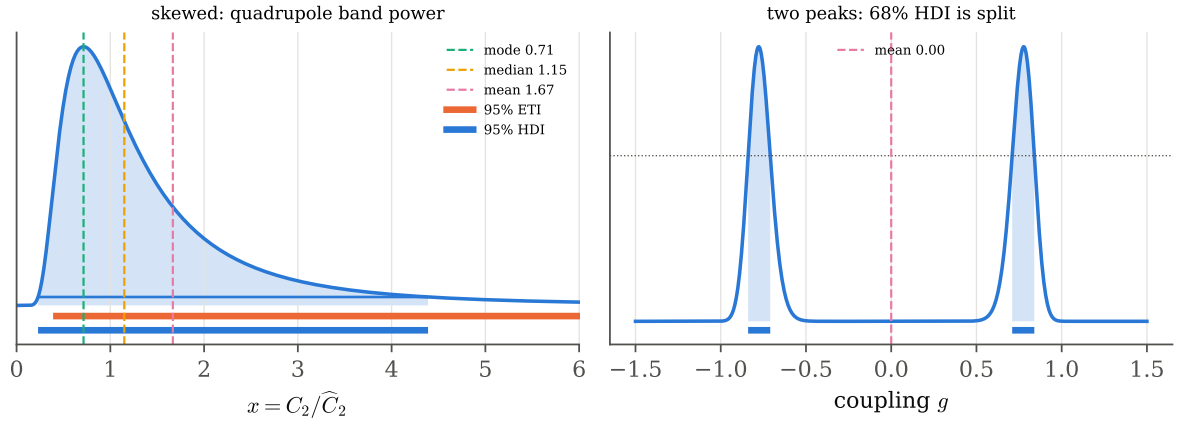


Figure 5.27: Summaries of two posteriors. Left: the skewed posterior of the quadrupole power relative to its estimate, with its mode, median and mean (vertical dashed lines), the 95% equal-tailed interval (upper bar, orange) and the 95% HDI (lower bar, blue, shaded, with its water level). Right: the two-peaked coupling posterior. Its 68% HDI is two intervals, and its mean falls in the gap between the peaks.

5.9.3 What a credible interval says, and what a confidence interval says

A 95% credible interval makes a direct statement about the parameter: given these data and this prior, the probability that θ lies in the interval is 0.95. A 95% confidence interval (Chapters 3 and 4) makes a statement about a procedure: if the experiment were repeated many times, intervals built this way would cover the fixed true value in 95% of the repetitions [Lambert §7.7]. The two are different statements and need not agree. Often, though, the numbers are the same. For a Gaussian likelihood and a flat or broad prior the posterior is $\mathcal{N}(\bar{\theta}, \tau^2)$ with $\bar{\theta} \approx \hat{\theta}$ and $\tau \approx \mathrm{se}$, where $\hat{\theta}$ is the maximum-likelihood estimate and se its standard error. So the 95% credible interval $\bar{\theta} \pm 1.96\tau$ is approximately the frequentist $\hat{\theta} \pm 1.96\,\mathrm{se}$ [Wasserman p. 179]. For large samples this agreement is general (Section 5.10). For small samples, skewed posteriors, parameters near a physical boundary, or informative priors, they can differ substantially, and in high dimensions a 95% credible region need not cover the truth 95% of the time [Wasserman p. 189].

Watch out: Credible is not confidence

"The parameter lies in $[a, b]$ with 95% probability" is a correct reading of a credible interval and an incorrect reading of a confidence interval. Conversely, "this interval covers the truth in 95% of repeated experiments" is guaranteed for a confidence interval and not, in general, for a credible interval. When a result is quoted, say which kind it is, and for a credible interval say which prior was used and whether it is an HDI or an equal-tailed interval.

5.10 The Gaussian approximation: a peak and its curvature

Most posteriors in a mature experiment look like a single smooth bump. Then two numbers describe the whole curve well: where the bump is and how wide it is. The result is quoted as "$\theta = \theta_0 \pm \sigma$". Where do these two numbers come from, in one parameter and in many? And when does this description mislead?

5.10.1 One parameter

We work with the logarithm of the posterior, $L(\theta) = \ln p(\theta \mid \boldsymbol{d})$, for two reasons. It turns products of likelihoods into sums. And near a peak it varies much more gently than the posterior itself, so a short Taylor series describes it well [Sivia (2.10)]. Let θ_0 be the peak, so $L'(\theta_0) = 0$ and $L''(\theta_0) < 0$, and Taylor-expand about it:

$$\begin{aligned} L(\theta) &\overset{(1)}{=} L(\theta_0) + L'(\theta_0)(\theta - \theta_0) + \tfrac{1}{2}L''(\theta_0)(\theta - \theta_0)^2 + \dots \\ &\overset{(2)}{=} L(\theta_0) - \frac{(\theta - \theta_0)^2}{2\sigma^2} + \dots, \qquad \sigma \equiv \left(-L''(\theta_0)\right)^{-1/2}. \end{aligned} \tag{5.33}$$

(1) Taylor's theorem to second order about θ_0.
(2) The linear term vanishes at a maximum, and $L''(\theta_0)$ is negative, so we may write it as $-1/\sigma^2$.

Exponentiate and drop the higher terms. The exponential of $-(\theta - \theta_0)^2/2\sigma^2$ is a Gaussian, so $p(\theta \mid \boldsymbol{d}) \approx \mathcal{N}(\theta_0, \sigma^2)$ [Sivia (2.13)–(2.15)]. This is the **Gaussian** or **Laplace approximation**. The approximate posterior standard deviation is the inverse square root of the curvature of the negative log-posterior at the mode. A sharper peak therefore gives a narrower credible interval. Within the approximation, $\theta_0 \pm \sigma$ holds 68% of the probability and $\theta_0 \pm 2\sigma$ holds 95%.

Example: Gaussian readings. For N readings of equal error σ_x and a flat prior, $L = \text{const} - \sum_k (x_k - \mu)^2/2\sigma_x^2$. Then $L' = \sum_k (x_k - \mu)/\sigma_x^2 = 0$ gives $\mu_0 = \bar{x}$, and $L'' = -N/\sigma_x^2$ gives $\sigma = \sigma_x/\sqrt{N}$ [Sivia (2.28)–(2.29)]. Here the expansion is exact, because L is a quadratic and all higher derivatives vanish.

Example: the coin. With a flat prior and R heads in N tosses, $L = \text{const} + R \ln\theta + (N-R)\ln(1-\theta)$ [Sivia (2.17)]. Then

$$\begin{aligned} L'(\theta) &\overset{(1)}{=} \frac{R}{\theta} - \frac{N-R}{1-\theta} = 0 \;\Rightarrow\; \theta_0 = \frac{R}{N}, \\ L''(\theta_0) &\overset{(2)}{=} -\frac{R}{\theta_0^2} - \frac{N-R}{(1-\theta_0)^2} \overset{(3)}{=} -\frac{N}{\theta_0} - \frac{N}{1-\theta_0} \overset{(4)}{=} -\frac{N}{\theta_0(1-\theta_0)}, \end{aligned}$$

(1) Differentiate the logs, and solve $R(1-\theta_0) = (N-R)\theta_0$.

(2) Differentiate once more.
(3) Substitute $R = N\theta_0$ and $N - R = N(1-\theta_0)$.
(4) Common denominator: $\frac{1}{\theta_0} + \frac{1}{1-\theta_0} = \frac{1}{\theta_0(1-\theta_0)}$.

so $\sigma = \sqrt{\theta_0(1-\theta_0)/N}$ [Sivia (2.19)–(2.20)], the familiar binomial error. The width shrinks like $1/\sqrt{N}$. At fixed N it is largest for $\theta_0 = \frac{1}{2}$, so it is easier to show that a coin is strongly biased than to show that it is fair.

How good is the approximation? It depends on the data. For 9 heads in 32 tosses it gives $\theta = 0.2812 \pm 0.07948$. The exact posterior is a Beta density (Section 5.8) with mean 0.2941 and standard deviation 0.07702, so the two agree well (Figure 5.28, left). For 1 head in 10 the approximation gives 0.1 ± 0.09487. It puts 0.1459 of its probability at $\theta < 0$, which is impossible. The exact posterior is squeezed against the boundary $\theta = 0$ and skewed, with mean 0.1667. So near a physical boundary, and with few data, an error bar is the wrong summary: quote an interval from the posterior itself (Section 5.9).

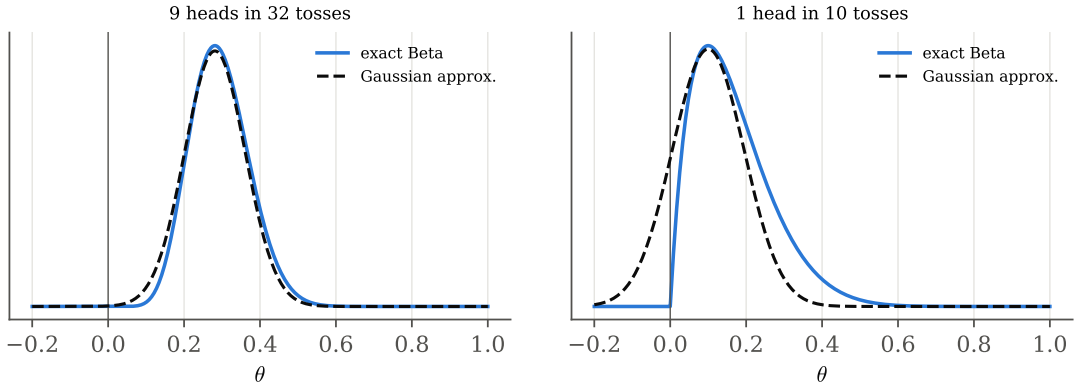


Figure 5.28: The Gaussian approximation (dashed) against the exact posterior (solid) for a coin with a flat prior. Left: 9 heads in 32 tosses, where the two nearly agree. Right: 1 head in 10 tosses, where the exact posterior is pressed against the boundary $\theta = 0$ and the Gaussian spills probability into the impossible region to its left.

5.10.2 Several parameters: the Hessian and its inverse

With parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_M)$ the same expansion about the peak $\boldsymbol{\theta}_0$ reads [Sivia (3.30)–(3.31)]

$$L(\boldsymbol{\theta}) \approx L(\boldsymbol{\theta}_0) + \tfrac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\mathsf{T}} \mathbf{H} (\boldsymbol{\theta} - \boldsymbol{\theta}_0), \qquad H_{ij} = \left.\frac{\partial^2 L}{\partial\theta_i\,\partial\theta_j}\right|_{\boldsymbol{\theta}_0}, \tag{5.34}$$

where the linear term has again vanished because $\boldsymbol{\theta}_0$ is the peak. The matrix $\mathbf{H}$ of second derivatives is the **Hessian**. Exponentiating, the posterior is approximately a multivariate Gaussian with covariance matrix $\mathbf{C} = -\mathbf{H}^{-1}$ [Sivia (3.32)], just as $\sigma^2 = -1/L''$ in one dimension. Its contours are ellipses (ellipsoids in more dimensions). Their axes point along the eigenvectors of $\mathbf{H}$, and their lengths are proportional to $|h_i|^{-1/2}$, where h_i are the eigenvalues of $\mathbf{H}$ [Sivia Fig. 3.6]. The peak really is a maximum only if all eigenvalues are negative.

Example: two parameters, done by hand. Write the second derivatives at the peak as $A = \partial_X^2 L$, $B = \partial_Y^2 L$, $C = \partial_X \partial_Y L$ [Sivia (3.19)]. A maximum needs $A < 0$, $B < 0$ and $AB > C^2$. The inverse of

a 2×2 matrix is $\left(\begin{smallmatrix} A & C \\ C & B \end{smallmatrix}\right)^{-1} = \frac{1}{AB-C^2}\left(\begin{smallmatrix} B & -C \\ -C & A \end{smallmatrix}\right)$, so

$$\mathbf{C} = -\mathbf{H}^{-1} \overset{(1)}{=} \frac{1}{C^2 - AB}\begin{pmatrix} B & -C \\ -C & A \end{pmatrix}, \qquad \sigma_X^2 \overset{(2)}{=} \frac{B}{C^2-AB}, \quad \sigma_Y^2 = \frac{A}{C^2-AB}, \quad \sigma_{XY} \overset{(3)}{=} \frac{C}{AB-C^2}. \tag{5.35}$$

(1) The 2×2 inverse, with the overall minus sign absorbed into the denominator.
(2) The diagonal entries are the variances. Both B and $C^2 - AB$ are negative, so $\sigma_X^2 > 0$ [Sivia (3.21)–(3.22)].
(3) The off-diagonal entry is the covariance [Sivia (3.27)].

The correlation coefficient is $\rho = \sigma_{XY}/(\sigma_X\sigma_Y) = C/\sqrt{AB}$.

There are two different error bars for X, and they answer different questions. The **marginal** error σ_X of (5.35) answers: how far can X move if Y may be anything the data permit (Section 5.11)? The **conditional** error answers: how far can X move if Y is held fixed at its best value? With Y fixed, only the curvature in X matters, so $\sigma_{X|Y} = (-A)^{-1/2}$. How much smaller is the second? Their ratio is

$$\frac{\sigma_X^2}{\sigma_{X|Y}^2} \overset{(1)}{=} \frac{B}{C^2-AB}(-A) \overset{(2)}{=} \frac{AB}{AB-C^2} \overset{(3)}{=} \frac{1}{1-\rho^2} \geq 1.$$

(1) Divide (5.35) by $1/(-A)$.
(2) Multiply numerator and denominator by -1.
(3) Divide numerator and denominator by AB and use $\rho^2 = C^2/(AB)$.

So holding a correlated parameter fixed always makes the error bar look smaller, by a factor $\sqrt{1-\rho^2}$. With $\rho = 0.95$ the conditional error is a third of the honest one [Sivia Fig. 3.8]. Figure 5.29 shows the geometry. The conditional error is the width of a horizontal slice through the centre of the ellipse. The marginal error is the width of the ellipse's shadow on the X axis.

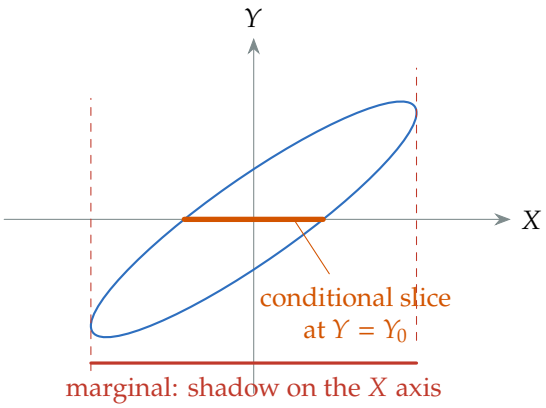


Figure 5.29: Conditional against marginal error for two correlated parameters. The ellipse is a contour of the Gaussian posterior. Holding Y at its best value and asking how far X may move gives the short orange slice, of half-width $(-A)^{-1/2}$. Letting Y take any value gives the full shadow of the ellipse on the X axis, between the red dashed tangents, of half-width proportional to $\sqrt{B/(C^2-AB)}$.

Key idea: error bars from the curvature

Near its peak, the log-posterior is a quadratic form with Hessian $\mathbf{H}$. The posterior is then approximately Gaussian with covariance $\mathbf{C} = -\mathbf{H}^{-1}$. The honest (marginal) error bar on θ_i is $\sqrt{\mathbf{C}_{ii}} = \sqrt{[(-\mathbf{H})^{-1}]_{ii}}$, the square root of a diagonal element of the *inverse*. The inverse of a

diagonal element, $1/\sqrt{-H_{ii}}$, is the conditional error with all other parameters frozen, and is always smaller or equal.

The link to the Fisher matrix. The Laplace approximation and the Fisher forecast of Chapter 10 are the same object, seen after and before the data are taken. Here is why. With a flat prior, $-\mathbf{H}$ is the curvature of the log-likelihood at its maximum, called the *observed information*. Average it over all the data sets the model could have produced, and we get the **Fisher matrix** $\mathbf{F}$, with $F_{ij} = -\mathbb{E}[\partial_i \partial_j \ln \mathcal{L}]$. So $\mathbf{F}^{-1}$ is the expected posterior covariance. Then $\sqrt{(\mathbf{F}^{-1})_{ii}}$ is the expected marginal error, and $1/\sqrt{F_{ii}}$ the expected conditional one [Padilla §3.8].

For a Gaussian likelihood with a model linear in the parameters, L is exactly quadratic. The Hessian then does not depend on the parameters, and the approximation is exact [Sivia §3.4, (3.52)]. For nonlinear models the peak itself is usually found step by step, by Newton–Raphson steps $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \mathbf{H}^{-1}\nabla L$ [Sivia (3.55)], as in Chapter 3.

Example: contours for two parameters. Which contour of a two-parameter posterior holds 68% of the probability? Measure the distance from the peak by $Q = (\boldsymbol{\theta}-\boldsymbol{\theta}_0)^\mathsf{T}\mathbf{C}^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)$. Within the Gaussian approximation, $Q = 2[L(\boldsymbol{\theta}_0) - L(\boldsymbol{\theta})]$, the "$\Delta\chi^2$" from the peak. The contours are the curves $Q = k$, so we need the distribution of Q. Change variables to $\boldsymbol{u} = \mathbf{C}^{-1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$. This turns the Gaussian into M independent standard normals, so $Q = \sum_i u_i^2$ has a χ^2 distribution with M degrees of freedom (Chapter 2). For $M = 2$ that distribution is an exponential, $p(Q) = \frac{1}{2}e^{-Q/2}$, so

$$\mathrm{P}(Q \le k) \overset{(1)}{=} \int_0^k \tfrac{1}{2}e^{-Q/2}\mathrm{d}Q \overset{(2)}{=} 1 - e^{-k/2}$$

$$\Rightarrow \quad k_{68.3\%} = -2\ln(1-0.683) = 2.30, \qquad k_{95.4\%} = 6.17.$$

(1) The χ^2_2 density is $\frac{1}{2}e^{-Q/2}$.
(2) Elementary integral, then solve for k.

These are the contour levels "$\Delta\chi^2 = 2.30$ and 6.17" used in every two-parameter plot in cosmology [Padilla Table III]. In one dimension the same argument gives $\Delta\chi^2 = 1$ and 4 for $\pm1\sigma$ and $\pm2\sigma$: a one-parameter interval and a two-parameter contour at the same confidence need different thresholds.

5.10.3 When the approximation fails, and when it comes back: the lighthouse

A lighthouse stands at unknown position α along a straight coast, a known distance β out to sea [Sivia §2.4]. It emits short flashes in random directions, uniformly in the azimuth $\vartheta \in (-\pi/2, \pi/2)$. Detectors along the shore record *where* each flash arrives, $x = \alpha + \beta\tan\vartheta$, but not its direction (Figure 5.30). Where is the lighthouse?

The likelihood of one arrival point follows from the change of variables for densities (Chapter 1):

$$p(x \mid \alpha, \beta) \overset{(1)}{=} p(\vartheta)\left|\frac{\mathrm{d}\vartheta}{\mathrm{d}x}\right| \overset{(2)}{=} \frac{1}{\pi}\cdot\frac{1}{\beta\,(1+\tan^2\vartheta)} \overset{(3)}{=} \frac{\beta}{\pi\,[\beta^2 + (x-\alpha)^2]}.$$

(1) The probability in a small range $\mathrm{d}\vartheta$ equals that in the matching $\mathrm{d}x$.
(2) $p(\vartheta) = 1/\pi$ on $(-\pi/2, \pi/2)$, and $\mathrm{d}x/\mathrm{d}\vartheta = \beta/\cos^2\vartheta = \beta(1+\tan^2\vartheta)$.
(3) $\tan\vartheta = (x-\alpha)/\beta$.

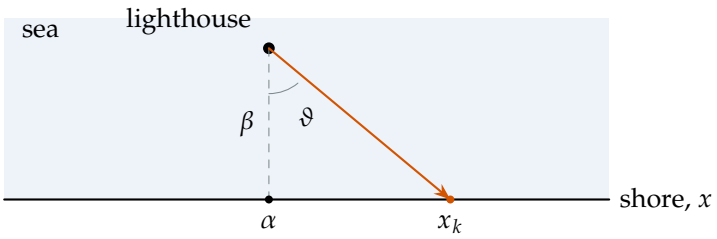


Figure 5.30: The lighthouse problem. A flash leaves at azimuth ϑ, uniformly distributed, and reaches the shore at $x_k = \alpha + \beta \tan \vartheta$. Only the arrival point is recorded.

This is a Cauchy (Lorentzian) density [Sivia (2.34)], with a peak at α and tails so heavy that its variance is infinite (Chapter 2). Take a flat prior on α and independent flashes at $x_1, \ldots, x_N$. The log-posterior is $L(\alpha) = \text{const} - \sum_k \ln[\beta^2 + (x_k - \alpha)^2]$ [Sivia (2.38)]. The condition $L' = 0$ for its peak cannot be solved in closed form, but on a grid the whole posterior is easy to compute.

Figure 5.31 shows the posterior as flashes accumulate, for a lighthouse at $\alpha = 1$ km with $\beta = 1$ km. After one or two flashes the posterior is broad with heavy tails. When two flashes land far apart, the posterior has two humps, one near each. In this run, by eight flashes it is a single bump. After 64 its standard deviation, 0.2039, agrees with the Laplace width 0.2026 from the curvature at the peak. After 512 flashes the posterior is $\alpha = 1.096 \pm 0.06506$ km, and the width matches the large-N prediction $\beta\sqrt{2/N} = 0.0625$ km that follows from the Fisher information of a Cauchy density.

The sample mean of the arrival points would be the answer for Gaussian data. Here it does not converge at all: after 64 flashes it is 3.578, and on average it is no closer to the truth than a single flash. The average of N independent Cauchy draws has exactly the same Cauchy distribution as one draw. Its variance is undefined, so the finite-variance central limit theorem of Chapter 2 does not apply. One wild flash, sent almost parallel to the shore, can drag the mean anywhere. The posterior never uses the mean. It weighs every flash through the likelihood. A wild flash far from α has a likelihood that is nearly flat in α, so it barely moves the posterior [Sivia p. 36].

The lighthouse shows both sides of the Gaussian approximation. With few data the posterior can be lopsided or have several peaks, and no "$\theta_0 \pm \sigma$" describes it. With many data from a well-behaved likelihood the posterior becomes a Gaussian of width $1/\sqrt{N\,I}$, where I is the Fisher information of one datum, whatever the shape of the single-datum likelihood. That second half is a theorem, the Bernstein–von Mises theorem of Section 5.13.

5.11 Nuisance parameters: integrating out what we do not care about

A spectrum shows a peak sitting on a smooth background: a Bragg peak on diffuse scattering, an emission line on a quasar's continuum, a resonance on the falling background of a collider mass spectrum. We want the amplitude A of the peak. The background level B is of no interest in itself, but it cannot be ignored, because the counts we measure are the sum of both, and a higher background leaves room for a smaller peak [Sivia §3.1]. Such a parameter, needed by the model and irrelevant to the question, is a **nuisance parameter**. How do we get a posterior for A alone that honestly includes our ignorance of B?

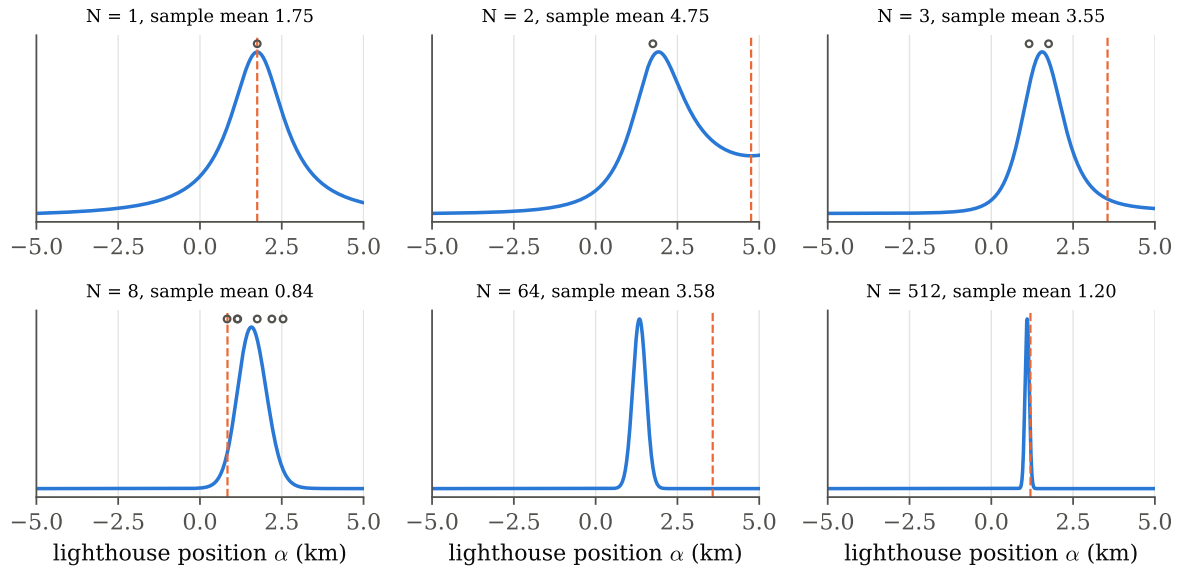


Figure 5.31: The posterior for the lighthouse position as flashes accumulate (N in each title). Circles at the top of the first four panels mark where the flashes landed, and the dashed vertical line is the sample mean of the arrival points. The posterior starts broad and sometimes two-peaked, becomes a narrow Gaussian-like peak near the true $\alpha = 1$ km, while the sample mean keeps jumping around.

5.11.1 Marginalising: the sum rule applied to parameters

Bayes' theorem gives the joint posterior $p(A, B \mid \boldsymbol{d})$, a density over pairs of values. Our question is about A alone: what is the probability that A lies in some small range, whatever B is? That can happen in many ways, one for each value of B, and the sum rule says to add up their probabilities. This is marginalisation (Chapter 1). A marginal distribution isolates the probability of a single variable, here A, by averaging out the "nuisance" variables, here B, from the joint distribution:

$$p(A \mid \boldsymbol{d}) = \int p(A, B \mid \boldsymbol{d})\, \mathrm{d}B \tag{5.36}$$

[Sivia (3.9)], [Wasserman p. 183, (11.9)], [Padilla (38)]. In the tree picture it is the sum over all branches of B that lead to the same A. With more nuisance parameters the integral runs over all of them. This is why the standard output of a cosmological analysis is a set of marginal posteriors, of one or two parameters out of dozens [HobsonBMC (2.6)].

Working definition 5.5 (marginal posterior)

For parameters of interest $\boldsymbol{\theta}$ and nuisance parameters $\boldsymbol{\nu}$, the **marginal posterior** is $p(\boldsymbol{\theta} \mid \boldsymbol{d}) = \int p(\boldsymbol{\theta}, \boldsymbol{\nu} \mid \boldsymbol{d})\, \mathrm{d}\boldsymbol{\nu}$ [the joint posterior summed over every value of the nuisance parameters, each weighted by how credible it is]. *Operationally:* on a grid, sum the joint posterior along the nuisance axes. From samples of the joint posterior, simply ignore the nuisance coordinates (Section 5.14).

How does not knowing B change what we learn about A? To see it, split the integrand with the product rule, $p(A, B \mid \boldsymbol{d}) = p(A \mid B, \boldsymbol{d})\, p(B \mid \boldsymbol{d})$:

$$p(A \mid \boldsymbol{d}) = \int p(A \mid B, \boldsymbol{d})\; p(B \mid \boldsymbol{d})\, \mathrm{d}B. \tag{5.37}$$

The marginal is an average of the conditional posteriors $p(A \mid B, \boldsymbol{d})$, one for each assumed background, weighted by how credible that background is. Now suppose an independent

calibration fixed the background exactly at b. Then $p(B \mid \boldsymbol{d}) = \delta(B - b)$ and

$$p(A \mid \boldsymbol{d}) \overset{(1)}{=} \int p(A \mid B, \boldsymbol{d})\, \delta(B - b)\, \mathrm{d}B \overset{(2)}{=} p(A \mid b, \boldsymbol{d}),$$

(1) The calibration puts all the probability for B at b.
(2) The delta function picks out the integrand at $B = b$.

So the conditional posterior is the special case in which the nuisance parameter is known perfectly [Sivia p. 44]. An analysis that "fixes" a parameter at a fiducial value is doing exactly this, without saying so: it gives that parameter a delta-function prior [HobsonBMC p. 42]. Any spread in $p(B \mid \boldsymbol{d})$ mixes conditionals centred at different places, so the marginal is broader.

Example: a peak on a background. Following Sivia, the expected count in channel k is $D_k = n_0[A\, e^{-x_k^2/2w^2} + B]$, a Gaussian peak of known width and position on a flat background, and the observed count N_k is Poisson with mean D_k [Sivia (3.1)–(3.4)]. With independent channels and a flat prior on $A, B \geq 0$,

$$\ln p(A, B \mid \{N_k\}) = \text{const} + \sum_k \big[N_k \ln D_k - D_k\big]$$

[Sivia (3.8)], the log of a product of Poisson probabilities with the factorials dropped. Here n_0 sets the overall count level, x_k is the position of channel k measured from the peak, and w is the peak width.

How much does not knowing B cost? It depends on whether the data can tell the peak from the background. Figure 5.32 shows two simulated set-ups, both with $A = 1$, $B = 2$.

- *Wide set-up* (top row): 15 channels spread well beyond the peak. The outer channels measure the background alone, so A and B are only moderately correlated, $\rho = -0.6688$. The marginal error on A is 0.2075, against 0.1538 if B were known to be 2.
- *Narrow set-up* (bottom row): 7 channels over a narrow range around the peak. No channel sees the background alone, so a little more background and a little less peak fit almost equally well. The correlation is $\rho = -0.9331$, and the marginal error on A grows to 0.4197, almost three times the conditional error 0.1517.

Both ratios agree with the factor $1/\sqrt{1-\rho^2}$ of Section 5.10. So when the data cannot separate signal from background, a calibration of the background is worth a lot. When they can, it adds little [Sivia p. 42].

Example: an unknown noise level. Readings x_k are Gaussian with mean μ and an error σ that nobody told us. Then σ is a nuisance parameter, and Chapter 3 showed that integrating it out with the prior $p(\sigma) \propto 1/\sigma$ gives $p(\mu \mid \boldsymbol{d}) \propto \big[\sum_k (x_k - \mu)^2\big]^{-N/2}$ [Sivia (3.38)], [HobsonBMC (2.26)]. Its shape in μ follows from splitting the sum about the sample mean $\bar{x}$:

$$\begin{aligned}\sum_k (x_k - \mu)^2 &\overset{(1)}{=} \sum_k \big[(x_k - \bar{x}) + (\bar{x} - \mu)\big]^2 \\ &\overset{(2)}{=} \sum_k (x_k - \bar{x})^2 + 2(\bar{x} - \mu) \sum_k (x_k - \bar{x}) + N(\bar{x} - \mu)^2 \overset{(3)}{=} V + N(\bar{x} - \mu)^2,\end{aligned}$$

(1) Add and subtract $\bar{x}$ inside each square.
(2) Expand the square. The middle factor $(\bar{x} - \mu)$ does not depend on k.

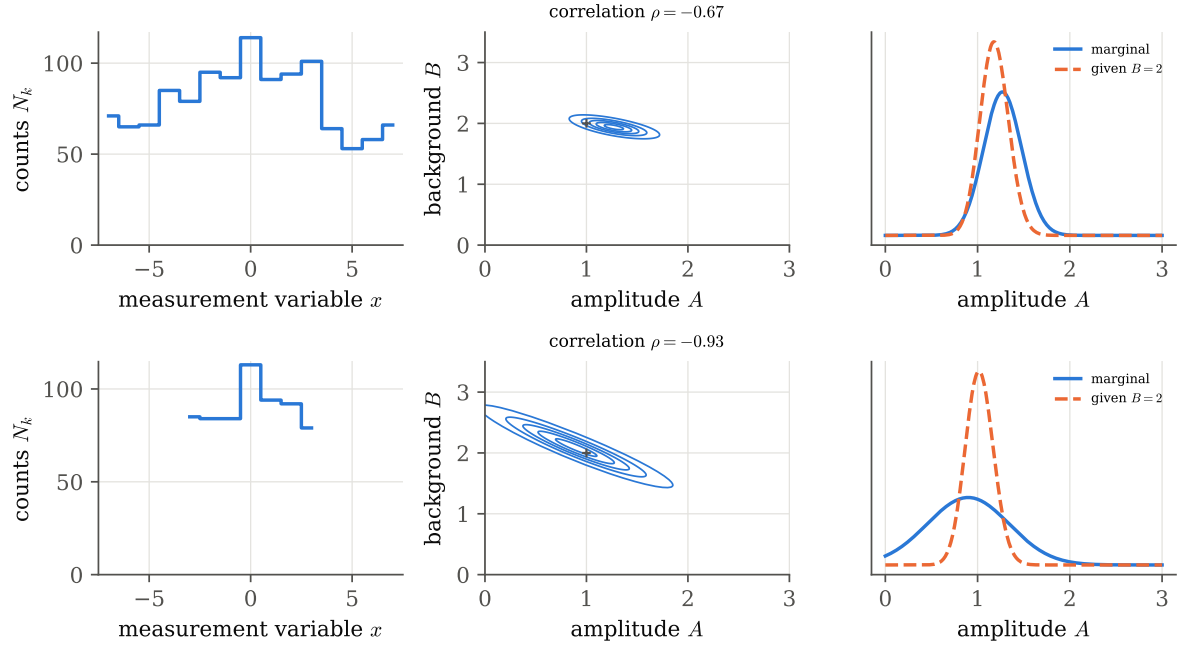


Figure 5.32: A peak of amplitude A on a flat background B, from Poisson counts. Top: 15 channels spanning the peak and the background around it. Bottom: 7 channels concentrated on the peak. Left: the simulated counts. Middle: contours of the joint posterior at 10, 30, 50, 70 and 90% of its maximum, with the true values marked by a cross. Right: the marginal posterior of A (solid) and the conditional posterior with B fixed at its true value (dashed). The narrow set-up correlates A and B strongly, and marginalising over B then widens the posterior of A a lot.

(3) $\sum_k (x_k - \bar{x}) = 0$ by the definition of $\bar{x}$, and $V \equiv \sum_k (x_k - \bar{x})^2$.

So $p(\mu \mid \boldsymbol{d}) \propto [N(\bar{x} - \mu)^2 + V]^{-N/2}$, a Student-$t$ curve centred on $\bar{x}$ [Sivia (3.41)–(3.44)]. For large N it approaches a Gaussian of width $\sqrt{V/N^2}$, the familiar $s/\sqrt{N}$ with $s^2 = V/N$. For small N its tails are much heavier than a Gaussian's. Not knowing the size of the noise costs precision, and marginalisation builds that cost in automatically. (Sivia's flat prior on σ gives the exponent $-(N-1)/2$ instead, a slightly wider curve. The choice matters only for very few data.)

5.11.2 Marginalising against profiling

The frequentist way to remove a nuisance parameter is to *maximise* over it: the profile likelihood $\mathcal{L}_p(A) = \max_B \mathcal{L}(A, B)$ (Chapter 3). Do integration and maximisation agree?

For a Gaussian posterior they do. Write the log-posterior about its peak as $L = \frac{1}{2}(ax^2 + 2cxy + by^2)$ in the deviations $x = A - A_0$, $y = B - B_0$, with $a, b < 0$ and $ab > c^2$. Maximising over y at fixed x gives $by + cx = 0$, so $y^* = -cx/b$, and

$$L(x, y^*) \overset{(1)}{=} \tfrac{1}{2}\left(ax^2 - \frac{2c^2x^2}{b} + \frac{c^2x^2}{b}\right) \overset{(2)}{=} \tfrac{1}{2}\,\frac{ab - c^2}{b}\,x^2, \qquad \int e^L \,\mathrm{d}y \overset{(3)}{=} e^{\frac{1}{2}\frac{ab-c^2}{b}x^2}\sqrt{\frac{2\pi}{-b}}.$$

(1) Substitute $y^* = -cx/b$.
(2) Collect the terms over the common denominator b.
(3) Complete the square in y, $by^2 + 2cxy = b(y + cx/b)^2 - c^2x^2/b$, and do the Gaussian integral over y. Its value $\sqrt{2\pi/(-b)}$ does not depend on x.

Profile and marginal have the same x-dependence, so after normalisation they are the same curve, with variance $b/(c^2 - ab)$ as in (5.35). The equality rests on step (3): the width of the

posterior in the nuisance direction, $(-b)^{-1/2}$, is the same for every x.

When that width changes with the parameter of interest, the two disagree. To see by how much, call the parameter of interest θ and the nuisance parameter ν. At each fixed θ, expand L about the best nuisance value $\hat{\nu}(\theta)$ and integrate the Gaussian in ν, exactly as in step (3). This gives

$$p(\theta \mid \boldsymbol{d}) = \int e^{L(\theta,\nu)}\,\mathrm{d}\nu \;\approx\; \underbrace{e^{L(\theta,\hat{\nu}(\theta))}}_{\text{profile}}\;\underbrace{\sqrt{2\pi}\,\sigma_\nu(\theta)}_{\text{volume}}, \qquad \sigma_\nu(\theta) = \left(-\partial_\nu^2 L\right)^{-1/2}\Big|_{\hat{\nu}(\theta)}. \tag{5.38}$$

The first factor is the profile, the best fit at this θ. The second measures how many values of ν fit almost as well. So the marginal rewards values of θ for which many values of the nuisance parameter fit the data, a large "volume" of good fits. The profile only asks how good the best fit is.

Example: a weak bump of unknown position. Readings y_i at 60 positions $x_i \in [0, 10]$ follow $y_i = A\,e^{-(x_i-\mu)^2/2w^2} + \epsilon_i$ with unit Gaussian noise and known narrow width $w = 0.4$. A bump of true amplitude $A = 1.2$ sits at $\mu = 5$, and we take flat priors on $A \in [0, 4]$ and on the position $\mu \in [0, 10]$ (Figure 5.33). Here the amplitude A is the parameter of interest and the position μ the nuisance parameter.

How wide is the range of good positions? It depends on A. When A is large, only positions near the true bump fit, so the posterior in μ is narrow. When A is near zero, every position fits equally well, because a bump of zero height is invisible wherever it is. Then the whole prior range of μ contributes. That large volume pulls the marginal posterior of A towards zero. For the data of Figure 5.33 its mode is at $A = 0$, and it puts probability 0.5375 on $A < 0.5$. The profile ignores volume. It peaks at $A = 0.83$ and gives 0.2077 to $A < 0.5$. One data set could be a fluke, so we repeat the analysis on 500 fresh data sets with the same true bump. On average the marginal gives $A < 0.5$ a probability of 0.20 and the profile 0.07 (each to within about 0.01), and the marginal peaks at $A = 0$ in 50 of them.

Neither is wrong; they answer different questions. The profile says "the best bump anywhere has amplitude about 0.83". The marginal says "averaged over where a bump could be, small amplitudes are more credible than the best fit alone suggests". This is the Bayesian face of the look-elsewhere effect (Chapter 4). It is also a first view of how the evidence penalises a model for its prior volume, the Occam factor of model comparison.

Watch out: The peak of the joint is not the peak of the marginal

The best-fit point of a joint posterior, projected onto one parameter, is the peak of the profile, not of the marginal. For skewed or banana-shaped posteriors, and whenever the nuisance width changes across the parameter of interest, the two differ, and quoting "the best fit" next to a marginal interval can put the best fit outside the interval. In cosmology these "projection" or "prior-volume" effects are routine in analyses with many nuisance parameters. Report which summary is quoted, and when they differ, report both [Sivia p. 54].

Marginals also hide structure. Two strongly correlated parameters may each have a broad marginal while their difference is sharply determined. And two pairs of parameters with the same marginals can have very different joint posteriors [Kruschke Fig. 12.1]. Whenever the question concerns a combination of parameters, compute the posterior of that combination from the joint, not from the marginals. With samples this is a one-line operation (Section 5.14).

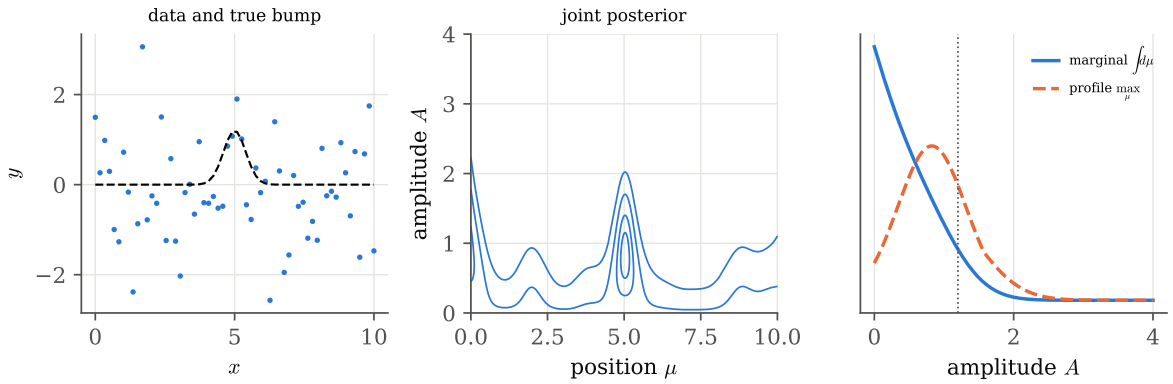


Figure 5.33: Marginal against profile for a weak bump whose position is unknown. Left: the data and the true bump (dashed). Middle: the joint posterior of amplitude and position. At small amplitude every position fits about equally well, so whatever probability the posterior keeps there is spread over the whole range of positions. Right: the marginal posterior of the amplitude (solid) integrates over position, so it gives small amplitudes more weight than the profile (dashed), which maximises over position and peaks near the best-fitting bump. The dotted line is the true amplitude.

5.12 Choosing a prior: subjective, flat and Jeffreys

Every posterior so far has needed a prior density $p(\theta)$ for the parameter θ we want, the weight each branch of the probability tree gets before the data arrive. Sometimes the choice is easy: an earlier experiment measured the same quantity, and its result is our prior. Often it is not. Suppose we are the first to measure the fraction θ of cosmic-ray muons that fire a new trigger. We believe it lies between 0 and 1 and not much more. What density should we write down, and how much does the answer depend on that choice?

Two schools answer differently [Wasserman pp. 181–182]. The first, called **subjectivism**, says that the prior should reflect our opinion about θ before the data are collected. The second looks for a prior that expresses no opinion at all, a "noninformative" prior. The obvious candidate is a flat prior, $p(\theta) \propto$ constant. Both are useful and both have traps. The trap of the flat prior is the more surprising one. A prior cannot be flat in every way of describing the same parameter, so "flat" always hides a choice of variable.

The questions, in words, are these.

1. How do we turn real but vague knowledge into a density?
2. What does a flat prior actually say, and when is a flat prior over an infinite range allowed?
3. Is there a rule that gives the same answer whichever variable we use to describe the parameter?

5.12.1 Priors that encode knowledge

A prior is a probability like any other. No probability, prior or likelihood, is ever simply "known". Each one is an assignment that reflects the information I we have. We call $p(X \mid I)$ a likelihood when X is data, and a prior when X is a parameter and no new measurement is involved. The difference is one of name, not of objectivity [Sivia §5.5]. What makes an assignment objective is consistency: two people with the same information should assign the same density. In practice there are three routes from knowledge to a density, which we take in turn.

Informal assessment. Think of ten points, each carrying 10% of the prior probability, and decide where they plausibly go [HobsonBMC §1.3.2]. That gives a rough centre, a rough width and a rough shape, which we then match to a standard density. For a *location* parameter [one that shifts the prediction, like a peak position or a mass] with centre c and width w, a Cauchy density

$$p(\theta) = \frac{w/\pi}{(\theta - c)^2 + w^2}$$

is a good choice. Its heavy tails still give real weight to values far from the guess, so the data can move the answer there if they demand it. For a positive *intensity* [a rate or an amplitude that cannot be negative] of plausible size a, an exponential $p(\theta) = a^{-1}e^{-\theta/a}$, or a Gamma density if we also know a width, does the same job [HobsonBMC (1.28)–(1.30)]. The prior does not have to be "right" in some undefinable sense. It has to be reasonable.

Moment matching. If earlier data exist, choose a standard density with the same mean and standard deviation as those data [Lambert §5.6.2]. For example, five earlier determinations of a detector gain, with mean 1.02 and scatter 0.05, give the prior $\mathcal{N}(1.02, 0.05^2)$. Which moments to match is an arbitrary choice, so this is a practical recipe rather than a principle.

Example: a prior elicited from quantiles. Experts find quantiles easier to state than standard deviations. Suppose a colleague says: "I would give even odds that the Hubble constant lies between 67 and 73 $\mathrm{km\,s^{-1}Mpc^{-1}}$, and I think it is equally likely to be below 67 as above 73." That is a statement of the 25% and 75% quantiles, $q_{25} = 67$ and $q_{75} = 73$. For a Gaussian prior $\mathcal{N}(\mu, s^2)$ each quantile is $q = \mu + z\,s$, where z is the matching quantile of the standard normal, $z_{75} = -z_{25} = 0.6745$ [Lambert §5.6.3]. Then

$$q_{75} - q_{25} \overset{(1)}{=} (\mu + z_{75}s) - (\mu + z_{25}s) \overset{(2)}{=} 2 \times 0.6745\, s \qquad \Rightarrow \quad s = \frac{73 - 67}{1.349} = 4.45,$$

$$q_{75} + q_{25} \overset{(3)}{=} 2\mu \qquad \Rightarrow \quad \mu = 70.$$

(1) Each quantile of $\mathcal{N}(\mu, s^2)$ is μ plus the standard-normal quantile times s.
(2) $z_{25} = -z_{75}$ because the standard normal is symmetric.
(3) Adding instead of subtracting, the z terms cancel for the same reason.

So the elicited prior is $\mathcal{N}(70, 4.45^2)$. With a panel of experts, the pairs (z, q) from all of them lie roughly on one straight line $q = \mu + zs$, and a linear regression through them gives a pooled μ (the intercept) and s (the slope).

5.12.2 Flat priors, and flat priors over an infinite range

A flat prior $p(\theta) =$ constant lets the likelihood alone set the shape of the posterior, since $p(\theta \mid \boldsymbol{d}) \propto \mathcal{L}(\theta) \times$ constant. For a bounded parameter such as our trigger fraction, the constant is 1 on $[0, 1]$, and after z firings in N muons the posterior is Beta$(z + 1, N - z + 1)$ (Section 5.8), which seems very reasonable [Wasserman p. 182].

For a parameter with no natural bound, such as the mean θ of Gaussian readings, a constant over the whole real line has infinite area. So it is not a probability density at all. Such a prior is called **improper**. Can we use it anyway? We can still multiply it by the likelihood, and then check whether the product can be normalised.

Example: the mean of Gaussian readings with a flat improper prior. Take n readings $x_1, \ldots, x_n$, each $\mathcal{N}(\theta, \sigma^2)$ with known σ, and $p(\theta) = c$ on the whole real line. Write $\bar{x} = \frac{1}{n}\sum_i x_i$.

Then

$$p(\theta \mid x_1, \ldots, x_n) \overset{(1)}{\propto} c \prod_{i=1}^{n} e^{-(x_i-\theta)^2/2\sigma^2} \overset{(2)}{\propto} \exp\Big[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\Big]$$
$$\overset{(3)}{=} \exp\Big[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\Big]\exp\Big[-\frac{n(\theta-\bar{x})^2}{2\sigma^2}\Big] \overset{(4)}{\propto} \exp\Big[-\frac{(\theta-\bar{x})^2}{2\sigma^2/n}\Big].$$

(1) Posterior $\propto$ likelihood $\times$ prior, dropping the factors $1/\sqrt{2\pi\sigma^2}$ that do not contain θ.
(2) The constant c drops out, and a product of exponentials is the exponential of the sum.
(3) Write $x_i - \theta = (x_i - \bar{x}) + (\bar{x} - \theta)$ and square. The cross term is $2(\bar{x} - \theta)\sum_i (x_i - \bar{x}) = 0$, because the deviations from the mean sum to zero.
(4) The first factor does not contain θ.

The posterior is $\mathcal{N}(\bar{x}, \sigma^2/n)$, a perfectly proper density, and its 95% interval $\bar{x} \pm 1.96\,\sigma/\sqrt{n}$ coincides exactly with the frequentist confidence interval of Chapter 3 [Wasserman p. 182]. An improper prior is harmless as long as the posterior it produces is a well-defined probability distribution.

The safe way to think about an improper prior is as the limit of proper ones [Sivia §5.5]. Use a prior that is flat between bounds A and B, compute the posterior, and watch what happens as $A \to -\infty$ and $B \to \infty$. Usually the posterior settles down and the bounds drop out, as they did above. If instead the conclusions depend strongly on A and B, the data carry very little information, and the prior range, which is prior knowledge, is doing the work. Every maximum-likelihood or least-squares fit uses this limit without saying so.

Improper priors are not harmless everywhere. Take the evidence $p(\boldsymbol{d}) = \int \mathcal{L}\, p \, \mathrm{d}\theta$ of a model whose prior is flat over a width W, so $p = 1/W$. The evidence carries this factor $1/W$, so as $W \to \infty$ it goes to zero. The model then loses any comparison with a model that knows even roughly what range to expect [HobsonBMC §1.3.2]. A flat prior over an enormous range is also not ignorance. It says we are almost certain that $|\theta|$ is enormous, since almost all of its area lies there. Are you really almost certain that $|\theta| > 10^{100}$?

5.12.3 Flat in which variable?

A flat prior is supposed to say "we know nothing". But knowing nothing about a parameter should also mean knowing nothing about any other way of describing it. Can one prior be flat in both? Return to the trigger fraction p with the flat prior $p \sim U(0,1)$, meant to represent our lack of information about p. A physicist might equally describe the trigger by its log-odds

$$\psi = \ln\frac{p}{1-p}, \qquad p = \frac{e^{\psi}}{1+e^{\psi}},$$

which runs over the whole real line. If we are ignorant about p, we are also ignorant about ψ, so we should use a flat prior for ψ too. But we have no freedom left: the flat prior on p already fixes the density of ψ. The rule is the change of variables for densities (Chapter 1): corresponding intervals must carry equal probability, $p_\Psi(\psi)\,|\mathrm{d}\psi| = p_P(p)\,|\mathrm{d}p|$. So

$$p_\Psi(\psi) \overset{(1)}{=} p_P\big(p(\psi)\big)\left|\frac{\mathrm{d}p}{\mathrm{d}\psi}\right| \overset{(2)}{=} 1 \times \frac{\mathrm{d}}{\mathrm{d}\psi}\frac{e^{\psi}}{1+e^{\psi}} \overset{(3)}{=} \frac{e^{\psi}(1+e^{\psi}) - e^{\psi}e^{\psi}}{(1+e^{\psi})^2} \overset{(4)}{=} \frac{e^{\psi}}{(1+e^{\psi})^2}. \tag{5.39}$$

(1) Change of variables for a density: equal probability in matching intervals.
(2) The flat density on $[0,1]$ is 1, and $p(\psi)$ increases, so the absolute value is not needed.
(3) Quotient rule.

(4) The $e^{2\psi}$ terms cancel.

This is a bump centred at $\psi = 0$, not a flat line [Wasserman p. 182]. How much of the probability does the bump hold in $|\psi| < 2$? That range corresponds to $p \in (1/(1+e^2), e^2/(1+e^2)) = (0.119, 0.881)$, which the flat prior gives probability 0.762. So a flat prior on p is a confident statement about ψ: the trigger is probably neither extremely efficient nor extremely inefficient.

Figure 5.34 shows why. Equal steps in p map onto steps in ψ that are narrow in the middle and wide at the ends. Each step carries the same probability, so probability gets squeezed into the middle of the ψ axis. The conclusion is unavoidable: "flat" has no meaning until we say flat in which variable, because a flat prior on a parameter is not flat on a nonlinear function of it. Flat priors are not **transformation invariant**.

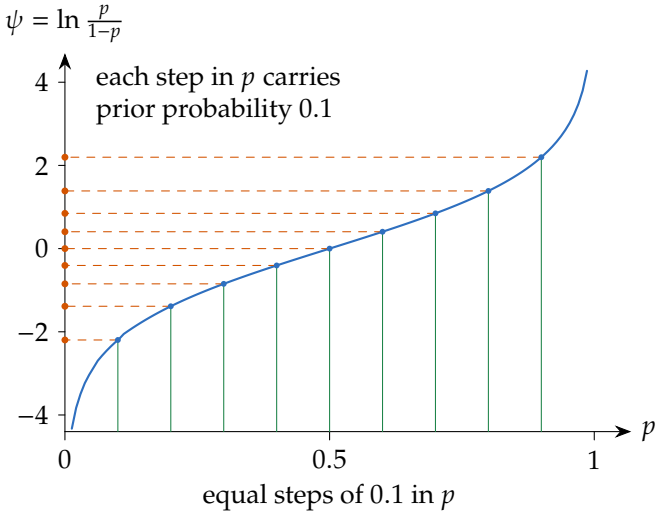


Figure 5.34: Why a flat prior in p is not flat in the log-odds ψ. The curve is $\psi = \ln[p/(1-p)]$. Nine equally spaced values of p (green lines) are carried up to the curve and across to the ψ axis (orange points). The orange points crowd together near $\psi = 0$ and spread apart towards the ends, so the same probability, 0.1 per step, is packed densely in the middle of the ψ axis and thinly at its ends: the bump of (5.39).

The same happens for any nonlinear function. Lambert's example is the probability $u = \theta^2$ that two randomly chosen individuals both carry a disease of prevalence θ [Lambert §5.6.1]. A flat prior on $\theta \in [0,1]$ gives u the density $p(u) = p(\theta)\,|\mathrm{d}\theta/\mathrm{d}u| = 1/(2\sqrt{u})$, which piles up at $u = 0$. For ten individuals, $v = \theta^{10}$, the pile-up is sharper still, $p(v) = \frac{1}{10}v^{-9/10}$. So a prior that looks uninformative about one quantity is quite informative about others. This is why "vague" or "diffuse" are better words than "uninformative".

Example: a scale parameter over four decades. The effect is largest for a positive *scale* parameter [one that sets a size, like a width, a lifetime or a noise level] whose order of magnitude is unknown. Let λ lie anywhere between 0.01 and 100, four decades. A prior flat in λ gives the decade $[10, 100]$ the probability $(100-10)/(100-0.01) = 0.9001$, and the decade $[0.01, 0.1]$ only 9.001×10^{-4}. So "complete ignorance" has become a 90% bet that λ is in the top decade. A prior flat in $\ln\lambda$ instead gives each of the four decades probability 0.25 (Figure 5.35, middle). For a quantity whose order of magnitude is unknown, the second is clearly what we mean.

Why flat in $\ln\lambda$ is special. A consistency argument makes this precise [Sivia §5.1]. The question is: what density says only "this is a location" or only "this is a scale"? Start with a location

parameter X. If that is all we know, then moving the origin by x_0 cannot change our density: $p(X)\,\mathrm{d}X = p(X + x_0)\,\mathrm{d}(X + x_0)$ for every x_0. Since $\mathrm{d}(X + x_0) = \mathrm{d}X$, this says $p(X) = p(X + x_0)$, so p is constant: complete ignorance about a location is a uniform density. If all we know about a length L is that it is a scale, then changing the units by a factor β (nanometres to ångströms) cannot change our density:

$$p(L)\,\mathrm{d}L = p(\beta L)\,\mathrm{d}(\beta L) \overset{(1)}{\Rightarrow} p(L) = \beta\, p(\beta L) \quad \text{for all } \beta > 0$$
$$\overset{(2)}{\Rightarrow} p(\beta) = \frac{p(1)}{\beta} \overset{(3)}{\Rightarrow} p(L) \propto \frac{1}{L}.$$

(1) $\mathrm{d}(\beta L) = \beta\,\mathrm{d}L$ for a constant β. Divide by $\mathrm{d}L$.
(2) Set $L = 1$ and solve for $p(\beta)$.
(3) Rename β as L: the relation holds for every positive value.

This is **Jeffreys' prior** for a scale parameter [Sivia (5.17)–(5.18)]. In terms of $s = \ln L$ it is $p(s) = p(L)\,\mathrm{d}L/\mathrm{d}s = (1/L) \times L =$ constant: flat in the logarithm, which is why problems involving magnitudes are best done on logarithmic graph paper. It is improper on $(0, \infty)$, so we use it between bounds or check that the posterior is normalisable. It is also the prior $p(C_\ell) \propto 1/C_\ell$ used for a power-spectrum amplitude in Section 5.8.

5.12.4 Jeffreys' rule: a prior that does not depend on the coordinates

Is there a recipe that gives the same prior whichever variable we choose, so that "Jeffreys in θ" transformed to ψ equals "Jeffreys computed in ψ"? Jeffreys found one [Wasserman pp. 182–183].

Working definition 5.6 (Jeffreys' prior)

For a model with density $f(x \mid \theta)$ for one datum, **Jeffreys' prior** is $p(\theta) \propto \sqrt{I(\theta)}$, where $I(\theta) = -\,\mathbb{E}_\theta\big[\partial_\theta^2 \ln f(X \mid \theta)\big]$ is the Fisher information [how sharply one datum pins down θ on average, Chapter 3]. For several parameters, $p(\boldsymbol{\theta}) \propto \sqrt{\det I(\boldsymbol{\theta})}$ with the Fisher information matrix. *Operationally:* compute the expected curvature of $\ln f$, take its square root, and normalise if the result has finite area.

Why is it invariant? Let $\psi = h(\theta)$ be any smooth, one-to-one reparametrisation. Use the equivalent form $I(\theta) = \mathbb{E}_\theta[(\partial_\theta \ln f)^2]$ of the Fisher information from Chapter 3. Then

$$I_\psi(\psi) \overset{(1)}{=} \mathbb{E}\Big[\Big(\frac{\partial \ln f}{\partial \psi}\Big)^2\Big] \overset{(2)}{=} \mathbb{E}\Big[\Big(\frac{\partial \ln f}{\partial \theta}\Big)^2\Big]\Big(\frac{\mathrm{d}\theta}{\mathrm{d}\psi}\Big)^2 \overset{(3)}{=} I_\theta(\theta)\Big(\frac{\mathrm{d}\theta}{\mathrm{d}\psi}\Big)^2,$$
$$\sqrt{I_\psi(\psi)} \overset{(4)}{=} \sqrt{I_\theta(\theta)}\,\Big|\frac{\mathrm{d}\theta}{\mathrm{d}\psi}\Big|.$$

(1) Fisher information in the new variable, as the mean square of the score.
(2) Chain rule, $\partial_\psi = (\mathrm{d}\theta/\mathrm{d}\psi)\,\partial_\theta$. The factor is not random and comes out of the expectation.
(3) The definition of I_θ.
(4) Square root of both sides.

The last line is exactly the change-of-variables law for a density, $p_\psi(\psi) = p_\theta(\theta)\,|\mathrm{d}\theta/\mathrm{d}\psi|$. So the square root of the Fisher information transforms as a density must. Computing Jeffreys' prior in θ and transforming it to ψ gives the same density as computing it in ψ directly. Geometrically, I is a metric on parameter space [it measures how distinguishable nearby parameter values are by their data], and $\sqrt{\det I}$ is its volume element. Jeffreys' prior spreads prior mass uniformly over distinguishable models, not over coordinate intervals [HobsonBMC (1.47)–(1.54)].

Example: the Bernoulli trigger. One muon fires the trigger, $x = 1$, with probability p, or not, $x = 0$. Then $\ln f = x \ln p + (1 - x) \ln(1 - p)$ and

$$I(p) \overset{(1)}{=} -\mathbb{E}\left[-\frac{X}{p^2} - \frac{1 - X}{(1 - p)^2}\right] \overset{(2)}{=} \frac{p}{p^2} + \frac{1 - p}{(1 - p)^2} \overset{(3)}{=} \frac{1}{p} + \frac{1}{1 - p} = \frac{1}{p(1 - p)}.$$

(1) Differentiate $\ln f$ twice in p.
(2) $\mathbb{E}[X] = p$ and $\mathbb{E}[1 - X] = 1 - p$.
(3) Cancel. Adding the fractions gives $1/[p(1 - p)]$.

Jeffreys' prior is $\sqrt{I(p)} = p^{-1/2}(1 - p)^{-1/2}$, the Beta$(\frac{1}{2}, \frac{1}{2})$ density [Wasserman p. 183, Ex. 11.6]. It is not flat; it rises at both ends. Compare it with the two other popular choices. The flat prior in p is $p^0(1 - p)^0$. The prior flat in ψ is the improper Haldane prior Beta$(0, 0) \propto p^{-1}(1 - p)^{-1}$. It is flat in ψ because its density in ψ is $p^{-1}(1 - p)^{-1} \times \mathrm{d}p/\mathrm{d}\psi = p^{-1}(1 - p)^{-1} \times p(1 - p) = 1$, using $\mathrm{d}p/\mathrm{d}\psi = p(1 - p)$ from (5.39). Jeffreys' exponents, $-\frac{1}{2}$, lie halfway between the flat prior's 0 and Haldane's -1.

Example: a Poisson rate and a Gaussian width. Take a count n with mean λT, where λ is the rate and T the exposure. Then $\ln f = n \ln \lambda - \lambda T + \text{const}$, so $\partial^2_\lambda \ln f = -n/\lambda^2$ and $I = \mathbb{E}[n]/\lambda^2 = T/\lambda$. Jeffreys' prior is therefore $\lambda^{-1/2}$. With zero observed events the posterior is $\propto \lambda^{-1/2} e^{-\lambda T}$, a Gamma$(\frac{1}{2}, T)$ density. For this density $2T\lambda$ has the χ^2 distribution with one degree of freedom, whose 95% point is 3.84. So the 95% upper limit is $\lambda < 3.84/(2T) = 1.92/T$, against $3.00/T$ from the flat prior (Section 5.8). Now take the width σ of Gaussian noise. Then $\ln f = -\ln \sigma - (x - \mu)^2/2\sigma^2$, so $\partial^2_\sigma \ln f = 1/\sigma^2 - 3(x - \mu)^2/\sigma^4$, whose expectation is $1/\sigma^2 - 3/\sigma^2 = -2/\sigma^2$. So $I = 2/\sigma^2$ and Jeffreys' prior is $1/\sigma$, the scale prior found above from units alone.

Does it matter? The three priors on real numbers. Figure 5.35 and Listing 5.7 compare the flat Beta$(1, 1)$, Jeffreys Beta$(\frac{1}{2}, \frac{1}{2})$ and Haldane Beta$(0, 0)$ priors for the trigger. With 1 firing in 10 muons the posterior means are 0.1667, 0.1364 and 0.1, and the 95% upper limits are 0.3644, 0.3306 and 0.2831. With so little data, the choice among "noninformative" priors changes the answer by a third. With 100 firings in 1000 muons the means are 0.1008, 0.1004 and 0.1 and the upper limits 0.1169, 0.1165 and 0.116: the choice has become irrelevant. A large sample makes the prior irrelevant for a reason, and the same reason tells us when it fails (Section 5.13).

Listing 5.7: Flat in which variable: a flat prior seen in the log-odds, mass per decade for a scale parameter, and three "noninformative" priors on the same data (`code/ch05/14_priors.py`)

```
# ------------------------------------------------------------------ flat p seen in logit
p = rng.uniform(size=100000)
psi = np.log(p / (1 - p))
ax[0].hist(psi, bins=120, range=(-8, 8), density=True, color=SERIES[0], alpha=0.6)
s = np.linspace(-8, 8, 400)
theory_line(ax[0], s, np.exp(s) / (1 + np.exp(s)) ** 2, label=r"$e^\psi/(1+e^\psi)^2$")
ax[0].set_xlabel(r"$\psi=\ln\frac{p}{1-p}$"); ax[0].set_title(r"$p\sim U(0,1)$, seen in
    $\psi$", fontsize=8)
ax[0].legend(fontsize=6); ax[0].set_yticks([])
frac_central = np.mean(np.abs(psi) < 2)

# ------------------------------------------------------------------ scale parameter: mass
    per decade
edges = 10.0 ** np.arange(-2, 3)          # 0.01, 0.1, 1, 10, 100
lo, hi = edges[0], edges[-1]
flat = np.diff(edges) / (hi - lo)
logflat = np.diff(np.log(edges)) / np.log(hi / lo)
xpos = np.arange(4)
```

```
ax[1].bar(xpos - 0.2, flat, 0.4, color=SERIES[1], label=r"flat in $\lambda$")
ax[1].bar(xpos + 0.2, logflat, 0.4, color=SERIES[0], label=r"flat in $\ln\lambda$")
ax[1].set_xticks(xpos, ["0.01-0.1", "0.1-1", "1-10", "10-100"], fontsize=7)
ax[1].set_ylabel("prior probability"); ax[1].set_title(r"scale $\lambda$: mass per decade",
    fontsize=8)
ax[1].legend(fontsize=6)

# ------------------------------------------------------------------ three priors, few data
th = np.linspace(1e-4, 0.6, 2000)
res = {}
for k, (name, a, b) in enumerate([("flat", 1, 1), ("Jeffreys", 0.5, 0.5), ("Haldane", 0,
    0)]):
    for z, N in [(1, 10), (100, 1000)]:
        post = stats.beta(a + z, b + N - z)
        res[(name, N)] = (post.mean(), post.ppf(0.95))
```

Lines 31–32 draw 10^5 values of p from the flat prior and transform each to its log-odds. This is the change of variables done by sampling, and line 35 overlays the exact density (5.39). Line 38 counts the fraction with $|\psi| < 2$, 0.763, against the exact 0.762. Lines 41–44 compute the prior mass per decade on $[0.01, 100]$ for the two priors: for the flat prior it is the width of the decade over the width of the range, and for the log-flat prior the log-width over the log-width of the range. Lines 55–58 update each of the three Beta priors with the conjugate rule Beta$(a + z, b + N - z)$ of Section 5.8 and read off the posterior mean and the 95% quantile.

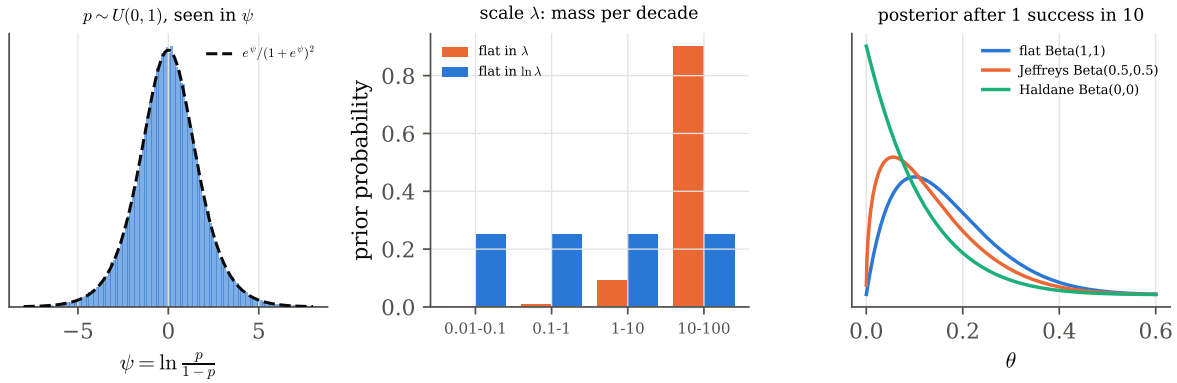


Figure 5.35: Flat depends on the variable. Left: 10^5 draws from a flat prior on p, histogrammed in the log-odds ψ, follow the bump $e^\psi/(1+e^\psi)^2$ (dashed). Middle: the prior probability of each decade of a scale parameter between 0.01 and 100. Flat in λ puts 90% in the top decade. Flat in $\ln\lambda$ gives each decade a quarter. Right: posteriors for a trigger efficiency after 1 firing in 10 muons, from the flat, Jeffreys and Haldane priors. With so few data they differ visibly.

Watch out: There is no prior that says nothing

Every prior says something about some function of the parameter. Jeffreys' rule is consistent across coordinates, but it depends on the likelihood, so it changes when the experiment changes, and in several dimensions $\sqrt{\det I}$ can behave badly or fail to normalise [HobsonBMC §1.3.2]. The general need for informal assessment means that the search for a universal "objective" prior is doomed. Choose a reasonable prior, say why, and check how much the answer depends on it [Lambert §5.7].

In practice most analyses use what Gelman calls **weakly informative** priors [Lambert §5.6.1]: smooth, proper densities that give zero weight only to impossible values (a negative rate, a lifetime longer than the age of the universe) and spread the rest broadly over physically sensible values. They avoid the arbitrary hard edges of a flat box, which produce kinked posteriors and

can trap the answer on the wrong side of the edge (Section 5.13).

5.13 Forgetting the prior: when the data win and when they cannot

Two physicists measure the same trigger efficiency θ, the fraction of muons that fire a new trigger, with true value 0.3. The first has no opinion and uses the flat prior Beta(1, 1). The second remembers a similar trigger with efficiency near 0.8 and is confident: Beta(80, 20), a prior worth $\kappa = a + b = 100$ muons (Section 5.8). Their priors barely overlap. They then record the same muons. After ten muons they still disagree. After a hundred thousand they agree to a fraction of an error bar (Figure 5.36). Is this agreement guaranteed? How fast does it come? And when does it never come, however much data we take?

A second kind of forgetting is easy to confuse with this one. A Markov chain Monte Carlo sampler (Chapter 9) started at a silly point forgets where it started. That is a different mechanism with a different cure. Keeping the two apart matters every time someone says "the prior washes out".

5.13.1 Why the prior is forgotten

Write the log-posterior for N independent data $x_1, \dots, x_N$ as

$$\ln p(\theta \mid x_1, \dots, x_N) = \underbrace{\ln p(\theta)}_{\text{one term}} + \underbrace{\sum_{i=1}^{N} \ln f(x_i \mid \theta)}_{N \text{ terms}} + \text{const.} \tag{5.40}$$

Here $f(x \mid \theta)$ is the density of one datum. The prior is a single term that does not grow with N. The log-likelihood $\ell_N(\theta) = \sum_i \ln f(x_i \mid \theta)$ is a sum of N terms, so it grows in proportion to N. As data accumulate, the likelihood therefore becomes a tall, narrow spike. The prior varies slowly across that spike, so it can only nudge it. How much? Expand both terms about the maximum-likelihood estimate $\hat\theta$, the peak of ℓ_N. The curvature of ℓ_N there is $N\hat I$, where $\hat I = -\ell_N''(\hat\theta)/N$ is the observed information per datum, close to the Fisher information $I(\theta)$ for large N (Chapter 3). Write $g = \frac{\mathrm{d}}{\mathrm{d}\theta} \ln p(\theta)\big|_{\hat\theta}$ for the slope of the log-prior there. Then

$$\begin{aligned} \ln p(\theta \mid x_1, \dots, x_N) &\overset{(1)}{\approx} \text{const} + g\,(\theta - \hat\theta) - \frac{N\hat I}{2}(\theta - \hat\theta)^2 \\ &\overset{(2)}{=} \text{const}' - \frac{N\hat I}{2}\Big(\theta - \hat\theta - \frac{g}{N\hat I}\Big)^2 . \end{aligned} \tag{5.41}$$

(1) Taylor expansion about $\hat\theta$: the log-likelihood to second order (its first derivative vanishes at its peak), the log-prior to first order (it changes little across the narrow likelihood peak).

(2) Complete the square: $-\frac{c}{2}u^2 + gu = -\frac{c}{2}(u - g/c)^2 + g^2/2c$ with $u = \theta - \hat\theta$ and $c = N\hat I$. The $g^2/2c$ goes into the constant.

The posterior is approximately Gaussian with

$$\text{mean } \hat\theta + \frac{g}{N\hat I}, \qquad \text{width } \frac{1}{\sqrt{N\hat I}}, \qquad \frac{\text{shift caused by the prior}}{\text{width}} = \frac{g}{\sqrt{N\hat I}}. \tag{5.42}$$

The prior moves the posterior by an amount that falls as $1/N$. The posterior width falls only as $1/\sqrt{N}$. So, measured in error bars, the influence of any smooth prior dies away as $N^{-1/2}$. Two analysts with log-prior slopes g_1 and g_2 at $\hat\theta$ disagree by $|g_1 - g_2|/\sqrt{N\hat I}$ posterior standard deviations.

Key idea: Bernstein–von Mises

If the model is regular, the parameter is identifiable and of fixed dimension, and the prior is smooth and nonzero at the true value, then for large N the posterior is approximately $\mathcal{N}\big(\hat{\theta},\, 1/[NI(\hat{\theta})]\big)$, whatever the prior. The posterior mean approaches the maximum-likelihood estimate, and a 95% credible interval approaches the 95% confidence interval $\hat{\theta} \pm 1.96/\sqrt{NI(\hat{\theta})}$ [Wasserman p. 181, Thm. 11.5].

This is the Gaussian approximation of Section 5.10 with a promise attached: for large N the curvature comes from the likelihood alone, so every reasonable prior gives the same answer. This is why frequentist and Bayesian error bars on well-measured quantities agree.

Example: two observers with Gaussian priors. Observers A and B have priors $\mathcal{N}(\mu_A, \Sigma_A^2)$ and $\mathcal{N}(\mu_B, \Sigma_B^2)$ for a quantity θ and share N readings with Gaussian noise σ and mean $\bar{m}$ [Padilla §3.4]. The rule "precisions add" of Section 5.8 gives each observer $i \in \{A, B\}$ a Gaussian posterior with

$$\hat{\mu}_i \overset{(1)}{=} \frac{N\bar{m}/\sigma^2 + \mu_i/\Sigma_i^2}{N/\sigma^2 + 1/\Sigma_i^2} \overset{(2)}{=} \frac{\bar{m} + r_i\,\mu_i}{1 + r_i}, \qquad r_i \equiv \frac{\sigma^2/N}{\Sigma_i^2},$$
$$\tau_i^2 \overset{(3)}{=} \frac{1}{N/\sigma^2 + 1/\Sigma_i^2} \overset{(4)}{=} \frac{\sigma^2/N}{1 + r_i}.$$

(1) Precision-weighted mean (5.32): the N readings count as one measurement $\bar{m}$ with variance σ^2/N.
(2) Divide numerator and denominator by N/σ^2.
(3) The posterior precision is the sum of the two precisions.
(4) Same division.

The ratio r_i compares the error of the data, $\sigma/\sqrt{N}$, with the width of the prior, Σ_i. When $r_i \ll 1$, the data dominate and $\hat{\mu}_i \to \bar{m}$ for both observers. How fast do they converge? Since $r_i \propto 1/N$, the difference $\hat{\mu}_A - \hat{\mu}_B \approx r_A\mu_A - r_B\mu_B - (r_A - r_B)\bar{m}$ falls as $1/N$. The common width $\tau \approx \sigma/\sqrt{N}$ falls only as $1/\sqrt{N}$. Their ratio falls as $N^{-1/2}$, the law of (5.42) again. With one reading the two posteriors can be far apart. After a hundred they are practically indistinguishable [Padilla Fig. 3].

Example: the confident and the open-minded physicist. Return to the two physicists of the opening: true efficiency 0.3, a flat prior, and a confident prior Beta(80, 20). For the trigger the conjugate update is exact, so we can follow the whole approach to agreement, not only its end. With z firings in N muons and prior Beta(a, b) of mean $m_0 = a/\kappa$, the posterior mean is

$$\frac{a + z}{\kappa + N} = \frac{N\hat{\theta} + \kappa m_0}{N + \kappa} \quad \Rightarrow \quad \text{mean} - \hat{\theta} = \frac{\kappa\,(m_0 - \hat{\theta})}{N + \kappa},$$

with $\hat{\theta} = z/N$. For the confident prior, $\kappa(m_0 - \hat{\theta}) \approx 100 \times (0.8 - 0.3) = 50$. The flat prior's own pull, $2(0.5 - \hat{\theta})/(N + 2)$, is tiny by comparison. Dividing by the posterior width $\sqrt{\hat{\theta}(1 - \hat{\theta})/N} \approx 0.458/\sqrt{N}$ gives the gap between the two analysts in units of the posterior width,

$$\frac{\text{gap}}{\text{width}} \approx \frac{50}{N + 100} \cdot \frac{\sqrt{N}}{0.458} = \frac{109\sqrt{N}}{N + 100}.$$

Surprisingly, this *grows* at first, because the posteriors narrow faster than they approach each other. It peaks at $N = \kappa = 100$, at about 5.5 widths. That is the moment of greatest

embarrassment: both analysts are confident, and they disagree. Then the gap falls as $109/\sqrt{N}$, the $N^{-1/2}$ law, and drops below one width only when $N \gtrsim 1.2 \times 10^4$. The simulation in Figure 5.36 follows this curve: its gap peaks at 5.885 widths near $N = 122$ and stays below one width from $N = 12999$ on.

The lesson: a prior worth κ observations needs many times κ observations to be forgotten. For a prior centred far from the truth, how many times is set by how many widths away its centre lies. This matches Sivia's finding that a fair-minded coin prior needs about a thousand flips before the data overrule it [Sivia §2.1], and Lambert's that a strong model is less sensitive to the prior [Lambert §5.7].

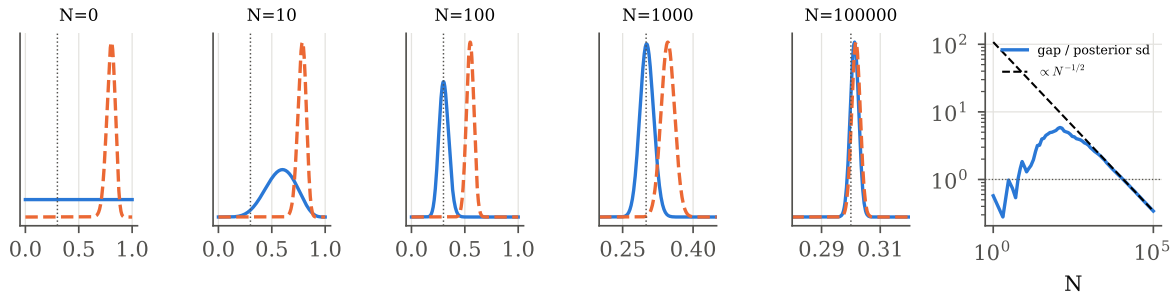


Figure 5.36: Two analysts forget their priors. First five panels: posteriors from the flat prior (solid) and from the confident prior Beta(80, 20) (dashed) after $N = 0$ to 10^5 tosses of a trigger with true efficiency 0.3 (dotted). The horizontal axis zooms in as the posteriors narrow. Last panel: the distance between the two posterior means in units of the posterior width. It rises until N is about the prior's weight of 100, then falls as $N^{-1/2}$ (dashed line) and crosses one width (dotted line) near $N = 10^4$.

5.13.2 When the prior is never forgotten

The derivation of (5.41) used every condition in the key idea: enough data, a prior nonzero at the truth, an identifiable parameter, a fixed number of parameters, and a truth away from the edge. Each one can fail, and each failure shows up in physics. We take them in turn.

Weak data. The theorem is about large N. For N comparable to the weight κ of an informative prior, or for a handful of events of any kind, the posterior is a genuine compromise and the prior matters. After one trigger firing in ten muons the three "noninformative" priors of Section 5.12 gave upper limits differing by a third, and in the $N = 10$ panel of Figure 5.36 the two analysts are 1.452 widths apart, with the gap still growing. With weak data the only honest course is to report the prior and show how the answer moves when it changes: a **prior sensitivity analysis** [Lambert §5.7].

A prior that is zero at the truth. A health minister is "morally certain" that the prevalence of a disease is below 25%, so the analysis uses a prior that is zero above 0.25 [Lambert §5.7.1]. The true value is 0.3. The posterior is prior times likelihood, so it is zero above 0.25 whatever the data say. As data accumulate the likelihood climbs steadily towards 0.3, and all the posterior can do is pile up against the wall at 0.25 (Figure 5.37, right). More data make the posterior *narrower at the wrong value*. The expansion (5.41) failed because $\ln p(\theta) = -\infty$ at the truth, so g does not exist. A prior should be zero only where a value is physically impossible.

A parameter the data cannot see. A detector records the total count n from two channels with rates λ_1 and λ_2 but cannot tell them apart, so $n \sim \text{Poisson}\big((\lambda_1 + \lambda_2)T\big)$ in exposure T. What can the count say about how the total is split between the channels? Reparametrise by the total rate $\Lambda = \lambda_1 + \lambda_2$ and the fraction $f = \lambda_1/\Lambda$ of the first channel, and take independent priors $p(\Lambda)\,p(f)$. Then

$$p(f \mid n) \overset{(1)}{=} \int p(\Lambda, f \mid n)\,\mathrm{d}\Lambda \overset{(2)}{=} \frac{\int \mathrm{P}(n \mid \Lambda)\,p(\Lambda)\,p(f)\,\mathrm{d}\Lambda}{\iint \mathrm{P}(n \mid \Lambda')\,p(\Lambda')\,p(f')\,\mathrm{d}\Lambda'\,\mathrm{d}f'}$$
$$\overset{(3)}{=} p(f)\,\frac{\int \mathrm{P}(n \mid \Lambda)\,p(\Lambda)\,\mathrm{d}\Lambda}{\int \mathrm{P}(n \mid \Lambda')\,p(\Lambda')\,\mathrm{d}\Lambda'} = p(f).$$

(1) Marginalise over the total rate (Section 5.11).
(2) Bayes' theorem. The likelihood depends on Λ only, because the detector sees only the total.
(3) $p(f)$ comes out of the Λ integral, and $\int p(f')\,\mathrm{d}f' = 1$.

So the posterior of f is its prior, exactly, for any number of counts. The data pin down Λ ever more precisely, but the prior alone decides the split. In the simulation, 5068 counts fix Λ to 1.4%, yet two different priors on f give two completely different posteriors for λ_1 (Figure 5.37, left). Such a parameter is **not identifiable** [different values of it predict exactly the same distribution of data]. Real cases are common: a cosmological amplitude and a galaxy bias that enter only as a product, two backgrounds with the same shape, a mass and a coupling that enter a rate only in combination.

More parameters than data. The theorem assumes the number of parameters stays fixed while N grows. Wasserman gives a simplified version of the Robins–Ritov example, a missing-data problem with $B \gg N$ unknown probabilities $\theta_1, \ldots, \theta_B$ [Wasserman pp. 186–188, Ex. 11.9]. With far fewer data than parameters, most θ_j never appear in the likelihood, so their posterior equals their prior. The posterior of their average $\psi = \frac{1}{B}\sum_j \theta_j$ therefore barely moves from its prior. Yet a simple frequentist weighted average, the Horvitz–Thompson estimator, does learn ψ: its error falls as $1/\sqrt{N}$ however large B is. It succeeds because it uses the known probability that each datum is observed. Those probabilities drop out of the likelihood, so no method built on the likelihood can use them. Bayesian methods are tied to the likelihood function, and in high-dimensional problems the likelihood may not give accurate inferences.

A truth on the edge. If the true value sits on the boundary of the allowed range, such as a signal strength of exactly zero with the prior restricted to nonnegative values, the posterior is a half-Gaussian piled against the edge, not a Gaussian, and the credible and confidence intervals need not agree (Chapter 4).

5.13.3 Forgetting the prior is not forgetting the start of a chain

When a posterior cannot be computed on a grid, we explore it with a Markov chain: a random walk in parameter space, biased so that it spends time in each region in proportion to the posterior there (Chapter 9). A chain started far from the bulk of the posterior wanders for a while before it settles. That early part, the **burn-in**, is thrown away, and after it the chain has forgotten where it started. It is tempting to say the chain has also "washed out" the prior. It has not, because the two forgettings work by different mechanisms.

- *A chain forgets its starting point* because every step is drawn from a rule that leaves one target distribution unchanged, and repeated steps pull any starting distribution towards

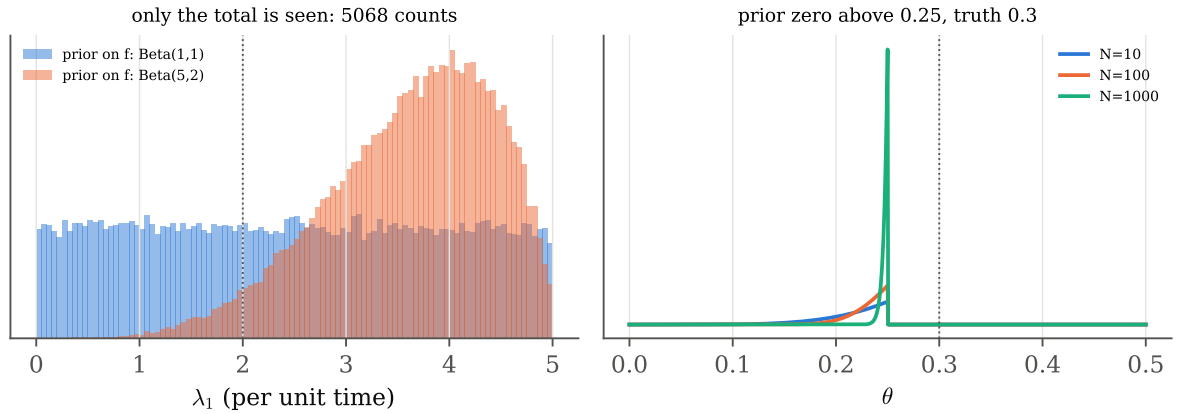


Figure 5.37: Two ways the prior is never forgotten. Left: a detector that sees only the total of two channels. With 5068 counts the total rate is known precisely, but the posterior of the first channel's rate λ_1 is set by the prior on its fraction: a flat prior gives the broad blue histogram, a Beta(5, 2) prior the orange one. The dotted line is the true $\lambda_1 = 2$. Right: a prior that is zero above 0.25 when the truth is 0.3 (dotted). With 10, 100 and 1000 data the posterior only piles up more sharply against the wall.

that target (Chapter 9 proves the contraction). Running the chain longer is the cure for a bad start.

- *A posterior forgets its prior* only through the likelihood, (5.42): more *data*. The target of the chain is prior×likelihood, prior included, so a chain run forever reproduces the prior's influence perfectly.

Example: four chains. Take 6 firings in 20 muons and the two priors of the trigger example. We run a Metropolis chain [a random walk whose proposed steps are accepted or rejected by comparing posterior values] from $\theta = 0.02$ and from $\theta = 0.98$ for each prior, 2×10^4 steps each, and discard the first 5000.

Listing 5.8: A Metropolis chain on the trigger posterior: only a ratio of posteriors enters (`code/ch05/15_bvm.py`)

```
94  def log_post(t, pa, pb):
95      if t <= 0 or t >= 1:
96          return -np.inf
97      return (Z + pa - 1) * np.log(t) + (NN - Z + pb - 1) * np.log(1 - t)
98
99
100 def metropolis(t0, pa, pb, n=20000, step=0.08):
101     chain = np.empty(n); t, lp = t0, log_post(t0, pa, pb)
102     for i in range(n):
103         tp = t + step * rng.standard_normal()
104         lpp = log_post(tp, pa, pb)
105         if np.log(rng.uniform()) < lpp - lp:          # only a ratio of posteriors enters
106             t, lp = tp, lpp
107         chain[i] = t
108     return chain
```

Lines 94–97 are the log-posterior, $\ln[\theta^{z+a-1}(1-\theta)^{N-z+b-1}]$ up to a constant, and $-\infty$ outside $(0, 1)$. Line 103 proposes a Gaussian step. Line 105 accepts it with probability $\min(1, p_{\text{new}}/p_{\text{old}})$. This is computed from the difference of log-posteriors, so the unknown evidence cancels and never appears.

The results separate the two forgettings (Figure 5.38). With the flat prior the two chains give means 0.3147 and 0.3182 from their two starts, against the exact posterior mean 0.3182. With the confident prior they give 0.7172 and 0.716, against the exact 0.7167. Within a few dozen steps each chain has forgotten whether it started at 0.02 or 0.98. But no number of steps brings the two priors together: they sample two different posteriors, both correctly. Only more muons would.

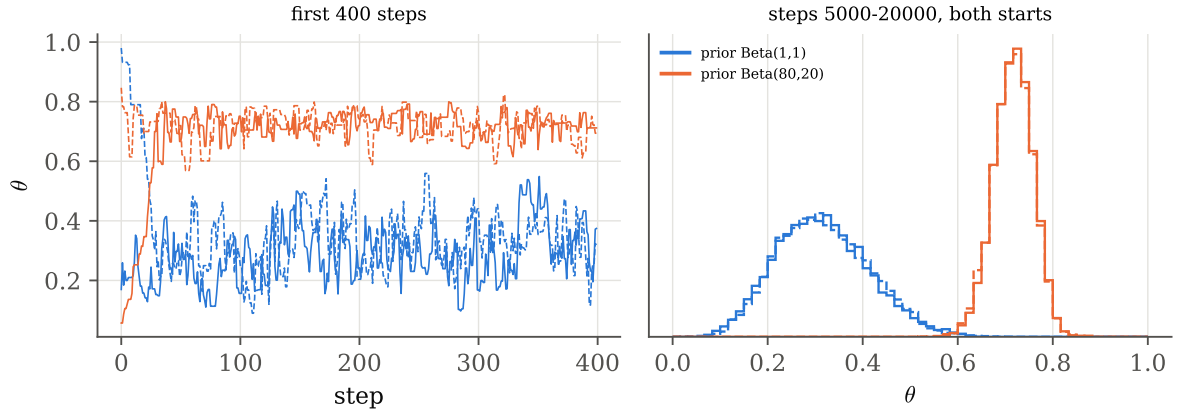


Figure 5.38: Two kinds of forgetting. Left: the first 400 steps of four Metropolis chains on the trigger posterior after 6 firings in 20 muons, started at 0.02 (solid) and 0.98 (dashed), with the flat prior (blue) and the confident prior (orange). Each pair forgets its start within a few dozen steps. Right: histograms of steps 5000 to 20000. Chains with the same prior agree whatever the start. Chains with different priors disagree however long they run.

Watch out: Burn-in does not remove a prior

Discarding burn-in removes the memory of the chain's starting point. It does nothing to the prior, which is part of the distribution the chain is built to sample. If two analyses with different priors disagree, running the chains longer will only make the disagreement more precise. The cure is more informative data, or a prior sensitivity analysis that reports how the result depends on the prior.

5.14 Sampling the posterior: a histogram instead of an integral

A mass on a spring oscillates as $x(t) = A\cos\omega t$. Where are we most likely to find it if we look at a random moment? "A random moment" already contains an assumption: the moment of looking is uniform in time, so the phase ωt is uniform over a cycle. With that assumption there are two ways to answer. One is a calculation: change variables from the uniform phase to the position and obtain the arcsine density $1/(\pi\sqrt{A^2 - x^2})$ (Chapter 1). The other needs no calculus at all. Pick a large set of random times, compute x at each, and histogram the values. The histogram piles up at the turning points, where the mass moves slowly, and it traces out the arcsine curve (Figure 5.40, bottom right). This idea can be used to find the probability distribution of position: we plot the histogram for x for a large set of random values of t.

The same idea works for a posterior. Suppose that instead of a formula for $p(\theta \mid \boldsymbol{d})$ we had a machine that produces random values $\theta_1, \ldots, \theta_B$ distributed according to it. Then a histogram of $\theta_1, \ldots, \theta_B$ approximates the posterior density $p(\theta \mid \boldsymbol{d})$ [Wasserman p. 180]. Every integral we have needed so far becomes a count or an average over the draws: a mean, a marginal, the probability of an interval. This is the bridge from Bayesian inference to Monte Carlo, and it is how almost all modern posteriors are handled.

If we can draw from the posterior, what can we compute from the draws, and how accurately?

Working definition 5.7 (posterior sample)

A **posterior sample** is a set of values $\theta_1, \dots, \theta_B$, each drawn from $p(\theta \mid \boldsymbol{d})$ [so that the fraction of draws in any interval approaches the posterior probability of that interval as B grows]. *Operationally:* the posterior probability of a set S is estimated by the fraction of draws in S, the posterior mean of any function $g(\theta)$ by $\frac{1}{B}\sum_i g(\theta_i)$, and a quantile by the corresponding order statistic of the sorted draws.

5.14.1 Why a histogram of draws is the density

Take a bin of width Δ around θ_j. Each draw lands in it independently with probability $P_j = \int_{\text{bin}} p(\theta \mid \boldsymbol{d})\,\mathrm{d}\theta \approx p(\theta_j \mid \boldsymbol{d})\,\Delta$, so the count n_j in the bin is binomial, $n_j \sim \text{Bin}(B, P_j)$ (Chapter 2). The histogram height, count divided by $B\Delta$, then has

$$\mathbb{E}\left[\frac{n_j}{B\Delta}\right] \overset{(1)}{=} \frac{BP_j}{B\Delta} \overset{(2)}{\approx} p(\theta_j \mid \boldsymbol{d}),$$
$$\frac{\text{sd}(n_j)}{\mathbb{E}[n_j]} \overset{(3)}{=} \frac{\sqrt{BP_j(1-P_j)}}{BP_j} \overset{(4)}{\approx} \frac{1}{\sqrt{BP_j}}.$$

(1) The mean of a binomial count is BP_j.
(2) $P_j \approx p(\theta_j \mid \boldsymbol{d})\Delta$ for a narrow bin.
(3) The standard deviation of a binomial count is $\sqrt{BP_j(1-P_j)}$.
(4) A single bin holds a small fraction of the probability, $P_j \ll 1$.

So the histogram is centred on the density. Its relative error in a bin is one over the square root of the expected count there. For example, with $B = 200000$ draws and a bin holding 1% of the probability, the expected count is 2000 and the height is good to about $1/\sqrt{2000} \approx 2\%$. The same counting argument says that averages over draws converge: the average of $g(\theta_i)$ has mean $\mathbb{E}[g(\theta) \mid \boldsymbol{d}]$ and standard deviation $\text{sd}(g)/\sqrt{B}$ for independent draws [Wasserman p. 181].

Watch out: Monte Carlo error is not the error bar

Two different widths appear. The *posterior width* $\text{sd}(\theta \mid \boldsymbol{d})$ is the physical uncertainty on the parameter. It is set by the data and does not shrink when we draw more samples. The *Monte Carlo error* $\text{sd}(\theta \mid \boldsymbol{d})/\sqrt{B}$ is how well the draws reproduce the exact posterior mean. It shrinks with B and says nothing about the physics. Quote the first as the error bar. Make sure the second is much smaller.

5.14.2 What the draws give for free

Figure 5.39 collects the operations. Each of them replaces an integral by a one-line manipulation of an array of draws.

Means and intervals. The posterior mean is estimated by $\frac{1}{B}\sum_i \theta_i$. The ends of the 95% equal-tailed interval are the 2.5% and 97.5% sample quantiles, and the HDI is the shortest window that contains 95% of the sorted draws (Section 5.9) [Wasserman p. 181].

Functions of the parameter. If $\tau = g(\theta)$, then $\tau_i = g(\theta_i)$ is a sample from the posterior of τ. No change of variables is needed [Wasserman pp. 180–181]. To see how much work this saves,

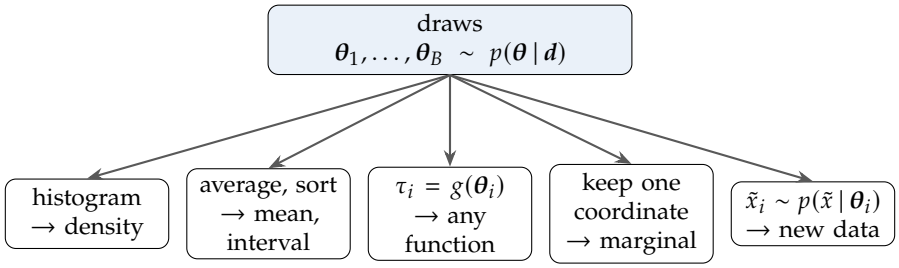


Figure 5.39: One array of posterior draws, five products. Histogramming it approximates the density. Averaging and sorting give the mean and the interval ends. Applying a function to each draw gives draws of that function. Ignoring all but one coordinate gives draws of that parameter's marginal. Simulating a new data set from each draw gives the distribution of future data.

take a coin's heads probability p after $s = 10$ heads in $n = 14$ tosses, with a flat prior. Write $x^n = (x_1, \ldots, x_n)$ for the tosses. The posterior is $p \mid x^n \sim \text{Beta}(s+1, n-s+1)$ (Section 5.8). What is the posterior density of the log-odds $\psi = \ln[p/(1-p)]$? The exact route goes through the cumulative distribution H of ψ, and then its derivative h:

$$H(\psi \mid x^n) \overset{(1)}{=} \mathrm{P}\Big(\ln \frac{P}{1-P} \le \psi \;\Big|\; x^n\Big) \overset{(2)}{=} \mathrm{P}\Big(P \le \frac{e^\psi}{1+e^\psi} \;\Big|\; x^n\Big) \overset{(3)}{=} \int_0^{e^\psi/(1+e^\psi)} f(p \mid x^n)\,\mathrm{d}p,$$

$$h(\psi \mid x^n) \overset{(4)}{=} f\Big(\frac{e^\psi}{1+e^\psi} \;\Big|\; x^n\Big)\,\frac{e^\psi}{(1+e^\psi)^2} \overset{(5)}{=} \frac{\Gamma(n+2)}{\Gamma(s+1)\Gamma(n-s+1)}\Big(\frac{e^\psi}{1+e^\psi}\Big)^{s+1}\Big(\frac{1}{1+e^\psi}\Big)^{n-s+1}.$$

(1) Definition of the cumulative distribution of $\Psi = \ln[P/(1-P)]$, with P the random heads probability.
(2) The log-odds is increasing in P, so the inequality can be inverted.
(3) The probability of an interval is the integral of the posterior density over it.
(4) Differentiate with respect to ψ: fundamental theorem of calculus and chain rule, with $\mathrm{d}p/\mathrm{d}\psi$ from (5.39).
(5) Insert the Beta density $f(p \mid x^n) = \frac{\Gamma(n+2)}{\Gamma(s+1)\Gamma(n-s+1)} p^s(1-p)^{n-s}$ at $p = e^\psi/(1+e^\psi)$, and note that $\frac{e^\psi}{(1+e^\psi)^2} = p(1-p)$ raises each power by one.

This is Wasserman's Example 11.3 [Wasserman p. 180]. The sampling route takes two lines: draw $P_i \sim \text{Beta}(11, 5)$, set $\psi_i = \ln[P_i/(1-P_i)]$, and histogram [Wasserman p. 181, Ex. 11.4]. The two agree (Figure 5.40, top left). For a function of several parameters, or one with no inverse in closed form, the sampling route is the only practical one.

Marginals and combinations. Take several parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ and B draws $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(B)}$. Their first coordinates $\theta_1^{(1)}, \ldots, \theta_1^{(B)}$ are a sample from the marginal posterior of θ_1. Keeping one coordinate and ignoring the rest replaces the $(p-1)$-dimensional integral of Section 5.11 [Wasserman p. 183, (11.9)].

A combination of parameters is just as easy. After a detector upgrade we want to know: did the efficiency improve? Before the upgrade $x_1 = 18$ of $n_1 = 40$ test pulses were detected, and after it $x_2 = 28$ of $n_2 = 40$. Call the two efficiencies p_1 and p_2. With a flat prior on the square $(p_1, p_2) \in [0,1]^2$,

$$p(p_1, p_2 \mid x_1, x_2) \propto p_1^{x_1}(1-p_1)^{n_1-x_1}\; p_2^{x_2}(1-p_2)^{n_2-x_2},$$

which factorises into $\text{Beta}(x_1+1, n_1-x_1+1)$ for p_1 times $\text{Beta}(x_2+1, n_2-x_2+1)$ for p_2: under the posterior the two efficiencies are independent [Wasserman pp. 183–184, Ex. 11.7]. The improvement is $\tau = p_2 - p_1$. Draw B values of each efficiency and subtract, $\tau_i = p_{2,i} - p_{1,i}$. The fraction of positive differences estimates $\mathrm{P}(p_2 > p_1 \mid d) = 0.9877$, and the 95% interval for the improvement

is (0.03117, 0.4355) (Figure 5.40, bottom left). If the parameters were correlated, the same subtraction of draws would use the correlation automatically, because each draw is a pair taken together. The separate marginals would have thrown that correlation away.

Predicting new data. The **posterior predictive distribution** of a future data set $\tilde{x}$ averages the forward model over our uncertainty about θ:

$$p(\tilde{x} \mid \boldsymbol{d}) \overset{(1)}{=} \int p(\tilde{x}, \theta \mid \boldsymbol{d})\, \mathrm{d}\theta \overset{(2)}{=} \int p(\tilde{x} \mid \theta, \boldsymbol{d})\, p(\theta \mid \boldsymbol{d})\, \mathrm{d}\theta \overset{(3)}{=} \int p(\tilde{x} \mid \theta)\, p(\theta \mid \boldsymbol{d})\, \mathrm{d}\theta.$$

(1) Sum rule: marginalise the joint distribution of new data and parameter.
(2) Product rule.
(3) Given θ, the new data are independent of the old data: the old data influence the prediction only through what they tell us about θ.

The sampling recipe mirrors the formula: for each draw θ_i, simulate one new data set $\tilde{x}_i$ from $p(\tilde{x} \mid \theta_i)$. The collection $\tilde{x}_1, \ldots, \tilde{x}_B$ is a sample from the predictive distribution [Lambert §7.8].

Example: two more heads. After 10 heads in 14 tosses, what is the probability that the next two tosses are both heads? The exact answer, found in Section 5.8, is $B(13,5)/B(11,5) = 0.4853$, not the plug-in $(10/14)^2 = 0.51$ [Padilla §2.3]. By sampling: for each draw p_i, toss two simulated coins with heads probability p_i and record whether both came up heads. The fraction is 0.4839, within the Monte Carlo error $\sqrt{0.485 \times 0.515/B} \approx 0.001$ of the exact value.

Example: failures in the next batch. One detector module in a batch of ten fails its burn-in test. How many failures should we expect in the next batch of ten?

- *Plug-in.* Take the failure rate to be exactly 0.1 and use Bin(10, 0.1). Its central 80% range is 0 to 2 failures.
- *Posterior predictive.* With a flat prior the posterior of the rate is Beta(1 + 1, 1 + 9) = Beta(2, 10). Draw a rate, then a batch of ten with that rate, and repeat. The central 80% range is 0 to 4 (Figure 5.40, top right) [Lambert §7.8].

The predictive range is wider because it includes two sources of spread: our uncertainty about the rate, and the binomial scatter of a batch. It also gives zero failures probability 0.2612, against the plug-in 0.3487, because the rate could well be larger than 0.1. Plugging in a best value always makes predictions look more certain than they are.

5.14.3 The same code for every posterior

Listing 5.9: Posterior draws put to work: a function of the parameter, two predictive calculations and a difference of two efficiencies (`code/ch05/16_sampling.py`)

```
# ------------------------------------------------------------- function of a parameter
s, n = 10, 14
p = rng.beta(s + 1, n - s + 1, B)
psi = np.log(p / (1 - p))
ps = np.linspace(-2.5, 4.5, 400)
e = np.exp(ps)
h = np.exp(-special.betaln(s + 1, n - s + 1)) * (e / (1 + e)) ** (s + 1) * (1 / (1 + e)) **
    (n - s + 1)
a = ax[0, 0]
```

```
41 a.hist(psi, bins=100, range=(-2.5, 4.5), density=True, color=SERIES[0], alpha=0.6,
       label="draws")
42 theory_line(a, ps, h, label=r"exact $h(\psi\,|\,x^n)$")
43 a.set_xlabel(r"log-odds $\psi=\ln\frac{p}{1-p}$"); a.set_yticks([]); a.legend(fontsize=6)
44 a.set_title("a function of the parameter", fontsize=8)
45
46 # ------------------------------------------------------------------ predictive
47 two_heads = np.mean((rng.uniform(size=B) < p) & (rng.uniform(size=B) < p))
48 exact2 = np.exp(special.betaln(13, 5) - special.betaln(11, 5))
49 th = rng.beta(1 + 1, 9 + 1, B)                        # 1 of 10, flat prior
50 xnew = rng.binomial(10, th)
51 k = np.arange(11)
52 pred = np.bincount(xnew, minlength=11) / B
53 plug = stats.binom(10, 0.1).pmf(k)
54 cdf = np.cumsum(pred)
55 lo80, hi80 = int(np.searchsorted(cdf, 0.1)), int(np.searchsorted(cdf, 0.9))
56 cdfp = np.cumsum(plug)
57 hi80p = int(np.searchsorted(cdfp, 0.9))
58 a = ax[0, 1]
59 a.bar(k - 0.2, pred, 0.4, color=SERIES[0], label="posterior predictive")
60 a.bar(k + 0.2, plug, 0.4, color=SERIES[1], label=r"plug-in, $\theta=0.1$")
61 a.set_xlabel("cases in the next sample of 10"); a.legend(fontsize=6)
62 a.set_title("predicting new data", fontsize=8)
63
64 # ------------------------------------------------------------------ difference of two rates
65 x1, n1, x2, n2 = 18, 40, 28, 40
66 p1 = rng.beta(x1 + 1, n1 - x1 + 1, B)
67 p2 = rng.beta(x2 + 1, n2 - x2 + 1, B)
68 tau = p2 - p1
69 a = ax[1, 0]
70 a.hist(tau, bins=100, density=True, color=SERIES[2], alpha=0.6)
71 a.axvline(0, color="k", lw=0.8)
72 a.set_xlabel(r"$\tau=p_2-p_1$"); a.set_yticks([])
73 a.set_title("a difference, by subtracting draws", fontsize=8)
74 p_pos = np.mean(tau > 0)
75 t_lo, t_hi = np.quantile(tau, [0.025, 0.975])
```

Lines 35–36 draw B values of the heads probability from Beta(11, 5) and turn each into a log-odds, and line 39 is the exact density from the chain above, used only as a check. Line 47 is the two-heads prediction: two uniform numbers per draw, each below p_i with probability p_i. Lines 49–50 draw a failure rate and then a whole new batch for each draw, and lines 54–55 read the 80% predictive range off the cumulative histogram. Lines 66–68 draw the two efficiencies independently and subtract them, and lines 74–75 turn the differences into a probability and an interval. The last panel of the figure is the oscillator: uniform phases, $x = A\cos\omega t$, histogram. It is the same three lines as the log-odds, with a random phase in the place of a random heads probability.

Here every draw came from a standard generator, because every posterior was a Beta density. That is the luxury of conjugacy (Section 5.8). For a realistic posterior in six or twenty dimensions we have neither a grid nor a generator. We can only evaluate $\mathcal{L}(\boldsymbol{\theta})\,p(\boldsymbol{\theta})$ at any point we choose. Monte Carlo methods turn uniform random numbers into draws from simple distributions (Chapter 6), and Markov chain Monte Carlo builds a random walk whose visited points are draws from a posterior known only up to its normalising constant (Chapter 9). Everything downstream of the draws, the histograms, averages, quantiles, functions, marginals and predictions of this section, stays exactly the same.

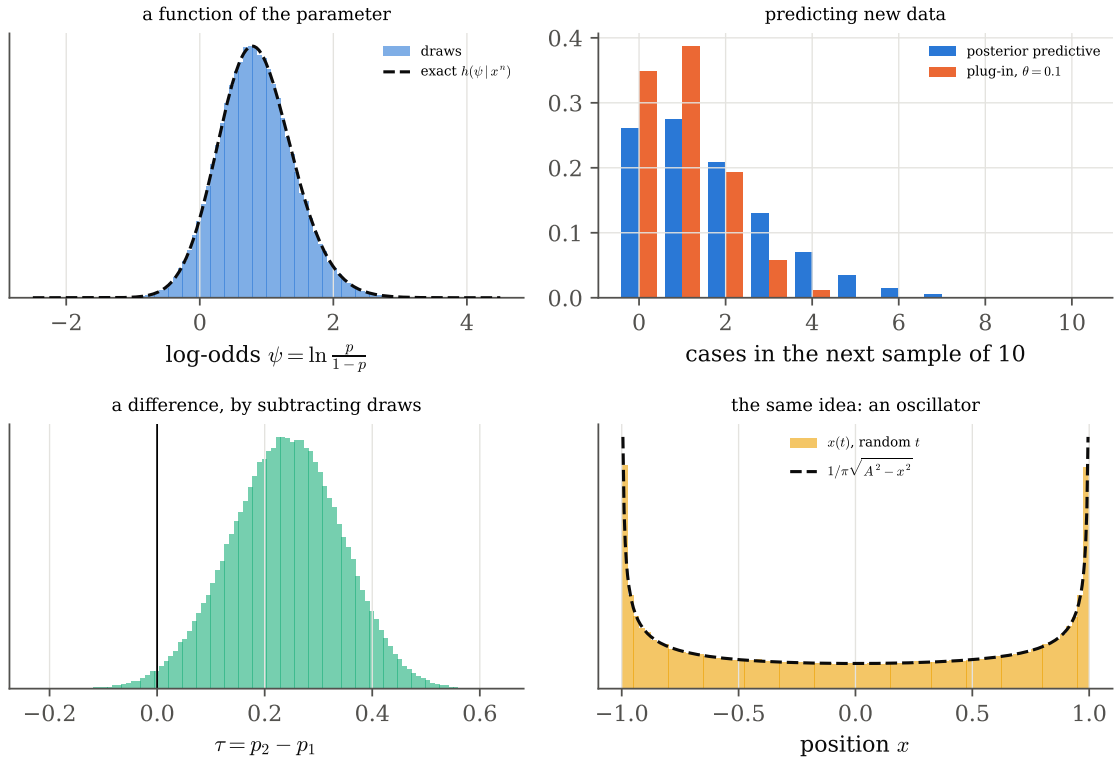


Figure 5.40: Posterior draws at work. Top left: draws of the heads probability after 10 heads in 14, turned into log-odds and histogrammed, against the exact density (dashed). Top right: the predicted number of failures in the next batch of ten modules after one failure in ten, from posterior draws (blue) and from plugging in the rate 0.1 (orange). Bottom left: the improvement in efficiency, from differences of draws. Bottom right: the same idea for an oscillator. Positions at uniformly random times follow the arcsine density (dashed).

Recap

- For a continuous parameter, the observed data are the event, the parameter values are the hypotheses, and $p(\lambda \mid \boldsymbol{d}) = \mathcal{L}(\lambda)p(\lambda)/p(\boldsymbol{d})$ is a density. On a grid, sum log-likelihoods and subtract the maximum before exponentiating.
- Conjugate pairs (Beta–Binomial, Gamma–Poisson, Gaussian–Gaussian) give exact posteriors. In the Gaussian case precisions add and means are precision-weighted.
- Summaries: the posterior mean, median and mode minimise squared, absolute and all-or-nothing loss. Equal-tailed intervals are invariant under reparametrisation. HDIs are the shortest.
- Near a single peak the posterior is Gaussian with covariance minus the inverse Hessian of $\ln p$. Marginal errors come from the diagonal of the inverse.
- Nuisance parameters are integrated out. Profiling maximises over them instead, and the two differ by a volume factor.
- No prior is flat in every parametrisation. Jeffreys' prior $\sqrt{I(\theta)}$ is coordinate independent: $1/\sigma$ for a scale, Beta$(\frac{1}{2}, \frac{1}{2})$ for a Bernoulli probability.
- With many data a smooth prior's influence dies as $N^{-1/2}$ posterior widths (Bernstein–von Mises). It is never forgotten if it is zero at the truth, if the parameter is not

identifiable, or if parameters outnumber data. A Markov chain forgets its start, not its prior.

- A histogram of posterior draws is the posterior. Functions, marginals and predictions of new data are one line each on the draws.

5.15 Which model produced the data?

Is the spectrum of primordial fluctuations exactly scale invariant, with spectral index $n_s = 1$, or is it tilted, with n_s a free number to be measured? Does a smooth falling background describe a diphoton mass spectrum, or does a resonance sit on top of it? Is a gravitational-wave event the merger of two black holes, or of two neutron stars whose tides leave a mark on the waveform? None of these asks for the value of a parameter inside a model we have agreed on. Each asks *which model* is right: which of several descriptions of nature, each with its own parameters, produced the data we have.

Until now the hypotheses were the possible values of a parameter θ inside one model. We observed data and asked which value of θ could have produced them. Now the hypotheses are whole models. The reasoning does not change. The four questions of Section 5.1 (what did we see, what could have produced it, how probable was it under each candidate, how do the prior probabilities change) get these answers: (1) we saw the whole data set $\boldsymbol{d}$; (2) the candidates are the models $M_1, M_2, \ldots$; (3) under each model the data have the probability $p(\boldsymbol{d} \mid M_i)$, the evidence derived below; (4) multiplying by the prior probabilities $\mathrm{P}(M_i)$ and normalising gives the posterior probability of each model.

Only step 3 is new, and it raises a question: how probable are the data under a model whose parameter we do not know? A model with a free parameter does not make one prediction for the data. It makes a different prediction for every value of its parameter, and before the data arrive we do not know which value nature chose. So the probability of the data under the model must be an average: over the parameter values the model allows, each weighted by how probable the model thought it was. Once this average is written down, Bayes' theorem chooses between models exactly as it chose between parameter values.

5.15.1 The probability of the data under a model

A model M, for our purposes, is two things [HobsonBMC §4.1], [Kruschke §10.1], [32, §4.1]:

1. a likelihood $p(\boldsymbol{d} \mid \boldsymbol{\theta}, M)$ [for each parameter value, the probability of the data this model would produce], and
2. a prior $p(\boldsymbol{\theta} \mid M)$ [how probable the model considered each parameter value before the data].

The two together are a complete recipe for generating data: draw $\boldsymbol{\theta}$ from the prior, then draw $\boldsymbol{d}$ from the likelihood at that $\boldsymbol{\theta}$. A tree shows the recipe (Figure 5.41). The first split chooses the model. The second chooses a parameter value inside the model. The leaves are the data. The observed data set $\boldsymbol{d}$ can be reached along many paths inside each model, one for every parameter value.

The probability of reaching the observed data inside model M is the sum over its paths. Each path contributes (probability of choosing $\boldsymbol{\theta}$)×(probability of producing $\boldsymbol{d}$ from $\boldsymbol{\theta}$). In

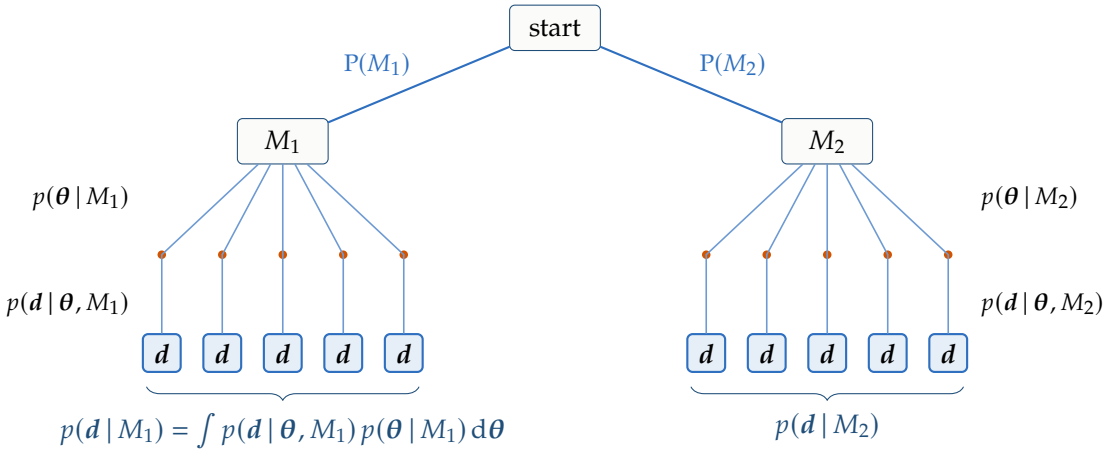


Figure 5.41: Two models as a probability tree. The first split picks the model, the second picks a parameter value inside it (orange dots stand for the continuum of values), and every path ends at the same observed data set $\boldsymbol{d}$. The probability of reaching $\boldsymbol{d}$ inside a model is the sum over all of that model's paths: the evidence.

symbols,

$$p(\boldsymbol{d} \mid M) \overset{(1)}{=} \int p(\boldsymbol{d}, \boldsymbol{\theta} \mid M)\, \mathrm{d}\boldsymbol{\theta} \overset{(2)}{=} \int p(\boldsymbol{d} \mid \boldsymbol{\theta}, M)\, p(\boldsymbol{\theta} \mid M)\, \mathrm{d}\boldsymbol{\theta}. \tag{5.43}$$

(1) Marginalisation (the sum rule for densities): the probability of $\boldsymbol{d}$ is the joint probability of $\boldsymbol{d}$ and $\boldsymbol{\theta}$ integrated over every $\boldsymbol{\theta}$, because the parameter must have *some* value.

(2) The product rule inside the model: the joint probability of a path is the probability of choosing $\boldsymbol{\theta}$ times the probability of producing $\boldsymbol{d}$ from it.

We have met this number before. It is the denominator of Bayes' theorem for the parameters of M, $p(\boldsymbol{\theta} \mid \boldsymbol{d}, M) = p(\boldsymbol{d} \mid \boldsymbol{\theta}, M)\, p(\boldsymbol{\theta} \mid M)/p(\boldsymbol{d} \mid M)$. It is also called the prior predictive probability of the data: the model's forecast of the data before they are taken [Sivia p. 82, (4.3)], [HobsonBMC p. 83, (4.5)], [Kruschke p. 268, (10.4)], [Padilla (12)], [32, eq. (17)]. In parameter estimation it did not depend on $\boldsymbol{\theta}$, so it only normalised the posterior and we could drop it. In model comparison it is the quantity we need.

The words that go with (5.43) are worth remembering: "For the distribution of parameter values you thought reasonable before the data came along, what was the average value of the likelihood?" [HobsonBMC p. 83]. The evidence is the likelihood averaged over the prior.

Working definition 5.8 (evidence of a model)

The **evidence** (or **marginal likelihood**) of a model M is

$$Z_M \equiv p(\boldsymbol{d} \mid M) = \int p(\boldsymbol{d} \mid \boldsymbol{\theta}, M)\, p(\boldsymbol{\theta} \mid M)\, \mathrm{d}\boldsymbol{\theta},$$

the probability, before the data were seen, that M would produce the data set actually observed [the likelihood averaged over the prior of that model]. *Operationally:* draw many parameter values from the prior and average the likelihood, or integrate likelihood × prior on a grid. It is a probability (or density) of the data, not of the parameters, so it carries the units and the normalisation of $\boldsymbol{d}$.

The prior is part of the model, because the evidence averages over it. Two models with the same likelihood and different priors are different models, and they have different evidences. Everything else about model comparison follows from this point [Kruschke p. 268].

5.15.2 Posterior model probabilities and the Bayes factor

With the evidence of each model in hand, steps 3 and 4 of the reasoning are Bayes' theorem with the models as the hypotheses. If we believed model M_i with probability $\mathrm{P}(M_i)$ before the data, then

$$\mathrm{P}(M_i \mid \boldsymbol{d}) = \frac{p(\boldsymbol{d} \mid M_i)\ \mathrm{P}(M_i)}{\sum_j p(\boldsymbol{d} \mid M_j)\ \mathrm{P}(M_j)}. \tag{5.44}$$

The logic is unchanged: nature produced one observed data set, several models could have produced it, and the posterior shares the probability among them according to how readily each would have produced what we saw. For two models the denominator cancels in the ratio:

$$\begin{aligned} \frac{\mathrm{P}(M_1 \mid \boldsymbol{d})}{\mathrm{P}(M_2 \mid \boldsymbol{d})} &\overset{(1)}{=} \frac{p(\boldsymbol{d} \mid M_1)\ \mathrm{P}(M_1)/p(\boldsymbol{d})}{p(\boldsymbol{d} \mid M_2)\ \mathrm{P}(M_2)/p(\boldsymbol{d})} \\ &\overset{(2)}{=} \underbrace{\frac{p(\boldsymbol{d} \mid M_1)}{p(\boldsymbol{d} \mid M_2)}}_{B_{12}} \times \frac{\mathrm{P}(M_1)}{\mathrm{P}(M_2)}. \end{aligned} \tag{5.45}$$

(1) (5.44) for $i = 1$ and $i = 2$, with the common denominator written as $p(\boldsymbol{d}) = \sum_j p(\boldsymbol{d} \mid M_j)\,\mathrm{P}(M_j)$.
(2) $p(\boldsymbol{d})$ cancels, and we group the evidences and the priors.

Posterior odds equal the **Bayes factor** B_{12} times the prior odds [Kruschke p. 268, (10.5)], [Sivia p. 82, (4.2)], [32, eqs. (19)–(20)]. It is the odds form of Bayes' theorem (posterior odds = likelihood ratio × prior odds), with each hypothesis now a whole model. The Bayes factor is the factor by which the data change the odds, whatever the odds were before.

Working definition 5.9 (Bayes factor)

The **Bayes factor** of model M_1 against model M_2 is the ratio of their evidences,

$$B_{12} = \frac{Z_1}{Z_2} = \frac{p(\boldsymbol{d} \mid M_1)}{p(\boldsymbol{d} \mid M_2)},$$

[how many times more probable the observed data were under M_1 than under M_2]. $B_{12} > 1$ means the data moved the odds towards M_1. It is usually quoted as $\ln B_{12} = \ln Z_1 - \ln Z_2$.

Model comparison is parameter estimation for one more parameter: the label of the model. Give the model a label m that takes the values $1, 2, \ldots$. The complete unknown is then the pair (m, the parameters of model m). Bayes' theorem on this joint space shares out the probability at both levels at once: between the models, and between parameter values inside each model [Kruschke pp. 266–267, (10.1)–(10.3)]. Summing the joint posterior over the parameters of each model leaves the posterior of m alone, which is (5.44). Cosmologists speak of three levels of inference [HobsonBMC §4.2]: parameter estimation inside a model, model comparison between models, and model averaging, which combines the parameter posteriors of all models weighted by the model probabilities (Section 5.18).

Example: which factory minted this coin? The example is Kruschke's [Kruschke §10.2]. A novelty company has two coin factories. One makes tail-biased coins: the bias θ [the probability

of heads] of its coins is spread as Beta($\theta \mid 3.5, 8.5$), with mode 0.25. The other makes head-biased coins, Beta($\theta \mid 8.5, 3.5$), mode 0.75. We are given one coin from one of the factories, we do not know which, and we toss it nine times.

The event. The coin shows a particular sequence with $z = 6$ heads in $N = 9$ tosses.

What could have produced it? Two models: M_1, the coin came from the tail-biased factory; M_2, from the head-biased factory. Inside each model the coin's own bias θ is unknown.

How probable was this sequence under each model? Under M_i with prior Beta($\theta \mid a, b$), a coin of bias θ produces this sequence with probability $\theta^z(1-\theta)^{N-z}$. Averaging over the factory's biases,

$$p(\boldsymbol{d} \mid M_i) \overset{(1)}{=} \int_0^1 \theta^z(1-\theta)^{N-z}\,\frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}\,\mathrm{d}\theta$$
$$\overset{(2)}{=} \frac{1}{B(a,b)}\int_0^1 \theta^{z+a-1}(1-\theta)^{N-z+b-1}\,\mathrm{d}\theta \overset{(3)}{=} \frac{B(z+a,\,N-z+b)}{B(a,b)}. \tag{5.46}$$

(1) The evidence (5.43): likelihood of the sequence times the Beta prior density, whose normalisation is the Beta function $B(a,b) = \int_0^1 \theta^{a-1}(1-\theta)^{b-1}\mathrm{d}\theta$.
(2) Collect the powers of θ and of $1-\theta$.
(3) The remaining integral is, by definition, the Beta function at the shifted arguments.

This is Kruschke's (10.6). For the tail-biased factory ($a = 3.5$, $b = 8.5$) it gives $p(\boldsymbol{d} \mid M_1) = 4.993 \times 10^{-4}$, and for the head-biased factory ($a = 8.5$, $b = 3.5$) it gives $p(\boldsymbol{d} \mid M_2) = 0.002339$. A 1001-point grid over θ reproduces both to the digits shown. The Bayes factor is $B_{12} = 0.2135$: six heads in nine were 4.683 times more probable from the head-biased factory.

Only then the priors, and the normalisation. If either factory was equally likely to have made our coin, the prior odds are one, and

$$\mathrm{P}(M_1 \mid \boldsymbol{d}) \overset{(1)}{=} \frac{B_{12}}{1+B_{12}} = \frac{0.2135}{1+0.2135} = 17.6\%, \qquad \mathrm{P}(M_2 \mid \boldsymbol{d}) = 82.4\%.$$

(1) From (5.45) with prior odds 1: $\mathrm{P}(M_1 \mid \boldsymbol{d})/\mathrm{P}(M_2 \mid \boldsymbol{d}) = B_{12}$ and the two posteriors add to one.

Figure 5.42 shows where the numbers come from. The evidence is the area under likelihood×prior. The head-biased factory puts its prior where the likelihood is large, so its area is almost five times bigger.

Notice what the comparison did *not* give us: any statement about the bias of this coin. For that we go back to parameter estimation inside each model. Inside the tail-biased factory the posterior is Beta($z+a, N-z+b$) with mean $(z+a)/(N+a+b) = 0.4524$; inside the head-biased factory the mean is 0.6905. If we must give one answer for the bias without deciding on a factory, we average the two posteriors with the model probabilities as weights, 17.6% and 82.4%, which gives the mean 0.6486 [Kruschke §10.4]. We return to this averaging in Section 5.18.

Example: background only, or background plus signal? A counter records events from a source that may or may not exist. The background is well measured: in the time of the run we expect $b = 2$ background counts.

The event. The run records $n = 5$ counts.

What could have produced it? Model M_0: background only, counts $\sim$ Poisson(b). Model M_1: background plus a signal of unknown strength s, counts $\sim$ Poisson($b+s$). Before the run, the physics of the source allows any s between 0 and $s_{\max} = 10$ counts, and we take s uniform in that range.

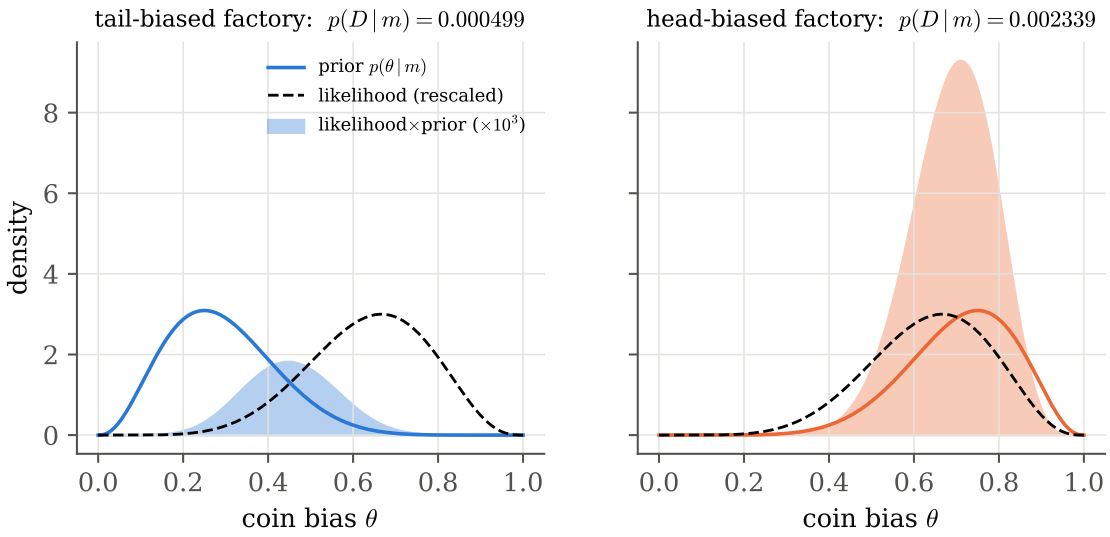


Figure 5.42: The evidence as an area. In each panel the solid curve is the factory's spread of coin biases (its prior), the dashed curve the likelihood of six heads in nine tosses (rescaled to fit), and the shaded region their product, whose area is the evidence printed above the panel. The likelihood is the same in both panels; only the prior differs, and with it the overlap.

How probable were five counts under each model? Under M_0 there is nothing to average: $p(n \mid M_0) = e^{-b} b^n / n! = 0.03609$. Under M_1 we average the Poisson probability over the prior of s:

$$p(n \mid M_1) \overset{(1)}{=} \frac{1}{s_{\max}} \int_0^{s_{\max}} \frac{e^{-(b+s)}(b+s)^n}{n!}\, \mathrm{d}s \overset{(2)}{=} \frac{1}{s_{\max}} \int_b^{b+s_{\max}} \frac{e^{-u} u^n}{n!}\, \mathrm{d}u$$
$$\overset{(3)}{=} \frac{1}{s_{\max}} \Big[F_n(b) - F_n(b + s_{\max}) \Big], \qquad F_n(a) \equiv \sum_{j=0}^{n} \frac{e^{-a} a^j}{j!}. \tag{5.47}$$

(1) The evidence (5.43) with the uniform prior density $1/s_{\max}$.
(2) Substitute $u = b + s$.
(3) The identity $\int_a^\infty e^{-u} u^n / n!\, \mathrm{d}u = F_n(a)$, the probability that a Poisson variable of mean a is at most n. It follows by integrating by parts n times; each step peels off one term $e^{-a} a^j / j!$. It says that the waiting time to the $(n+1)$-th event of a unit-rate Poisson process exceeds a exactly when at most n events happened before a.

With $F_5(2) = 0.9834$ and $F_5(12) = 0.02034$ this gives $p(n \mid M_1) = 0.09631$ (quadrature agrees). The Bayes factor is $B_{10} = 2.669$ in favour of a signal, and with even prior odds the probability that the source is there is 0.7274.

The Bayes factor is smaller than the best fit would suggest, and the gap is the price of not knowing s. The best signal value is $\hat{s} = n - b = 3$. At that value five counts are 4.862 times more probable than under background only. The Bayes factor is only 2.669 because the signal model averages over possible values of s. Its prior also assigns weight to values such as $s = 9$, under which observing only five counts would be unlikely. These values pull the average down. The ratio of the two numbers, 0.5489, is the price the signal model pays for not knowing s in advance; Section 5.16 derives this price in general. The price depends on how wide the prior range is. If the physics had allowed anything up to $s_{\max} = 100$, the same five counts would have given $B_{10} = 0.2725$, and the data would then favour background only. The prior range of a parameter is part of what the model claims.

Key idea: The evidence judges a model with its prior

A model is a likelihood *and* a prior. Its evidence is the likelihood averaged over its prior, and the Bayes factor, the ratio of two evidences, multiplies the prior odds of the models. A model that spreads its prior over parameter values the data do not favour gets a smaller average, even if its best fit is excellent.

Two cautions about reading a quoted Bayes factor. First, conventions differ. $B_{01} = Z_0/Z_1$ and $B_{10} = Z_1/Z_0 = 1/B_{01}$ are both in use, and some authors call $\ln(Z_0/Z_1)$ itself "the Bayes factor" [Padilla (13)–(14)]. Always check which model is in the numerator and whether a logarithm has been taken. Second, a Bayes factor is not the posterior odds. It multiplies the prior odds, so a model with prior probability zero stays at zero whatever its Bayes factor [Kruschke §10.5.1].

5.15.3 No model is rejected on its own

Can we ask how probable one model is, without naming any competitor? No. Look at the denominator of (5.44): it is a sum over the competitors, each with its evidence. Suppose we try to make do with "M is false" as the only alternative. The denominator then needs $p(\boldsymbol{d} \mid \text{not } M)$, the probability of the data given that M is false. That number cannot be computed, because "M is false" makes no prediction for the data [Sivia pp. 87–88]. We need well-defined alternatives, each making its own prediction. Once we have a list of them, the question has become model comparison.

A bad fit, on its own, does not make a model improbable. A large χ^2 means a small likelihood $p(\boldsymbol{d} \mid \boldsymbol{\theta}, M)$ at the best parameters. The posterior probability of M also needs the prior and the evidence of the alternatives, and a bad fit says nothing about those. What a bad fit does do is tell us to go and look for an alternative [Sivia p. 88]. The history of Neptune shows that support for a theory is always support against something [32, §4.1]. Finding Neptune near the predicted position strengthened Newtonian gravity against a theory that predicted no planet there. It did nothing to Newton's standing against Einstein's gravity, which predicts the same orbit. How much an observation supports a theory depends on what it is compared with.

5.16 Occam's razor from the evidence

To see why comparing best fits can favour a more elaborate model even when the extra feature is absent, consider forty measurements y_i are taken at equally spaced points x_i between 0 and 10, each with Gaussian noise of standard deviation $\sigma = 1$. One physicist says the points scatter about a straight line, $y = a + bx$. A second says there is also a Gaussian bump on top of the line, the signature of a resonance, with known width $W = 0.5$, height A and unknown position μ. Whatever the data, the second physicist's best fit is at least as good as the first's, because setting $A = 0$ turns the bump model into the line. So the best-fit χ^2 never prefers the simpler model, and a rule that picks the model with the smallest χ^2 would find a bump in every data set. For the same reason a tenth-order polynomial always beats a straight line, and nobody believes it [Sivia p. 81].

The evidence settles this without any extra rule. The more flexible model spread its prior over many parameter values, and the data rule most of them out. So its *average* likelihood, which is the evidence, can be low even when its *best* likelihood is high.

5.16.1 Mr A and Mr B

The simplest version is a story due to Jeffreys, as told by Gull and Sivia [Sivia §4.1]. Mr A has a theory with no adjustable parameter. Mr B has a theory with one adjustable parameter λ. Whose theory should we prefer, given data $\boldsymbol{d}$?

The event is the data $\boldsymbol{d}$. *The hypotheses* are A and B. *How probable was* $\boldsymbol{d}$ *under each?* For Mr A this is a straightforward likelihood, $p(\boldsymbol{d} \mid A)$, since his theory predicts the data with nothing left to choose. Mr B cannot predict the data without a value of λ, so his probability of the data is the evidence (5.43), an average over his prior for λ. To make it concrete, Sivia makes two simple assumptions (Figure 5.43):

1. Before seeing the data, Mr B will only say that λ lies between $\lambda_{\min}$ and $\lambda_{\max}$, so his prior is uniform: $p(\lambda \mid B) = 1/(\lambda_{\max} - \lambda_{\min})$ in that range, zero outside [Sivia (4.4)].
2. His likelihood has a single peak at the best value λ_0, of height $p(\boldsymbol{d} \mid \lambda_0, B)$, and near the peak it is Gaussian with width $\delta\lambda$ [Sivia (4.5)]:

$$p(\boldsymbol{d} \mid \lambda, B) \approx p(\boldsymbol{d} \mid \lambda_0, B)\, \exp\!\left[-\frac{(\lambda - \lambda_0)^2}{2\,\delta\lambda^2}\right].$$

The likelihood is a function of λ but not a density in λ: its area is not one. Its height at the peak is the probability of the data at the best value.

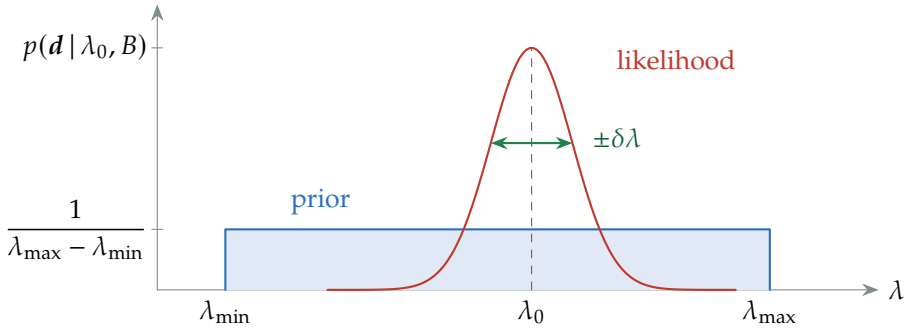


Figure 5.43: Mr B's prior and likelihood for his parameter. The prior (blue box) spreads unit probability evenly over the range $\lambda_{\min}$ to $\lambda_{\max}$; its height is drawn exaggerated. The likelihood (red) is large only within about $\delta\lambda$ of its peak. The evidence is the likelihood averaged over the box, so it is roughly the peak height times the fraction of the box the peak occupies.

With these two assumptions the average over Mr B's prior can be done by hand:

$$\begin{aligned} p(\boldsymbol{d} \mid B) &\overset{(1)}{=} \int p(\boldsymbol{d} \mid \lambda, B)\, p(\lambda \mid B)\, \mathrm{d}\lambda \overset{(2)}{=} \frac{1}{\lambda_{\max} - \lambda_{\min}} \int_{\lambda_{\min}}^{\lambda_{\max}} p(\boldsymbol{d} \mid \lambda, B)\, \mathrm{d}\lambda \\ &\overset{(3)}{=} \frac{p(\boldsymbol{d} \mid \lambda_0, B)}{\lambda_{\max} - \lambda_{\min}} \int_{-\infty}^{\infty} e^{-(\lambda-\lambda_0)^2/2\delta\lambda^2}\, \mathrm{d}\lambda \overset{(4)}{=} \frac{p(\boldsymbol{d} \mid \lambda_0, B)\; \delta\lambda\sqrt{2\pi}}{\lambda_{\max} - \lambda_{\min}}. \end{aligned} \tag{5.48}$$

(1) The evidence (5.43) for Mr B.
(2) The uniform prior is a constant inside the range and zero outside, so it comes out of the integral and restricts the limits [Sivia (4.6)].
(3) The Gaussian form of the likelihood. We also assumed the peak lies well inside the range, so extending the limits to $\pm\infty$ loses almost nothing.
(4) The Gaussian integral $\int e^{-u^2/2s^2}\mathrm{d}u = s\sqrt{2\pi}$ [Sivia (4.7)].

Put this into the odds (5.45) with $M_1 = A$ and $M_2 = B$:

$$\frac{\mathrm{P}(A \mid \boldsymbol{d})}{\mathrm{P}(B \mid \boldsymbol{d})} = \underbrace{\frac{\mathrm{P}(A)}{\mathrm{P}(B)}}_{\text{prior odds}} \times \underbrace{\frac{p(\boldsymbol{d} \mid A)}{p(\boldsymbol{d} \mid \lambda_0, B)}}_{\text{best-fit ratio}} \times \underbrace{\frac{\lambda_{\max} - \lambda_{\min}}{\delta\lambda\sqrt{2\pi}}}_{\text{Occam factor}} . \tag{5.49}$$

This is Sivia's (4.8). Each of the three factors answers its own question. The first says what we thought before the data. The second says how well each theory does at its best; since Mr B can always tune λ, this ratio can only favour him. The third is new. It compares the range of λ that Mr B's theory allowed before the data, $\lambda_{\max} - \lambda_{\min}$, with the range the data allow, about $\delta\lambda$. The first is usually much wider, so this factor is large and favours Mr A. It is called the **Occam factor** (Sivia writes Ockham), because it puts a number on William of Ockham's rule that "it is vain to do with more what can be done with fewer" [Sivia p. 84].

Working definition 5.10 (Occam factor)

For a model with parameters, the **Occam factor** is the ratio of its evidence to its best-fit likelihood,

$$\frac{Z}{\mathcal{L}_{\max}} \approx \frac{\text{volume of parameter space allowed by the posterior}}{\text{volume allowed by the prior}} \leq 1,$$

[the fraction of the prior's parameter range that the data leave standing]. The evidence is the best fit times the Occam factor.

Reading (5.49) in a few limits tells us how the razor behaves [Sivia pp. 84–86]:

- *Good data, genuine effect.* If Mr B's extra parameter is really needed, the best-fit ratio is tiny (his fit is much better) and it overwhelms the Occam factor. Most of the time the goodness of fit decides.
- *Good data, no effect.* If Mr B's fit is not much better, the best-fit ratio stays near one, while more data make $\delta\lambda$ smaller and the Occam factor larger. Mr A is favoured ever more strongly as the data improve.
- *Poor data.* If the data barely constrain λ, then $\delta\lambda$ approaches the prior range and the likelihood is flat, so both the best-fit ratio and the Occam factor approach one. The posterior odds return to the prior odds: poor data have taught us nothing.
- *Unbounded prior.* If $\lambda_{\max} - \lambda_{\min} \to \infty$, the Occam factor is infinite and Mr B always loses. Jeffreys worried about this. In practice the range is always bounded by what we know. A new fifth force with a huge coupling would have been noticed long ago, so its strength has an upper bound before any new experiment.

Two variations make the razor sharper.

- *Both theories have one parameter.* Give Mr A a parameter μ, with a uniform prior and a Gaussian likelihood of width $\delta\mu$, just like Mr B's. Each evidence now carries its own Occam factor. With equal prior ranges the ranges cancel, and the posterior odds are the best-fit ratio times $\delta\mu/\delta\lambda$ [Sivia (4.9)]. So for equally good fits the theory with the *larger* error bar wins. That sounds backwards, but a larger error bar means that more of the theory's parameter values agree with the data.
- *The same theory, two prior ranges.* Let Mr A and Mr B hold the same theory, but let Mr A claim a narrower prior range, still wide enough to contain the fit. Everything cancels except the ranges, and the odds are the ratio of the ranges in Mr A's favour. He predicted the parameter more precisely, and the evidence rewards him for it.

None of this touches parameter estimation. Inside Mr B's theory the posterior for λ is likelihood times prior divided by the evidence [Sivia (4.10)]. The prior is the constant $1/(\lambda_{\max} - \lambda_{\min})$, and it appears in both the numerator and the evidence (5.48), so it cancels; so does the peak height. What is left is the Gaussian $\lambda = \lambda_0 \pm \delta\lambda$, whatever the range, as long as the range contains the peak. The range matters only when we compare models. The same equation serves both purposes; we are asking it different questions.

Example: the counting experiment again. In Section 5.15.2 five counts on an expected background of $b = 2$ gave a best-fit likelihood ratio of 4.862 for a signal, but a Bayes factor of only 2.669. The Occam factor explains the difference. We need the width δs of the likelihood. The log-likelihood is $\ln p(n \mid s) = -(b+s) + n\ln(b+s) - \ln n!$. Its second derivative is $-n/(b+s)^2$, and at the peak $\hat{s} = n - b$, where $b + s = n$, this is $-1/n$. A Gaussian of width δs has second derivative of its logarithm $-1/\delta s^2$, so $\delta s = \sqrt{n} = \sqrt{5}$. The prior range is $s_{\max} = 10$, and (5.48) gives the Occam factor

$$\frac{\delta s\sqrt{2\pi}}{s_{\max}} = \frac{\sqrt{10\pi}}{10} \approx 0.56,$$

close to the exact ratio 0.5489 of evidence to best fit. The Gaussian approximation is rough for a Poisson likelihood with five counts, but the size of the penalty is right: a signal prior of width 10 against a data width of about 2.2.

Example: must be fair, or anything goes? Kruschke's version compares two coin factories that differ only in how widely they spread their coin biases [Kruschke §10.5]. The "must-be-fair" factory makes coins with $\theta \sim \text{Beta}(500, 500)$, all very close to 0.5. The "anything-goes" factory makes $\theta \sim \text{Beta}(1, 1)$, uniform on $[0, 1]$. The second model is the more flexible: it contains a coin matching any observed fraction of heads. The evidence of each is (5.46). For $z = 15$ heads in $N = 20$ tosses the Bayes factor of fair against anything is 0.3229: the fair model has no bias near 0.75, and loses. For $z = 11$ heads it is 3.337: now the fair model wins, although the anything model contains the bias 0.55 that matches the data exactly. The anything model pays for spreading its prior over biases like 0.1 and 0.9 that these data rule out. Figure 5.44 shows the pattern for every possible z: the simpler model wins whenever the data are consistent with it, here for $7 \le z \le 13$.

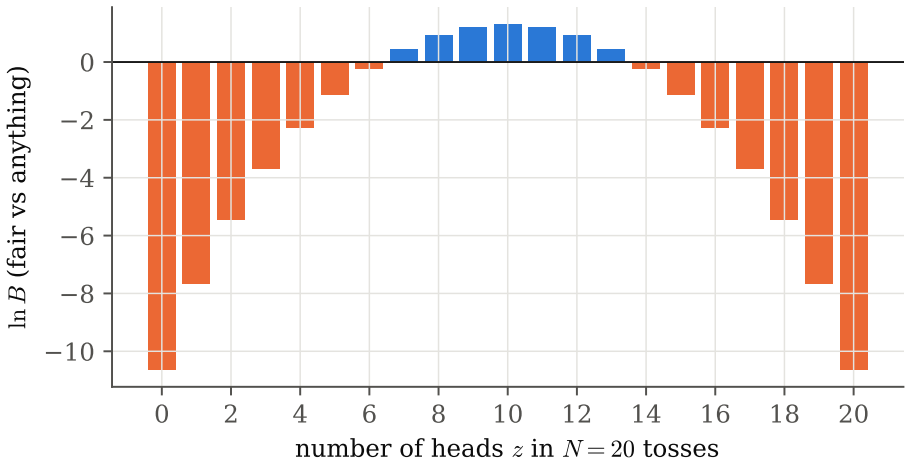


Figure 5.44: Natural Occam's razor for coins. For each possible number of heads in twenty tosses, the bar is the log Bayes factor of the must-be-fair factory against the anything-goes factory. Near ten heads the simple model wins by a modest factor (blue); far from ten it loses by a large one (orange), because it simply cannot produce such data.

The picture is asymmetric, and that is typical. A simple model that is wrong is punished hard, because it has no parameter value near the data. A flexible model that is unnecessary is punished gently, by a factor of a few, the size of its Occam factor.

5.16.2 Many parameters: the Laplace approximation

The same argument works in D dimensions. Near its peak $\hat{\boldsymbol{\theta}}$ the logarithm of the likelihood is approximately quadratic,

$$\ln \mathcal{L}(\boldsymbol{\theta}) \approx \ln \mathcal{L}_{\max} - \tfrac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^{\mathsf{T}}\mathbf{F}\,(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}),$$

where $\mathbf{F}$ is the matrix of second derivatives of $-\ln\mathcal{L}$ at the peak [the curvature of the log-likelihood, the Fisher matrix of Chapter 10 evaluated on these data]. Its inverse is the covariance of the Gaussian approximation to the posterior. If the prior $\pi(\boldsymbol{\theta})$ changes little across the peak,

$$\begin{aligned} Z &\overset{(1)}{=} \int \mathcal{L}(\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\,\mathrm{d}^D\boldsymbol{\theta} \overset{(2)}{\approx} \mathcal{L}_{\max}\,\pi(\hat{\boldsymbol{\theta}}) \int e^{-\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^{\mathsf{T}}\mathbf{F}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})}\,\mathrm{d}^D\boldsymbol{\theta} \\ &\overset{(3)}{=} \mathcal{L}_{\max}\,\pi(\hat{\boldsymbol{\theta}})\,\frac{(2\pi)^{D/2}}{\sqrt{\det\mathbf{F}}} \overset{(4)}{=} \mathcal{L}_{\max} \times \frac{(2\pi)^{D/2}\sqrt{\det\mathbf{C}}}{V_\pi}. \end{aligned} \tag{5.50}$$

(1) The evidence, with $\mathcal{L}(\boldsymbol{\theta}) = p(\boldsymbol{d}\,|\,\boldsymbol{\theta}, M)$ and $\pi(\boldsymbol{\theta}) = p(\boldsymbol{\theta}\,|\,M)$.
(2) The quadratic expansion of $\ln\mathcal{L}$, and the prior taken out at the peak.
(3) The D-dimensional Gaussian integral. Rotate to the eigenvectors of the symmetric positive matrix $\mathbf{F}$; the integral factorises into D one-dimensional integrals $\sqrt{2\pi/f_k}$, one per eigenvalue f_k, and the product of the eigenvalues is $\det\mathbf{F}$.
(4) For a uniform prior on a region of volume V_π, $\pi(\hat{\boldsymbol{\theta}}) = 1/V_\pi$; and $\mathbf{C} = \mathbf{F}^{-1}$, so $1/\sqrt{\det\mathbf{F}} = \sqrt{\det\mathbf{C}}$.

This is the **Laplace approximation** to the evidence [Sivia (4.20)], [32, eqs. (24)–(25)]. The factor $(2\pi)^{D/2}\sqrt{\det\mathbf{C}}$ is the volume of the posterior's "error ellipsoid", so the Occam factor of (5.49) generalises to the ratio of the posterior volume to the prior volume. In one dimension it is $\sqrt{2\pi}\,\delta\lambda$ over $\lambda_{\max}-\lambda_{\min}$ again.

Key idea: Evidence ≈ best fit × Occam factor

$$Z \;\approx\; \mathcal{L}_{\max} \times \frac{V_{\text{posterior}}}{V_{\text{prior}}}.$$

A model is rewarded for fitting the data well and penalised by the fraction of its prior parameter space that the data rule out. Every parameter the data constrain adds a factor $\sqrt{2\pi}\,\sigma_k/(\text{prior range}_k)$ (for uncorrelated parameters); a model is favoured only if its better fit pays for all of them.

Two remarks keep the formula honest. First, a parameter the data do not constrain at all has a posterior as wide as its prior, contributes a factor one, and costs nothing. Adding the price of fish at Billingsgate market to a cosmological model changes no prediction, so the likelihood is flat in it and the evidence is unchanged however wide its prior [HobsonBMC p. 85], [32, p. 16]. What the evidence penalises is not the number of parameters but wasted predictive range: parameter values the model allowed and the data rule out. Hobson and collaborators put it as "the Bayesian evidence rewards model predictiveness" [HobsonBMC p. 83]. Second, if the likelihood has several separate peaks, each contributes its own term. With M identical bumps whose labels can be permuted, the likelihood has $M!$ equivalent peaks and the Laplace sum

must count all of them [Sivia (4.20)]. An exact integral over the whole prior box counts them automatically.

5.16.3 A line, or a line plus a bump?

Back to the forty points. The two models are

$$M_0 \text{ (line)}: \quad y_i = a + b x_i + \epsilon_i,$$
$$M_1 \text{ (line + bump)}: \quad y_i = a + b x_i + A\, e^{-(x_i-\mu)^2/2W^2} + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1),$$

with the bump width $W = 0.5$ known. The priors are part of the models and are fixed before looking at the data: a uniform on $[0, 20]$ and b uniform on $[-2, 2]$ in both models; in M_1 also A uniform on $[0, A_{\max}]$ with $A_{\max} = 10$ (a generous ceiling, ten times the noise) and μ uniform on $[0, 10]$ (the range we measured). We simulate two data sets, one with a real bump of height 2.5 at $\mu = 6.2$ and one with no bump, both on the line $y = 5 + 0.3x$ (Figure 5.45).

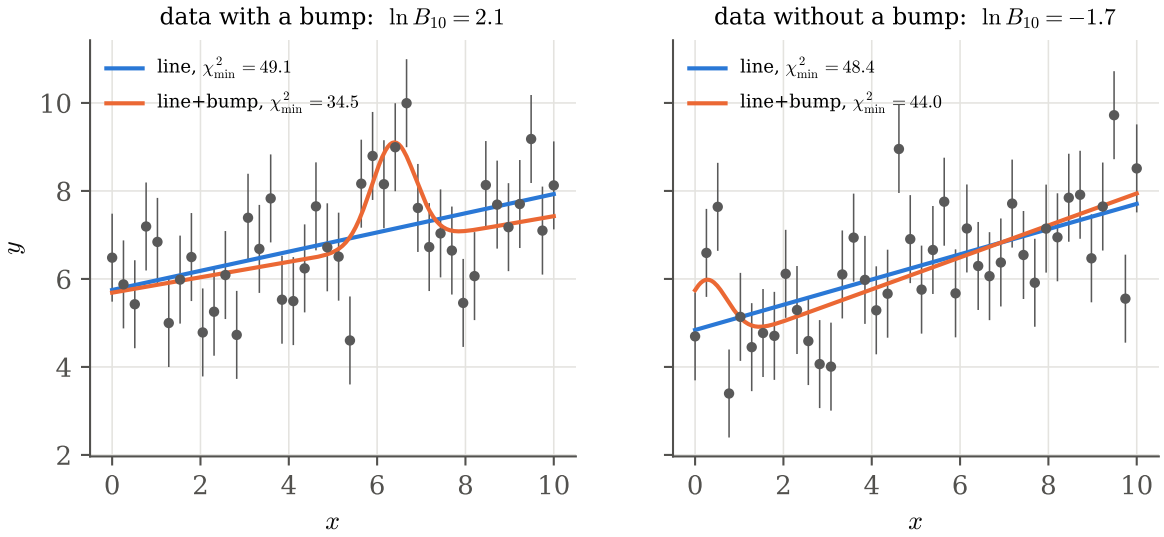


Figure 5.45: Two simulated data sets of forty points with unit error bars, and the best fits of both models. Left: a bump of height 2.5 is present near $x = 6.2$; the bump model finds it and its χ^2 drops by 14.53. Right: there is no bump; the bump model still finds a small improvement by fitting a noise fluctuation. The log Bayes factor of bump against line is printed above each panel.

Integrating out the line exactly. The evidence of M_1 is a four-dimensional integral over (a, b, A, μ). The line parameters a, b enter the model linearly, so the likelihood is Gaussian in them and their integral can be done exactly. That leaves a two-dimensional integral over (A, μ) for the grid. The question for each bump is: once the bump (A, μ) is subtracted, how well can the best line fit what remains? Write the data as the vector $\boldsymbol{y}$ (here $N = 40$), the two line shapes as the columns of the $N \times 2$ matrix $X = (\mathbf{1}, \boldsymbol{x})$, and the bump shape at position μ as the vector $\boldsymbol{g}_\mu$. For fixed (A, μ) the line fits $\boldsymbol{y} - A\boldsymbol{g}_\mu$. What it cannot fit is the part of $\boldsymbol{y} - A\boldsymbol{g}_\mu$ that points in directions no line can reach, $P_\perp(\boldsymbol{y} - A\boldsymbol{g}_\mu)$, where $P_\perp = \mathbb{1} - X(X^\mathsf{T}X)^{-1}X^\mathsf{T}$ is the projector onto those directions. Writing $\boldsymbol{r} = P_\perp \boldsymbol{y}$ and $\boldsymbol{u}_\mu = P_\perp \boldsymbol{g}_\mu$,

$$\chi^2_{\min}(A,\mu) - \chi^2_{\min,\text{line}} \overset{(1)}{=} \|\boldsymbol{r} - A\boldsymbol{u}_\mu\|^2 - \|\boldsymbol{r}\|^2 \overset{(2)}{=} -2A\, \boldsymbol{r}\cdot\boldsymbol{u}_\mu + A^2\, \boldsymbol{u}_\mu\cdot\boldsymbol{u}_\mu. \tag{5.51}$$

(1) With $\sigma = 1$, the minimum of χ^2 over the line is the squared length of the residual the line cannot absorb; for the line model alone that residual is $\boldsymbol{r}$.

(2) Expand the square: $\|\boldsymbol{r} - A\boldsymbol{u}\|^2 = \|\boldsymbol{r}\|^2 - 2A\,\boldsymbol{r}\cdot\boldsymbol{u} + A^2\|\boldsymbol{u}\|^2$.

The integral over (a, b) at fixed (A, μ) is the same Gaussian integral in both models, because both use the same line shapes and the same prior box for the line. It cancels in the Bayes factor, which is therefore

$$B_{10} = \frac{1}{A_{\max}\cdot 10}\int_0^{A_{\max}}\int_0^{10} \exp\!\left[-\tfrac{1}{2}\left(\chi^2_{\min}(A,\mu) - \chi^2_{\min,\text{line}}\right)\right] \mathrm{d}\mu\,\mathrm{d}A. \tag{5.52}$$

The Bayes factor averages the likelihood ratio of each candidate bump to the best straight line over the prior range of bump heights and positions. In taking that average we assumed that the prior bounds for the line parameters enclose their likelihood peak. The bounds used here extend many standard deviations beyond it.

Listing 5.10: The Bayes factor of bump against line on a grid (`code/ch05/lib_bump.py`)

```
72 def dchi2_grid(y, A, mu):
73     """chi2_min over (a, b) at each (A, mu), minus the line's chi2_min.  Shape (len(A),
       len(mu))."""
74     ry = PPERP @ y
75     U = PPERP @ bump(X[:, None], mu[None, :])                    # projected bump shapes,
       (N, n_mu)
76     uy = ry @ U / SIGMA**2                                       # (n_mu,)
77     uu = np.sum(U * U, axis=0) / SIGMA**2
78     return -2 * A[:, None] * uy[None, :] + A[:, None] ** 2 * uu[None, :]
79
80
81 def ln_bayes_factor(y, a_max=A_MAX, mu_fixed=None, nA=2001, nmu=2001, return_grid=False):
82     """ln B_10 = ln Z(bump) - ln Z(line), priors A ~ U[0, a_max], mu ~ U[MU_RANGE] (or mu
       fixed)."""
83     A = np.linspace(0.0, a_max, nA)
84     mu = np.array([mu_fixed]) if mu_fixed is not None else np.linspace(*MU_RANGE, nmu)
85     dc = dchi2_grid(y, A, mu)
86     f = np.exp(-0.5 * dc)                                        # likelihood ratio to the
       best line
87     inner = f[:, 0] if mu_fixed is not None else np.trapezoid(f, mu, axis=1) / L_MU
88     lnB = np.log(np.trapezoid(inner, A) / a_max)
89     if return_grid:
90         return lnB, A, mu, dc
91     return lnB
```

Lines 74–78 build (5.51) for every (A, μ) on the grid at once: the projected data $\boldsymbol{r}$, the projected bump shapes $\boldsymbol{u}_\mu$ (one column per position), their dot products, and the quadratic in A. Line 86 exponentiates to get the likelihood ratio to the best line; lines 87–88 average it over μ and then over A with the trapezoid rule, which is (5.52). Working with the ratio to the best line keeps the numbers of order one; the raw likelihoods are about e^{-60}.

The data with a bump. The line has $\chi^2_{\min} = 49.07$ for 38 degrees of freedom, the bump model 34.54 at $\hat{A} = 2.307$, $\hat{\mu} = 6.373$. The best bump improves the fit by $\Delta\chi^2 = 14.53$, a best-fit likelihood ratio of $e^{7.265}$. The Bayes factor is $\ln B_{10} = 2.096$ ($B_{10} = 8.133$; a grid of half the resolution changes $\ln B_{10}$ by 1.236×10^{-8}). The difference, $\ln B_{10} - \frac{1}{2}\Delta\chi^2 = -5.169$, is the logarithm of the Occam factor. Here the bump is favoured, but by odds of 8.133 to 1, not by the $e^{7.265}$ that the fit improvement alone suggests.

The Laplace approximation (5.50) lets us read the Occam factor off the map of Figure 5.46. At the best fit the posterior standard deviations are $\sigma_A = 0.6053$ and $\sigma_\mu = 0.1415$, so the posterior area in the (A, μ) plane is $2\pi\sqrt{\det \mathbf{C}_{A\mu}} = 0.5381$, against a prior area of $A_{\max}\times 10 = 100$. Their ratio, 0.005381, is the Occam factor, and the Laplace estimate of the Bayes factor is $\ln B_{10} = 2.04$,

close to the exact 2.096. The bump model allowed 100 square units of bumps and the data kept only 0.5381 of them.

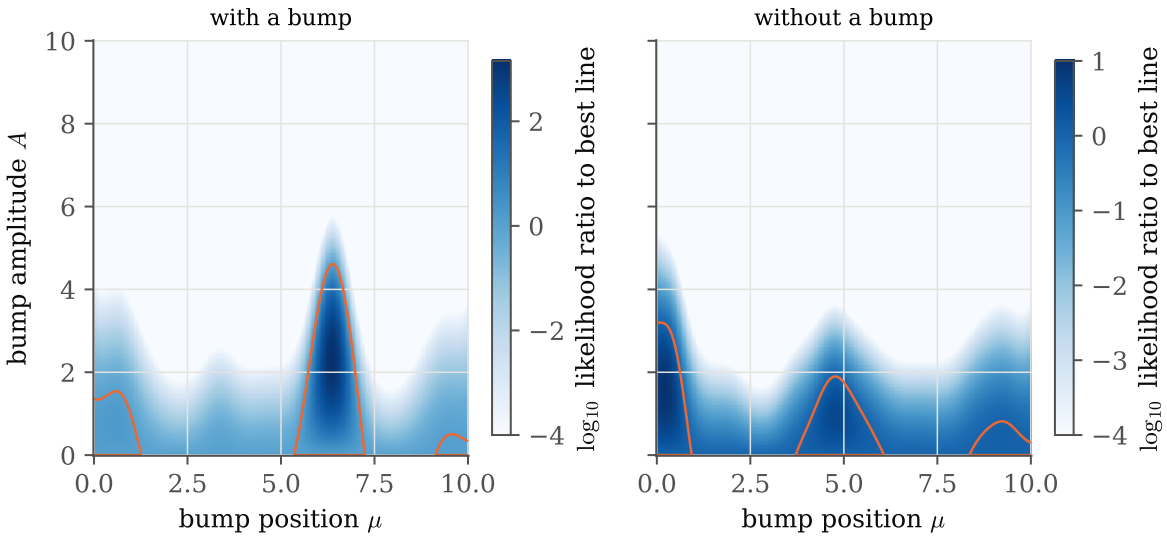


Figure 5.46: The likelihood of every bump in the prior box, relative to the best line, in powers of ten. With a bump in the data (left) a small island near $\mu = 6.373$, $A = 2.307$ is far better than the line; almost all the rest of the box is worse. Without a bump (right) nothing is much better than the line. The orange contour marks a likelihood ratio of one. The Bayes factor is the average of this map over the box.

The data without a bump. Here the bump model still finds something: a noise fluctuation at $\hat{\mu} = 0.2214$ improves χ^2 from 48.37 to 43.98. A χ^2 drop of a few units for two extra parameters is what noise does. The evidence agrees: $\ln B_{10} = -1.675$, so the line is favoured by a factor $1/0.1873$. The Occam factor, $e^{-3.868}$, outweighs the modest improvement of the fit. For both data sets the razor cuts in the right direction, and it needed no rule beyond the sum and product rules.

If we had known where to look. Not knowing where the bump sits has a price, and the evidence charges it automatically. Suppose theory had predicted the resonance at $\mu = 6.2$, so that M_1 has only the amplitude free. Then the prior no longer wastes probability on positions the data rule out. For the data with a bump, $\ln B_{10} = 4.668$ ($B_{10} = 106.5$), which is 13.09 times larger than with the position free. At the predicted position the amplitude is 2.18 ± 0.6018, a local significance of 3.622σ. So not knowing the position costs a factor of 13.09. This is the look-elsewhere effect of Chapter 4. There it needed a separate trials factor; here it appears by itself, as the Occam factor of the parameter μ.

5.17 Reading a Bayes factor

With three years of WMAP data and the large-scale-structure surveys of the time, a cosmological model with a free spectral index n_s was found to have a log evidence 1.99 ± 0.26 higher than the Harrison–Zel'dovich model with $n_s = 1$ [HobsonBMC Table 4.4]. The same data gave $n_s = 0.958 \pm 0.016$, nearly three standard deviations from one. How impressed should we be by $\ln B = 2$? Is it the same statement as "three sigma", the significance of Chapter 4? And would another reasonable choice of prior have given a different number?

5.17.1 The Jeffreys scale

A Bayes factor multiplies the prior odds. With even prior odds and only two models, the posterior probability of the favoured model is

$$\mathrm{P}(M_1 \mid \boldsymbol{d}) = \frac{B_{10}}{1 + B_{10}}, \qquad \text{e.g.} \quad B_{10} = e^1 \approx 2.7 \;\to\; 0.73, \quad e^{2.5} \approx 12 \;\to\; 0.92, \quad e^5 \approx 150 \;\to\; 0.993, \tag{5.53}$$

from (5.45) with $\mathrm{P}(M_1) = \mathrm{P}(M_0)$ and $\mathrm{P}(M_1 \mid \boldsymbol{d}) + \mathrm{P}(M_0 \mid \boldsymbol{d}) = 1$. Jeffreys proposed words for these levels, and a version of his scale is quoted in almost every cosmology paper that computes an evidence [HobsonBMC §4.5], [Padilla Table II], [32, Table 1], [33]:

$\lvert\ln B\rvert$	odds	probability	Jeffreys	Hobson et al.	Trotta
< 1	$< 3:1$	< 0.75	not worth more than a bare mention	—	inconclusive
1–2.5	3 : 1 to 12 : 1	0.75–0.92	significant	weak	weak
2.5–5	12 : 1 to 150 : 1	0.92–0.993	strong to very strong	significant	moderate
> 5	$> 150:1$	> 0.993	decisive	strong	strong

The probability column assumes even prior odds and two models, and the table rounds $e^1 \approx 2.7$ up to 3 : 1. The thresholds are a convention, calibrated on experience, not a theorem. Three cautions come with them.

- *The scale is logarithmic.* Moving up one level needs roughly ten times more support. Evidence accumulates slowly [32, p. 14].
- *It applies to posterior odds.* The words are meant for the ratio of posterior model probabilities; they coincide with the Bayes factor only when the prior odds are even [HobsonBMC p. 90]. A Bayes factor of 20 for an exotic model that we believed at 1 : 1000 beforehand still leaves it at 1 : 50.
- *The priors are uncertain too.* Because a Bayes factor depends on the prior widths (Section 5.17.2), Hobson and collaborators soften Jeffreys' words: $\ln B$ between 1 and 2.5 is only "weak", between 2.5 and 5 "significant", above 5 "strong" (third column of the table) [HobsonBMC p. 90]. Trotta's column is similar. Kruschke's rule of thumb, "substantial" above $B = 3$, is the same first threshold [Kruschke p. 268].

An evidence computed numerically also has an error, like ±0.26 above, and a difference in $\ln B$ smaller than its error means nothing.

Example: is the primordial spectrum tilted? The two models are the tilted one (n_s free) and Harrison–Zel'dovich ($n_s = 1$). In the WMAP analysis [HobsonBMC pp. 91–93] the log Bayes factor of tilted against Harrison–Zel'dovich grew as the data improved:

- first-year WMAP with other data: $\ln B = -0.58 \pm 0.12$, inconclusive, with a slight preference for the simpler model;
- three-year WMAP alone: 0.34 ± 0.26, still inconclusive;
- three-year WMAP with other data: 1.99 ± 0.26, odds of about 7 : 1, "significant" in Jeffreys' words and "weak" in the softer scale.

Then inflation was taken seriously. Slow roll predicts a tilt *and* a tensor-to-scalar ratio r, so this model has two new parameters. Its verdict depended on the prior for r. With r uniform between 0 and 1, the n_s–r model came out at $\ln B = -1.45 \pm 0.45$, disfavoured against Harrison–Zel'dovich. With $\ln r$ uniform from -80 to 0 it came out at 1.90 ± 0.24, about as good as the n_s-only model. The data were identical. Why did the prior on r matter so much? The data allow only small r. A uniform prior on $[0, 1]$ puts most of its weight on larger values of r, which the data exclude, so it pays an Occam factor. The logarithmic prior puts most of its weight on tiny r, which the data cannot distinguish from zero. There the likelihood is flat, so this prior pays almost nothing: this is the Billingsgate effect of Section 5.16.2.

5.17.2 The prior width sets the price

The wider the prior range of an extra parameter, the less the data favour it. The reason is the Occam factor of (5.49): it is the data's range divided by the prior range. So once the prior range is much wider than the likelihood, the Bayes factor in favour of the extra parameter falls in proportion to the range: every factor of ten in the range subtracts $\ln 10 \approx 2.3$ from $\ln B_{10}$. Take the bump of Section 5.16.3 and vary only the prior range $A_{\max}$ of its amplitude (Figure 5.47). For the data with a bump, $\ln B_{10}$ first rises, because a prior that stops below the fitted amplitude $\hat{A} = 2.307$ cannot reach the bump; it peaks once the range just covers it; then it falls along a straight line of slope -1 in $\ln A_{\max}$. At $A_{\max} = 100$ it is -0.2067, against 2.096 at $A_{\max} = 10$: the same bump, with the same local significance of 3.622σ, is now worth ten times less evidence. For the data without a bump $\ln B_{10}$ starts at zero (for a tiny range the two models are the same model) and falls steadily, reaching -3.977 at $A_{\max} = 100$.

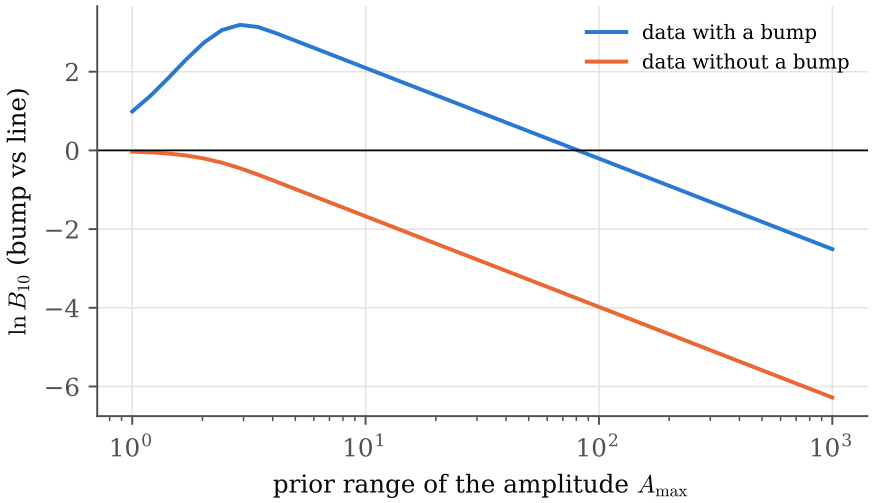


Figure 5.47: The Bayes factor of bump against line as the prior range of the bump amplitude is widened, for the two data sets of Figure 5.45. Past the fitted amplitude both curves fall by $\ln 10$ per decade of $A_{\max}$: each widening of the prior spreads the model's probability over more bumps the data rule out.

So which range is right? The answer has to come from physics, before the data: the largest resonance the experiment would already have noticed, the range of n_s that slow-roll inflation allows ($0.8 \lesssim n_s \lesssim 1.2$), the coupling a fifth force could have without being seen [Sivia p. 85], [32, §4.5]. If several plausible ranges give the same qualitative verdict, the verdict is robust. If not, the honest report is the Bayes factor as a function of the range, which is what Figure 5.47 is. In the (information, significance) plane of Section 5.17.3 a change of prior width is a horizontal shift [32, §4.4].

Example: two vague priors, one posterior, two verdicts. Two priors that both claim to know nothing can give opposite verdicts [Kruschke §10.6]. A coin gives $z = 65$ heads in $N = 100$ tosses. Model one is must-be-fair, $\theta \sim \text{Beta}(500, 500)$, where θ is the probability of heads. Model two is anything-goes, and we must choose a vague prior for it.

- With $\text{Beta}(1, 1)$, uniform, the Bayes factor of fair against anything is 0.1253: anything-goes wins by $8 : 1$.
- With the Haldane prior $\text{Beta}(0.01, 0.01)$, recommended by some statisticians as maximally uninformative, the Bayes factor is 5.728: must-be-fair wins by nearly $6 : 1$.

The verdict reversed. Yet the posteriors of θ under the two vague priors are practically the same: their 95% HDIs are $[0.5543, 0.7382]$ and $[0.5564, 0.7418]$, and both exclude 0.5. How can the posterior stay put while the evidence flips? The evidence (5.43) is the likelihood averaged over the prior. The Haldane prior puts almost all its probability extremely close to $\theta = 0$ and $\theta = 1$. So its density near $\theta = 0.65$, where the likelihood lives, is tiny, and the average of the likelihood over the prior is tiny with it. The posterior, by contrast, is renormalised: it does not care about the overall level of the prior near 0.65, only about its shape there, and there both priors are nearly flat.

Lambert shows the same thing for a binomial sample [Lambert §10.6, Fig. 10.12]: fifteen counts, each out of ten trials, with a $\text{Beta}(a, a)$ prior on the success probability (Figure 5.48). As a goes from 0.01 to 1 to 10 the posterior hardly moves (mean 0.5, standard deviation 0.04069 against 0.04043), while $\ln p(\boldsymbol{d})$ goes from -32.78 to -28.86 to -27.66: a factor 50.36 in the evidence between the first two "vague" choices.

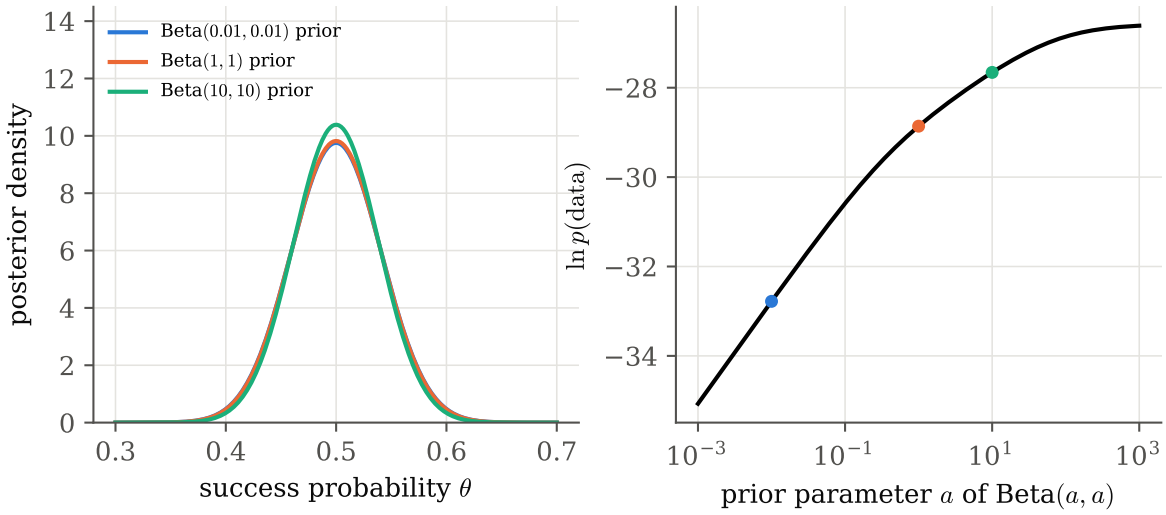


Figure 5.48: Left: the posterior of a success probability under three symmetric Beta priors, for fifteen binomial counts out of ten trials each; the three curves nearly coincide. Right: the log evidence of the same data as a function of the prior parameter a; the three dots are the three priors on the left. The posterior barely notices the prior, the evidence notices it a lot.

Kruschke's remedy is to put the competing models on an equal footing: inform the priors of *both* models with the same small, representative part of the data, and compute the Bayes factor from the rest [Kruschke §10.6.1]. Why does this help? Even a few data points overwhelm a vague prior and bring each model's parameters "into the ballpark", so the arbitrary choice of vague prior no longer matters. For the coin, use 6 heads in the first 10 tosses to update both priors, then compare on the remaining 90. The Bayes factor of fair against anything becomes 0.05571 starting from the uniform prior and 0.05748 starting from the Haldane prior. The two vague

priors now agree, and both say the coin is not fair, as the HDIs did. Intrinsic and fractional Bayes factors are more systematic versions of the same idea.

Watch out: Vague priors are harmless for estimation, not for comparison

A wide or "uninformative" prior is usually safe when estimating a parameter, because the likelihood dominates the posterior. In model comparison the overall height of the prior where the likelihood lives enters directly, and changing from one vague prior to another can change the Bayes factor by orders of magnitude. An improper prior, one that cannot be normalised (flat on the whole real line), gives an evidence with an arbitrary constant in it, so Bayes factors with improper priors on the extra parameters are undefined [Wasserman p. 185], [HobsonBMC §1.3].

5.17.3 Lindley's paradox

The simplest nested comparison can be done exactly, and it holds a surprise. We have a measurement $\hat{\theta}$ of a quantity θ with Gaussian error σ. Model M_0 says the effect is absent, $\theta = 0$. Model M_1 says it is present, with a size we can only bound: $\theta \sim \mathcal{N}(0, \Sigma^2)$ with Σ the scale of effects the theory allows. Call $\lambda = |\hat{\theta}|/\sigma$ the significance of the measurement, the "number of sigma".

The event: the value $\hat{\theta}$. *The hypotheses:* M_0 and M_1. *How probable was $\hat{\theta}$ under each?* Under M_0 it is the Gaussian density $\mathcal{N}(\hat{\theta}; 0, \sigma^2)$. Under M_1 we average over θ:

$$p(\hat{\theta} \mid M_1) \overset{(1)}{=} \int \mathcal{N}(\hat{\theta}; \theta, \sigma^2)\, \mathcal{N}(\theta; 0, \Sigma^2)\, \mathrm{d}\theta \overset{(2)}{=} \mathcal{N}(\hat{\theta}; 0, \sigma^2 + \Sigma^2),$$

$$B_{01} = \frac{p(\hat{\theta} \mid M_0)}{p(\hat{\theta} \mid M_1)} \overset{(3)}{=} \sqrt{\frac{\sigma^2 + \Sigma^2}{\sigma^2}}\, \exp\!\left[-\frac{\hat{\theta}^2}{2\sigma^2} + \frac{\hat{\theta}^2}{2(\sigma^2 + \Sigma^2)}\right] \overset{(4)}{=} \sqrt{1 + \frac{\Sigma^2}{\sigma^2}}\, \exp\!\left[-\frac{\lambda^2}{2}\, \frac{\Sigma^2}{\sigma^2 + \Sigma^2}\right]. \quad (5.54)$$

(1) The evidence of M_1: the likelihood of the measurement averaged over the prior of the effect.

(2) The measurement is the true effect plus independent Gaussian noise, so it is a sum of two independent Gaussians of variances Σ^2 and σ^2, and its density is the Gaussian of the summed variance (the convolution of two Gaussians; complete the square in θ to check).

(3) Ratio of the two Gaussian densities: the prefactors give the square root, the exponents subtract.

(4) Combine the exponents over the common denominator and write $\hat{\theta}^2/\sigma^2 = \lambda^2$.

This is Trotta's (21) [32, p. 13]. When the data are much more precise than the prior, $\sigma \ll \Sigma$, the last factor becomes $e^{-\lambda^2/2}$ and

$$\ln B_{01} \approx \ln \frac{\Sigma}{\sigma} - \frac{\lambda^2}{2} \quad (5.55)$$

[32, eq. (26)]: the Occam term $\ln(\Sigma/\sigma)$, the logarithm of the ratio of prior to posterior width, against the goodness-of-fit term $\lambda^2/2$. The more the data have taught us (Σ/σ large), the more significance is needed to prefer the extra parameter. Figure 5.49 maps (5.54) over the plane of information $I_{10} = \log_{10}(\Sigma/\sigma)$ and significance λ. For moderately informative data, I_{10} between 1 and 2, the measurement must lie about 4σ from zero before the extra parameter is favoured robustly; results at 2–3σ are inconclusive or even favour the simpler model [32, §4.4].

What happens if the significance stays fixed while the data grow? With n measurements of unit noise, $\sigma = 1/\sqrt{n}$; take $\Sigma = 1$ and a result exactly at $\lambda = 1.96$, the two-sided $p = 0.05$ of Chapter 4. The Bayes factor (5.54) is $\sqrt{1+n}\, \exp[-\tfrac{1}{2}(1.96)^2\, n/(1+n)]$, which grows like $\sqrt{n}$. With even prior odds the posterior probability of the null (Figure 5.50) is 0.3512 at $n = 1$,

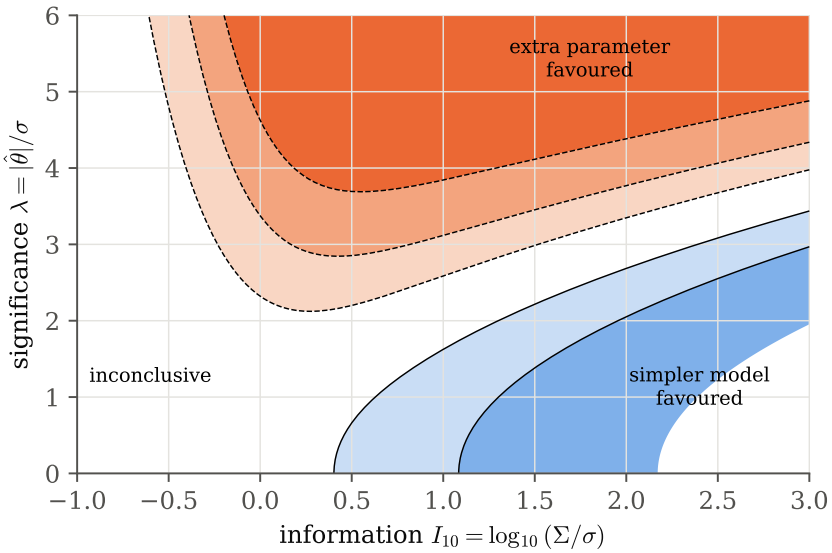


Figure 5.49: The Bayes factor for one extra parameter with Gaussian prior and likelihood, over the plane of information (how much narrower the likelihood is than the prior, in powers of ten) and significance (how many standard deviations the measurement lies from zero). Contours are at $|\ln B| = 1, 2.5, 5$. Blue: the simpler model is favoured; orange: the extra parameter is favoured; white: inconclusive. A wider prior moves a result to the right.

0.8228 at $n = 10^3$ and 0.9932 at $n = 10^6$. At a fixed "1.96σ" the data favour the null once n exceeds about 41.25, and overwhelmingly so for large samples. Even at $p = 0.01$ ($\lambda = 2.58$) the probability of the null at $n = 10^6$ is 0.9731. This is **Lindley's paradox** [34]: the same p-value means evidence *against* the null in a small experiment and evidence *for* it in a large one.

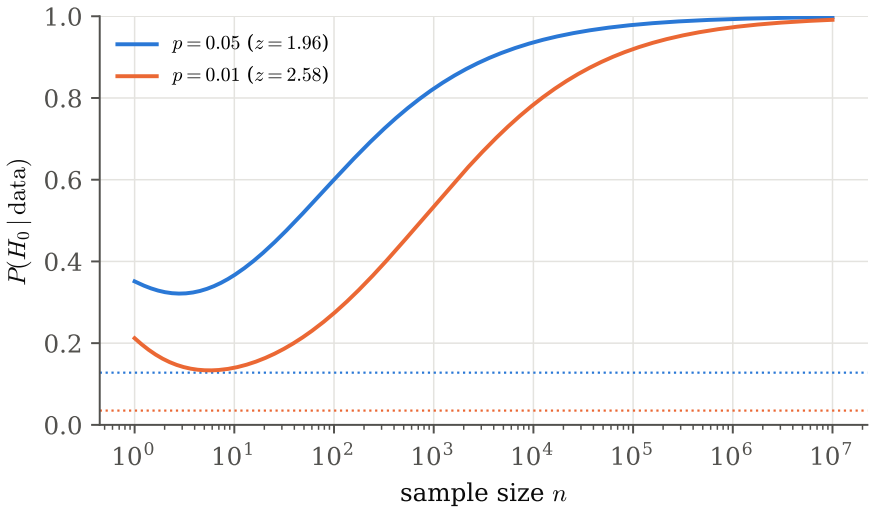


Figure 5.50: Lindley's paradox. The posterior probability of "no effect" for a result held at fixed significance (1.96σ, blue; 2.58σ, orange) as the sample size grows, with a unit-width prior on the effect and even prior odds. The dotted lines are the lowest value any prior could give (the bound of Section 5.17.4).

There is nothing paradoxical in the mechanism. At fixed significance the measured effect $\hat{\theta} = 1.96/\sqrt{n}$ shrinks as n grows. The alternative spread its prior over effects of size $\Sigma = 1$, and the data increasingly say that the effect, if there is one, is far smaller than anything M_1 expected. The measurement becomes a successful prediction of M_0. The p-value, which only asks how

surprising the data are under M_0, cannot see this. Large modern data sets (Planck maps, LHC luminosities, years of detector time) are exactly where the two answers part company, which is why "how many sigma" is not by itself a good measure of the evidence for a new effect [32, §5.4].

5.17.4 What a p-value can and cannot say

Wasserman's version of the Bayesian test makes the link to Chapter 4 explicit [Wasserman pp. 184–185]. Take the data $x^n = (x_1, \dots, x_n)$ with likelihood $\mathcal{L}(\theta)$, and test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. Give each hypothesis prior probability $\frac{1}{2}$, and give θ a prior density $f(\theta)$ under H_1. Then

$$\mathrm{P}(H_0 \mid x^n) = \frac{\mathcal{L}(\theta_0)}{\mathcal{L}(\theta_0) + \int \mathcal{L}(\theta) f(\theta)\, \mathrm{d}\theta}, \tag{5.56}$$

which is (5.44) with $Z_0 = \mathcal{L}(\theta_0)$ and Z_1 the average of the likelihood over f. The prior f matters here in a way it did not in estimation, and it must be proper. Is there anything we can say whatever f is? Yes: a bound. An average cannot exceed the maximum, so $\int \mathcal{L} f \, \mathrm{d}\theta \leq \mathcal{L}(\hat{\theta})$, where $\hat{\theta}$ is the maximum-likelihood value. Putting this into the denominator,

$$\mathrm{P}(H_0 \mid x^n) \;\geq\; \frac{\mathcal{L}(\theta_0)}{\mathcal{L}(\theta_0) + \mathcal{L}(\hat{\theta})}. \tag{5.57}$$

For a Gaussian measurement at significance z, $\mathcal{L}(\hat{\theta})/\mathcal{L}(\theta_0) = e^{z^2/2}$, which at $z = 1.96$ is 6.826. So a result at "$p = 0.05$" leaves the null with posterior probability at least 0.1278, whatever prior the alternative is given. The most favourable prior for H_1 is a spike at $\hat{\theta}$, chosen after seeing the data, which is cheating; any honest prior does worse.

A less extreme and more useful bound restricts the alternative to priors that are symmetric about θ_0 and do not increase away from it. Sellke, Bayarri and Berger showed that over this class [35]

$$B_{10} \;\leq\; \bar{B}_{10} = -\frac{1}{e\, p \ln p} \qquad (p \leq e^{-1}) \tag{5.58}$$

for a p-value p [32, §4.5, eq. (27)]. It turns a significance into the best odds it could possibly support:

p-value	0.05	0.01	0.003	0.001	3×10^{-4}	5.7×10^{-7}
significance z	1.96	2.576	2.968	3.291	3.615	5.001
best odds $\bar{B}_{10}$	2.456	7.988	21.11	53.26	151.2	44889
$\ln \bar{B}_{10}$	0.8985	2.078	3.05	3.975	5.018	10.71

A "99%" result supports odds of 8 : 1 at best, short of the moderate level. Strong evidence at best needs about 3.6σ. A useful rule of thumb is to read an s-sigma result as roughly an $(s-1)$-sigma one [32, p. 18]. The 5σ discovery convention of particle physics corresponds to best odds of about $4.5 \times 10^4 : 1$.

Key idea: Significance is not evidence

A p-value is the probability, computed as if H_0 were true, of the observed result *together with every result lying further out in the same direction* (Section 4.1.3). A Bayes factor compares how probable the result actually observed was under H_0 and under a specified alternative. Even with the alternative's prior chosen as favourably as honesty allows, a 2σ result supports odds of at most about 2.5 : 1, and at fixed significance larger samples favour the null

(Lindley).

The difference lies in those further-out results. The p-value collects the probability of outcomes that were not observed, and it never asks what the alternative predicted. The Bayes factor uses only the observed outcome, under both hypotheses. When a paper reports "3σ evidence", the Bayes factor is the number that says how much the data should move us, and the prior on the new effect is what the claim is really about.

5.18 Nested models, shortcuts and model averaging

Almost every model comparison in cosmology has the shape of our bump hunt. The base model is ΛCDM with its six parameters. The candidate extension adds a parameter that is zero, or some other fixed value, in the base model: the spatial curvature Ω_k, the running $\mathrm{d}n_s/\mathrm{d}\ln k$, the tensor ratio r, the non-Gaussianity f_{NL}, the dark-energy equation of state through $w+1$, an extra number of neutrino species $N_\nu - 3.04$, and upwards of twenty more [HobsonBMC Table 4.2]. In the bump hunt the base model is the line and the extension adds a bump that disappears at $A = 0$. The simpler model sits inside the bigger one as a special value of the new parameter. Such models are called **nested**. Nesting makes the Bayes factor easier to compute, makes it natural to combine the models, and leads to some useful approximations.

5.18.1 The Savage–Dickey density ratio

Write the parameters of the bigger model M_1 as $(\boldsymbol{\phi}, \psi)$: the common parameters $\boldsymbol{\phi}$ (the line, a and b) and the new parameter ψ (the amplitude A). The simpler model M_0 is M_1 with ψ fixed at ψ_* ($A = 0$). Two conditions make the models genuinely nested [HobsonBMC (4.12)], [32, eq. (22)]:

1. the likelihoods agree at the special value: $p(\boldsymbol{d} \mid \boldsymbol{\phi}, \psi_*, M_1) = p(\boldsymbol{d} \mid \boldsymbol{\phi}, M_0)$;
2. the prior of M_1 factorises into the prior of M_0 for the common parameters times a prior for the new one: $p(\boldsymbol{\phi}, \psi \mid M_1) = p(\boldsymbol{\phi} \mid M_0)\, p(\psi \mid M_1)$.

Now ask a question about the bigger model alone: after the data, how probable is it that the new parameter sits at its special value? That is the posterior density of ψ at ψ_*, and it turns out to contain the Bayes factor:

$$
\begin{aligned}
p(\psi_* \mid \boldsymbol{d}, M_1) &\overset{(1)}{=} \int p(\boldsymbol{\phi}, \psi_* \mid \boldsymbol{d}, M_1)\, \mathrm{d}\boldsymbol{\phi} \overset{(2)}{=} \frac{1}{Z_1} \int p(\boldsymbol{d} \mid \boldsymbol{\phi}, \psi_*, M_1)\, p(\boldsymbol{\phi} \mid M_0)\, p(\psi_* \mid M_1)\, \mathrm{d}\boldsymbol{\phi} \\
&\overset{(3)}{=} \frac{p(\psi_* \mid M_1)}{Z_1} \int p(\boldsymbol{d} \mid \boldsymbol{\phi}, M_0)\, p(\boldsymbol{\phi} \mid M_0)\, \mathrm{d}\boldsymbol{\phi} \overset{(4)}{=} p(\psi_* \mid M_1)\, \frac{Z_0}{Z_1}.
\end{aligned}
\tag{5.59}
$$

(1) The marginal posterior of ψ: integrate the joint posterior over the common parameters.
(2) Bayes' theorem in M_1: likelihood times prior over the evidence Z_1, with the factorised prior of condition 2.
(3) Condition 1 replaces the likelihood by that of M_0; the prior density of ψ at ψ_* is a constant and comes out.
(4) The remaining integral is the evidence Z_0 of the simpler model, (5.43).

Rearranged, this is the **Savage–Dickey density ratio** [HobsonBMC (4.13)], [32, eq. (23)].

Working definition 5.11 (Savage–Dickey density ratio)

For nested models with a factorised prior,

$$B_{01} = \frac{Z_0}{Z_1} = \frac{p(\psi_* \mid \boldsymbol{d}, M_1)}{p(\psi_* \mid M_1)},$$

[how much the data raised (or lowered) the probability density of the new parameter at its "absent" value]. *Operationally:* run the bigger model only, histogram the samples of ψ near ψ_*, and divide the height of the normalised histogram there by the prior density.

The ratio has a direct reading. If the data pile up the posterior of the new parameter at its null value, higher than the prior put it, the simpler model gains; if they push the posterior away from it, the simpler model loses. Only a one-dimensional marginal density is needed, and its normalisation is a one-dimensional integral, so there is no high-dimensional evidence integral left to do.

Example: the bump amplitude at zero. Take the data without a bump. The new parameter is $\psi = A$ with $\psi_* = 0$ and prior density $p(A = 0 \mid M_1) = 1/A_{\max} = 0.1$. The bump position μ exists only in M_1, but at $A = 0$ the likelihood does not depend on it and its prior integrates to one, so the derivation goes through with μ integrated along with $\boldsymbol{\phi}$. On the grid, the posterior density of A at $A = 0$ times $A_{\max}$ is 5.338, which is the exact $B_{01} = 1/B_{10} = 5.338$ of Section 5.16.3. From a Markov chain (Section 5.19.3) we estimate the density by the fraction of samples with $A < 0.1$, divided by 0.1; this gives $B_{01} \approx 5.571 \pm 0.5489$, where the error comes from splitting the chain into twenty batches. How close this comes to the exact value varies from run to run, for two reasons. The estimate is an average over the first bin, and the posterior density changes across it (the grid average over the bin is 5.049); a finer bin needs more samples. And the chain moves into and out of the first bin slowly, so twenty batches can understate its random error.

For the data *with* a bump the same estimate is much harder. There $B_{01} = 1/8.133$, so the posterior density at $A = 0$ is small and few samples land there. A histogram is accurate only where samples land, so the estimate is noisy. This is the known weakness of the method: it is noisy exactly when the simpler model is disfavoured [HobsonBMC p. 89]. The factorisation condition can also fail quietly, for instance when a prior boundary of the common parameters depends on the new one.

5.18.2 Model averaging

Sometimes no model wins clearly, and we still want the best statement about a quantity θ that all the models share. Probability theory gives it directly: we do not know which model is true, so we sum the joint posterior over the models,

$$p(\theta \mid \boldsymbol{d}) = \sum_i p(\theta \mid \boldsymbol{d}, M_i)\ \mathrm{P}(M_i \mid \boldsymbol{d}), \tag{5.60}$$

the posterior of θ inside each model, weighted by the probability of that model [Kruschke §10.4], [HobsonBMC §4.2]. This is **Bayesian model averaging**. For the two coin factories of Section 5.15.2 it gave the averaged mean bias 0.6486, between the two factories' 0.4524 and 0.6905, with the tail-biased factory showing up as a long left tail of the averaged posterior.

A model in which θ is fixed still takes part. The line model fixes the amplitude A at 0, so it contributes a spike $\mathrm{P}(M_0 \mid \boldsymbol{d})\,\delta(\theta)$ to the average. Hobson and collaborators compare the averaged posterior to a quantum superposition that later data may collapse onto fewer models

[HobsonBMC p. 82]. Predictions of future data are averaged the same way. Basing them on the winning model alone ignores the probability that it is wrong.

Example: which dark-energy model? Five models of the dark-energy equation of state w were compared with the same supernova, CMB and large-scale-structure data [HobsonBMC Table 4.5]. Model I is ΛCDM, $w = -1$ exactly. Models II and III let w take a constant value with priors $-1 \le w \le -0.33$ and $-2 \le w \le -0.33$. Model IV gives w two parameters, $w(a) = w_0 + w_a(1-a)$ with $-2 \le w_0 \le -0.33$ and $-1.33 \le w_a \le 1.33$, and model V lets $w(a)$ vary with redshift, between -1 and 1 for $0 \le z \le 2$. The log Bayes factors against ΛCDM are 0, -1.3, -1.8, -2.0 and -4.1. With equal prior probability for each model, the posterior probabilities are $e^{\ln B_{iI}}$ divided by their sum:

$$\mathrm{P}(M_\mathrm{I} \mid \boldsymbol{d}) \overset{(1)}{=} \frac{1}{1 + e^{-1.3} + e^{-1.8} + e^{-2.0} + e^{-4.1}} = \frac{1}{1 + 0.273 + 0.165 + 0.135 + 0.017} = 0.63,$$

(1) Equation (5.44) with equal priors: each model's share is its evidence over the total, and dividing every evidence by that of ΛCDM turns the evidences into $e^{\ln B}$.

and the other four models get 0.17, 0.10, 0.09 and 0.01 in the same way. Every extension fits at least as well as ΛCDM at its best, since each contains $w = -1$. The evidence prefers ΛCDM because the extensions spread their priors over values of w the data exclude, and the more parameters or the wider the range, the larger the penalty. None of the extensions is decisively ruled out (no $|\ln B|$ exceeds 5), and the probabilities depend on the prior ranges chosen for w, which is the lesson of Section 5.17.2.

Does the answer you need depend on the model? Averaging is the formal remedy for model uncertainty. The informal remedy is a **sensitivity analysis**: fit several reasonable models and check whether the quantity you actually care about changes [Lambert §10.8]. Lambert's example is a rainfall model for sizing a dam. A two-state model (dry, wet) and a three-state model (dry, wet, persistently wet) fit 100 days of rain equally well on every standard test. But let each generate long rain series, and the three-state model gives far more probability to long runs of wet days. Long wet runs are the very quantity the dam depends on. So fit quality cannot choose between the models, and the conclusion that matters differs between them. The cautious report gives the answer of the pessimistic model and says why. The same habit applies to a forecast of a background tail under different fit functions, or of a cosmological parameter under different sets of nuisance parameters. A sensitivity analysis matters most when data are few or the decision is sensitive to the modelling choice.

5.18.3 How many bumps?

Sivia's "how many lines are there?" is the multi-model version of the bump hunt [Sivia §4.2]. A spectrum is a sum of M peaks of known shape on a background, M is unknown, and each value of M is a model with $2M$ peak parameters (amplitudes and positions). With a uniform prior on M over a few values, $\mathrm{P}(M \mid \boldsymbol{d})$ is proportional to the evidence Z_M [Sivia (4.15)], a $2M$-dimensional integral over a uniform box of amplitudes and positions [Sivia (4.16)–(4.19)].

We do it for $M = 0, 1, 2$ bumps on our line, with the line integrated out exactly as in (5.51) and the priors of Section 5.16.3 for each bump. For two bumps, $\chi^2_{\min}$ minus the line's $\chi^2_{\min}$ is a quadratic form in (A_1, A_2) with coefficients $\boldsymbol{r} \cdot \boldsymbol{u}_{\mu_j}$ and $\boldsymbol{u}_{\mu_j} \cdot \boldsymbol{u}_{\mu_k}$, so the four-dimensional grid (81 amplitudes and 161 positions per bump) costs only multiplications; refining it changes $\ln Z$ by 7.188×10^{-7}. The box includes both orderings of the two positions, so the $M!$ equivalent peaks of Sivia's (4.20) are counted automatically. Besides the one-bump data we simulate a data set with two bumps of height 2.5 at $\mu = 4$ and 5.4 (Figure 5.51).

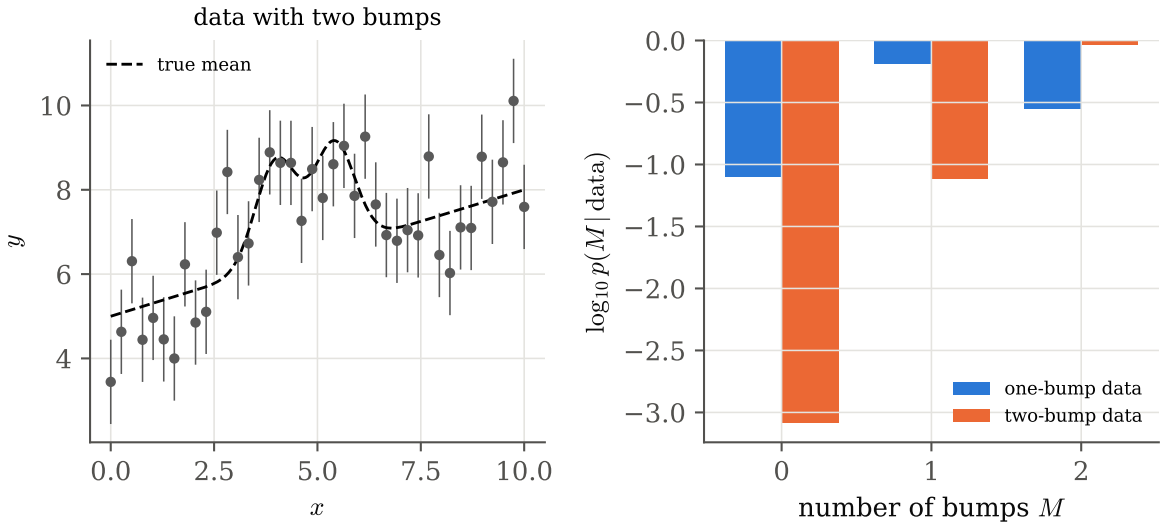


Figure 5.51: How many bumps? Left: forty points containing two overlapping bumps, with the true mean curve. Right: the posterior probability of each number of bumps, in powers of ten, for the one-bump data of Figure 5.45 and for the two-bump data.

	one-bump data					two-bump data				
M	$\chi^2_{\min}$	AIC	BIC	$\ln Z_M$	$P(M \mid \boldsymbol{d})$	$\chi^2_{\min}$	AIC	BIC	$\ln Z_M$	$P(M \mid \boldsymbol{d})$
0	49.07	53.07	56.44	−68.61	0.07901	63.32	67.32	70.7	−75.73	8.192×10^{-4}
1	34.54	42.54	49.29	−66.51	0.6425	44.62	52.62	59.38	−71.2	0.07661
2	30.1	42.1	52.24	−67.35	0.2784	30.98	42.98	53.11	−68.71	0.9226

For the one-bump data the evidence gives one bump the probability $P(M = 1 \mid \boldsymbol{d}) = 0.6425$, although the two-bump model fits better ($\chi^2_{\min} = 30.1$ against 34.54). For the two-bump data it gives two bumps 0.9226. The χ^2 column alone would always choose the most bumps. Two data sets could be lucky, so we repeat the count on 100 fresh data sets of each kind. The evidence picks one bump in 80 of the one-bump sets and two bumps in 80 of the two-bump sets.

5.18.4 Information criteria

The table also shows two quantities that are popular because they need no integral. Both are built from $-2 \ln \mathcal{L}_{\max}$, which for Gaussian noise is $\chi^2_{\min}$ plus a constant, plus a penalty for the number of parameters k [Lambert §10.5], [32, §4.7], [HobsonBMC (4.14)]:

$$\text{AIC} = -2 \ln \mathcal{L}_{\max} + 2k, \qquad \text{BIC} = -2 \ln \mathcal{L}_{\max} + k \ln N,$$

with N the number of data points; smaller is better. (The table uses $\chi^2_{\min}$ for $-2 \ln \mathcal{L}_{\max}$ and counts the line's two parameters, $k = 2, 4, 6$.)

The **Bayesian information criterion** is a crude evidence. Start from the Laplace approximation (5.50) with a general prior density [32, eqs. (39)–(41)]:

$$\begin{aligned} \ln Z &\overset{(1)}{\approx} \ln \mathcal{L}_{\max} + \ln \pi(\hat{\boldsymbol{\theta}}) + \frac{k}{2} \ln 2\pi - \frac{1}{2} \ln \det \mathbf{F} \\ &\overset{(2)}{\approx} \ln \mathcal{L}_{\max} + \ln \pi(\hat{\boldsymbol{\theta}}) + \frac{k}{2} \ln 2\pi - \frac{k}{2} \ln N - \frac{1}{2} \ln \det \mathbf{F}_1 \\ &\overset{(3)}{\approx} \ln \mathcal{L}_{\max} - \frac{k}{2} \ln N = -\tfrac{1}{2}\,\text{BIC}. \end{aligned} \tag{5.61}$$

(1) The logarithm of (5.50), with k parameters.
(2) For N independent data points the curvature of the log-likelihood grows in proportion to N: $\mathbf{F} = N\mathbf{F}_1$, with $\mathbf{F}_1$ the curvature per data point, and $\det(N\mathbf{F}_1) = N^k \det \mathbf{F}_1$.
(3) Keep only the terms that grow with N: $\ln \mathcal{L}_{\max}$ grows like N, $\ln N$ grows slowly, and the prior, 2π and $\mathbf{F}_1$ terms stay constant. They are dropped.

So $-\frac{1}{2}\Delta\text{BIC}$ approximates $\ln B$ for large N. The dropped constant terms are exactly the ones that carry the prior. That is why BIC is attractive (no prior to choose) and also why it is limited. It charges every parameter the same price $\frac{1}{2}\ln N$, as if every prior range were one data point's worth of information wide, whether the parameter is constrained or not. In our table $-\frac{1}{2}\Delta\text{BIC}$ gives 3.576 for one bump against none on the one-bump data, where the exact $\ln B$ is 2.096. Our prior ranges for the bump are wider than BIC assumes, so BIC under-charges.

The **Akaike information criterion** comes from a different question: which model will predict new data best, measured by a Kullback–Leibler divergence [Lambert §10.5.5]. Its penalty, 2 per parameter, is weaker still. In the 100 fresh one-bump data sets, where the truth has one bump, AIC chooses two bumps over one in 33, BIC in 8, and the evidence in 6.

Lambert describes the Bayesian descendants of AIC: DIC, which uses the posterior mean and the posterior variance of the log-likelihood as the penalty; WAIC, which averages the likelihood of each data point over the posterior; and leave-one-out cross-validation (LOO-CV), which predicts each point from a fit to all the others [Lambert §§10.5.6–10.5.8]. Trotta relates DIC to the Bayesian complexity, the effective number of parameters the data actually constrain [32, §4.6], [HobsonBMC (4.9)–(4.11)]. All of these answer "which model predicts best?", not "which model produced the data?". Lambert prefers them to Bayes factors because of the prior sensitivity of Section 5.17.2 [Lambert §10.6]; cosmologists mostly compute evidences, because their models come with physically motivated prior ranges.

Taxonomy: Which number answers which question

tool	what it computes	when to trust it
evidence Z, Bayes factor	probability of the data under each model, prior included	always the target; needs the prior and an integral
Laplace	$\mathcal{L}_{\max}\times$ posterior volume / prior volume	one Gaussian-like peak well inside the prior
Savage–Dickey	posterior over prior density of the new parameter at its null value	nested models, factorised prior, enough samples near the null
BIC	$\ln Z$ keeping only terms that grow with N	many independent data, all parameters constrained, priors about one datum wide
AIC, DIC, WAIC, LOO-CV	expected accuracy of predictions for new data	when prediction, not the generating model, is the question
$\Delta\chi^2$, p-value	improvement of the best fit; surprise under the null	never on its own for comparing models

5.18.5 Estimating a difference, or comparing models?

A trigger is upgraded. Before the upgrade $x_1 = 28$ of $n_1 = 40$ test pulses fired it; after, $x_2 = 35$ of $n_2 = 40$. Did the efficiency change? There are two different questions hidden in these words.

Estimation: by how much did it change? Give the two efficiencies p_1, p_2 independent flat priors. Their posteriors are independent, $p_j \mid x_j \sim \text{Beta}(x_j + 1, n_j - x_j + 1)$, and the posterior of the change $\tau = p_2 - p_1$ is obtained by drawing pairs and subtracting, exactly as in Wasserman's example of two binomials [Wasserman pp. 183–184]. With 200000 draws (Figure 5.52), $\text{P}(\tau > 0 \mid \boldsymbol{d}) = 0.9694$, the posterior mean of τ is 0.1666, and the central 95% interval is $[-0.007736, 0.3402]$.

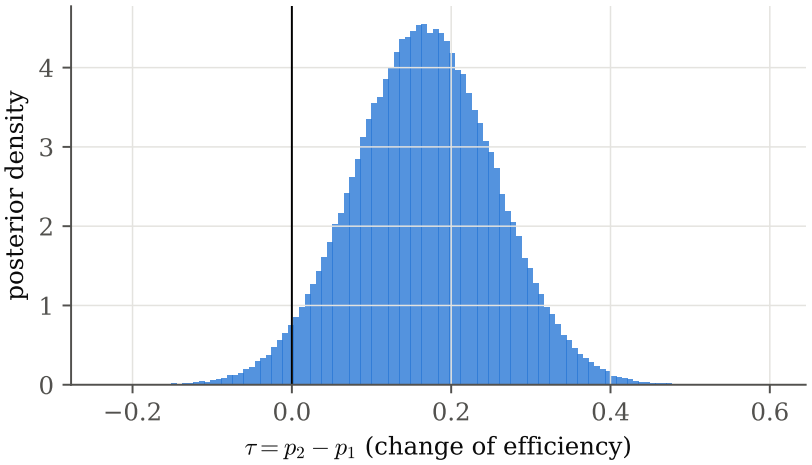


Figure 5.52: The posterior of the change of trigger efficiency, from 2×10^5 pairs of draws of the two Beta posteriors. About 97% of the area lies to the right of zero. The picture answers "by how much did it change?" and says nothing about whether the change is exactly zero, which has zero probability under this continuous posterior.

Comparison: did it change at all? Now let the hypotheses be two models. M_{same}: one efficiency p for both periods, flat prior. M_{diff}: independent p_1, p_2, flat priors. The evidence of each is a Beta integral like (5.46); the binomial coefficients are the same in both and cancel in the ratio:

$$B_{\text{same,diff}} \overset{(1)}{=} \frac{\int_0^1 p^{x_1+x_2}(1-p)^{n_1+n_2-x_1-x_2}\,\mathrm{d}p}{\int_0^1 p_1^{x_1}(1-p_1)^{n_1-x_1}\mathrm{d}p_1 \int_0^1 p_2^{x_2}(1-p_2)^{n_2-x_2}\mathrm{d}p_2} \overset{(2)}{=} \frac{B(x_1+x_2+1,\, n_1+n_2-x_1-x_2+1)}{B(x_1+1, n_1-x_1+1)\,B(x_2+1, n_2-x_2+1)}. \tag{5.62}$$

(1) The evidences (5.43) with flat priors (density one on $[0,1]$ and on $[0,1]^2$); the likelihood of the two samples is a product of two binomials.

(2) Each integral is a Beta function, as in (5.46).

Numerically $B_{\text{same,diff}} = 0.7517$, so with even prior odds $\text{P}(M_{\text{same}} \mid \boldsymbol{d}) = 0.4291$.

The two answers look contradictory: probability 0.9694 that the efficiency went up, and probability 0.4291 that it did not change. They are not, because they answer different questions with different priors.

- The estimation put zero prior probability on "exactly no change", since a continuous prior gives any single value probability zero.
- The comparison put half of all prior probability on exactly that value, and spread the other half over all pairs (p_1, p_2), most of which the data exclude.

Which is right depends on the physics. If the upgrade could plausibly have done nothing at all (a firmware change that might not have been loaded), the point hypothesis deserves its prior mass and the comparison is the right question. If the upgrade certainly changed something, and we want to know how much, estimation is the right question. Kruschke makes the same point about restricted models with negligible prior plausibility: testing every possible equality among nine baseball positions gives 21147 restricted models, most of which nobody believes, and estimation of the full model is more informative [Kruschke §10.5.1]. Lambert's advice is similar: when one model extends another, estimate within the bigger model and check its predictions [Lambert §10.7].

Watch out: A point null is a strong claim

Comparing "the effect is exactly zero" with "the effect is something" gives the point null a finite share of prior probability on a single value. That is appropriate when theory singles the value out ($n_s = 1$, $\Omega_k = 0$, $w = -1$, a symmetry that forbids a coupling). It is not appropriate for effects that are surely nonzero but possibly small; there the question is the size of the effect, and a credible interval answers it.

5.19 Why we need sampling

The denominator of Bayes' theorem, the integral over all parameter values, is the part of Bayesian inference that does not scale to many parameters. Every evidence so far was computed exactly or on a small grid. The coin factories had a Beta-function formula. The bump model had four parameters, but two of them, the line, entered linearly and could be integrated by hand, which left a two-dimensional grid. Real problems are not so kind. The base cosmological model has six parameters. Each evaluation of its CMB likelihood runs a Boltzmann code and takes of order a second. Its extensions add more parameters, none of them linear [HobsonBMC p. 87]. There are two ways out. For parameters, never compute the denominator. For models, compute it with an algorithm that spends its effort where the likelihood is.

5.19.1 The cost of a grid

A grid needs about w/Δ points along each axis to cover a width w with resolution Δ, so $(w/\Delta)^D$ points in D dimensions [HobsonBMC p. 58], [Sivia §3.4], [Kruschke §5.4]. The exponent is the whole problem. We can watch it happen by computing the bump evidence of Section 5.16.3 the brute-force way, on a grid over all four parameters (a, b, A, μ) with n points per axis and no analytic help (Figure 5.53). The wall time grows as n^4. With $n = 12$ the answer is $\ln B_{10} = 4.558$, wrong by more than a unit because the grid is too coarse to resolve the peak; with $n = 24$ it is 2.379; with $n = 64$ it is 2.1, against the exact 2.096, after 9.43 seconds.

At the measured cost of 5.621×10^{-7} seconds per grid point, a six-parameter problem with a thousand points per axis, 10^{18} points, would take $1.784e+04$ years. With a CMB likelihood at one second per call it would take 3.175×10^{10} years, more than twice the age of the Universe. And almost all of that work is wasted, because almost every grid point sits where the likelihood is negligible. How small is the region that matters? The bulk of the posterior occupies a fraction of about e^{-H} of the prior volume. Here H is the information gained from the data (the

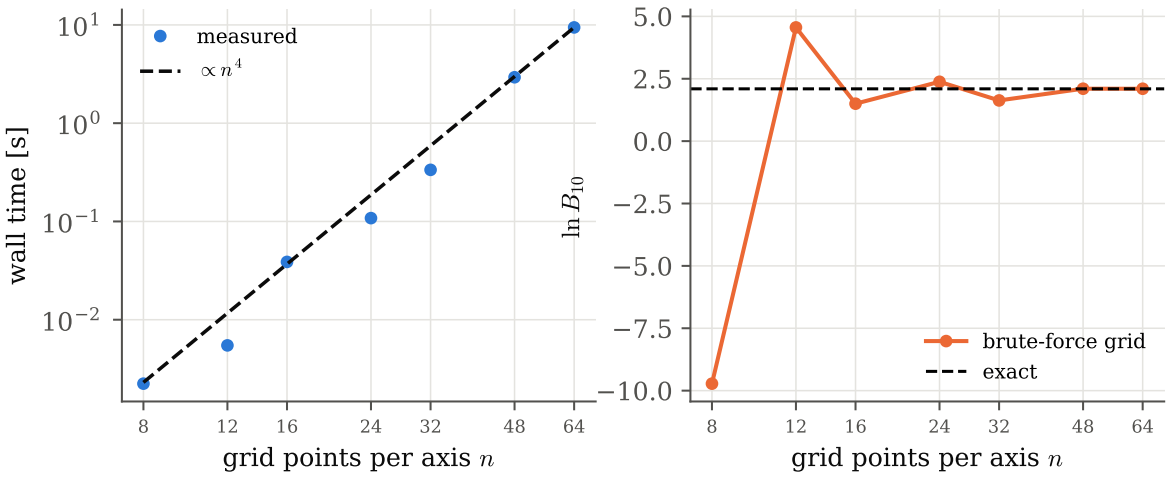


Figure 5.53: The bump evidence by brute force over the four-dimensional prior box. Left: the wall time grows as the fourth power of the number of points per axis (dashed line). Right: the log Bayes factor from the grid settles on the exact value only once the grid is fine enough to resolve the likelihood peak, which occupies a tiny corner of the box.

Kullback–Leibler divergence from prior to posterior), and it grows in proportion to the number of well-measured parameters [HobsonBMC p. 20].

5.19.2 Sampling the prior does not rescue us

A grid is one way to do an integral; random points are another. The evidence is an average of the likelihood over the prior, so draw $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K$ from the prior and average: $\hat{Z} = \frac{1}{K}\sum_k \mathcal{L}(\boldsymbol{\theta}_k)$ [Kruschke (10.7)]. The estimate is unbiased, and its error falls like $1/\sqrt{K}$ whatever the dimension. The catch is the constant in front of $1/\sqrt{K}$. For the simplest test case, a Gaussian likelihood of width s centred in the unit cube $[0,1]^D$ with a uniform prior, the relative variance of one term is

$$\frac{\mathrm{Var}_\pi[\mathcal{L}]}{Z^2} \overset{(1)}{=} \frac{\mathbb{E}_\pi[\mathcal{L}^2]}{Z^2} - 1 \overset{(2)}{=} \frac{\int e^{-|\boldsymbol{\theta}-\boldsymbol{\theta}_0|^2/s^2}\,\mathrm{d}^D\boldsymbol{\theta}}{\left(\int e^{-|\boldsymbol{\theta}-\boldsymbol{\theta}_0|^2/2s^2}\,\mathrm{d}^D\boldsymbol{\theta}\right)^2} - 1 \overset{(3)}{=} \frac{(s\sqrt{\pi})^D}{(s\sqrt{2\pi})^{2D}} - 1 = \left(\frac{1}{2\sqrt{\pi}\,s}\right)^D - 1. \quad (5.63)$$

(1) The variance is the mean square minus the square of the mean, and the mean of $\mathcal{L}$ over the prior is Z.

(2) The prior density is one on the cube, so averages over the prior are integrals; squaring the Gaussian halves its width parameter. We neglect the parts of the Gaussians outside the cube, which is exact enough for $s \ll 1$.

(3) Gaussian integrals: $\int e^{-u^2/s^2}\mathrm{d}u = s\sqrt{\pi}$ and $\int e^{-u^2/2s^2}\mathrm{d}u = s\sqrt{2\pi}$, once per dimension.

The relative error of $\hat{Z}$ is the square root of (5.63) divided by $\sqrt{K}$. For $s = 0.05$ the base is $1/(2\sqrt{\pi}\,s) \approx 5.6$: every extra dimension multiplies the required number of draws by more than five. With $K = 100000$ draws the simulation (Figure 5.54) finds a relative error of 0.01888 for $D = 2$ and 0.2925 for $D = 5$, roughly as (5.63) predicts. For $D = 8$ the estimate is so lopsided that our 20 repeated runs cannot pin down its error: most runs fall short, and the typical run gives only 0.1252 of the true value. By $D = 10$ the typical estimate is 8.293×10^{-4} of the true value: almost no draw lands where the likelihood is, and the rare draw that does dominates the sum. The fraction of prior draws inside the bulk of the posterior tells the same story: 0.04682 at $D = 2$, 7.315×10^{-4} at $D = 5$, 9.500×10^{-6} at $D = 8$. Kruschke's verdict on prior sampling is that it may need "a ginormous number of samples" [Kruschke p. 275].

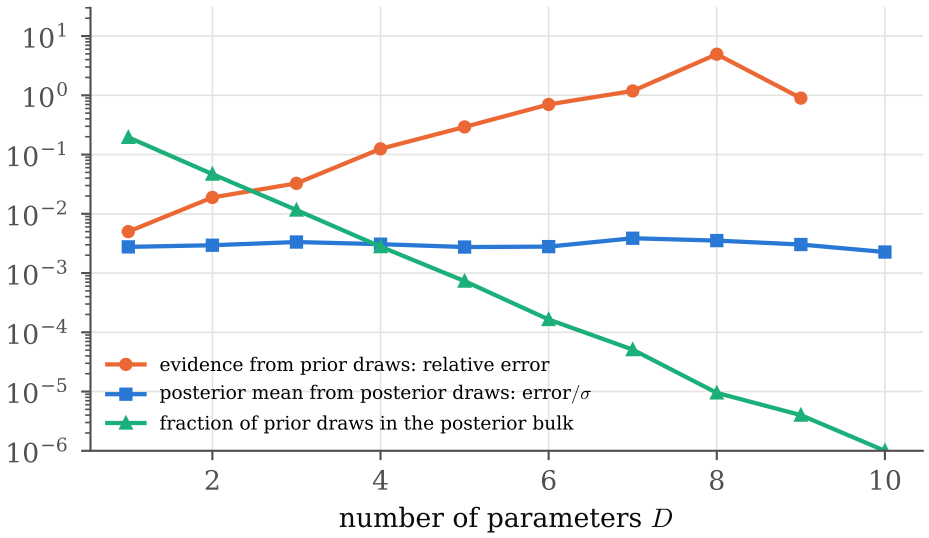


Figure 5.54: The curse of dimension for a Gaussian likelihood of width 0.05 in the unit cube, with 10^5 draws each time. Orange: relative error of the evidence estimated from draws of the prior, shown only while some runs still reach the likelihood peak. Green: fraction of prior draws that land in the bulk of the posterior. Blue: error of the posterior mean of one parameter estimated from draws of the posterior, in units of the posterior width; it does not grow with the dimension.

The blue curve of Figure 5.54 is the hopeful one. If instead we had draws from the *posterior*, the posterior mean of a parameter would have error $\sigma/\sqrt{K}$, independent of D: $0.002947\,\sigma$ at $D = 2$ and $0.002263\,\sigma$ at $D = 10$, both of the order of $1/\sqrt{10^5} \approx 0.003$. This is the central limit theorem at work on the samples [HobsonBMC §3.1], and it is why cosmologists sample: a few thousand posterior draws summarise a posterior in any number of dimensions [HobsonBMC p. 58]. What we need is a way to draw from the posterior without knowing its normalisation.

5.19.3 Ratios are all we need for parameters

The normalisation of the posterior is the evidence, and it cancels from every ratio of posterior densities:

$$\frac{p(\boldsymbol{\theta}' \mid \boldsymbol{d})}{p(\boldsymbol{\theta} \mid \boldsymbol{d})} \overset{(1)}{=} \frac{p(\boldsymbol{d} \mid \boldsymbol{\theta}')\,p(\boldsymbol{\theta}')/p(\boldsymbol{d})}{p(\boldsymbol{d} \mid \boldsymbol{\theta})\,p(\boldsymbol{\theta})/p(\boldsymbol{d})} \overset{(2)}{=} \frac{\mathcal{L}(\boldsymbol{\theta}')\,\pi(\boldsymbol{\theta}')}{\mathcal{L}(\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})}. \tag{5.64}$$

(1) Bayes' theorem at the two points $\boldsymbol{\theta}'$ and $\boldsymbol{\theta}$.
(2) The evidence $p(\boldsymbol{d})$ is the same number at both points and cancels.

A method that only ever asks "how much more probable is this point than that one?" can therefore move through parameter space using likelihood times prior alone. The Metropolis algorithm does exactly this [36]. From the current point $\boldsymbol{\theta}$, propose a random step to $\boldsymbol{\theta}'$; accept it with probability $\min(1, \text{ratio (5.64)})$; otherwise stay where you are and count the current point again. We don't want the unbiased random walker, but instead a biased random walker which spends more time in the region where probability is higher: the accept–reject step, which uses only the ratio, is the bias. That the fraction of time the walker spends in each region converges to the posterior probability of that region is proved in Chapter 9; here we only watch it work.

Listing 5.11: A Metropolis walk on the bump posterior that uses only ratios (`code/ch05/28_sampling_evidence.py`)

```
37 def metropolis(y, rng, n_steps=1_000_000, step=(0.25, 0.04, 0.35, 0.12)):
38     ab = np.linalg.lstsq(lb.DESIGN, y, rcond=None)[0]
39     theta = np.array([ab[0], ab[1], 1.0, 5.0])                    # start: best line, small
       bump mid-range
40     lp = lb.ln_post_unnorm(theta, y)
41     chain = np.empty((n_steps, 4)); acc = 0
42     for k in range(n_steps):
43         prop = theta + np.array(step) * rng.standard_normal(4)
44         lp_new = lb.ln_post_unnorm(prop, y)
45         if np.log(rng.random()) < lp_new - lp:                     # only the RATIO of
       posteriors enters
46             theta, lp = prop, lp_new; acc += 1
47         chain[k] = theta
48     return chain, acc / n_steps
```

Line 39 starts the walker at the best line with a small bump in the middle of the box. Line 43 proposes a Gaussian step in all four parameters. Line 44 evaluates the logarithm of likelihood times prior, which is $-\infty$ outside the prior box, so steps out of the box are always rejected. Line 45 is the whole algorithm: accept if a uniform random number is below the ratio, in logarithms. Nothing in the loop knows the evidence.

Example: the bump posterior from a chain. For the data with a bump, 1000000 steps take 43.73 seconds, with acceptance rate 0.3533; we discard the first 50000 as the walker finds the peak. The histograms of the chain (Figure 5.55) lie on the grid posteriors. The posterior mean position is 6.37 from the chain, 6.371 from the standard ensemble sampler `emcee` run as a cross-check (640000 samples), and 6.349 on the grid; the mean amplitude is 2.215 (chain), 2.212 (`emcee`) and 2.199 (grid).

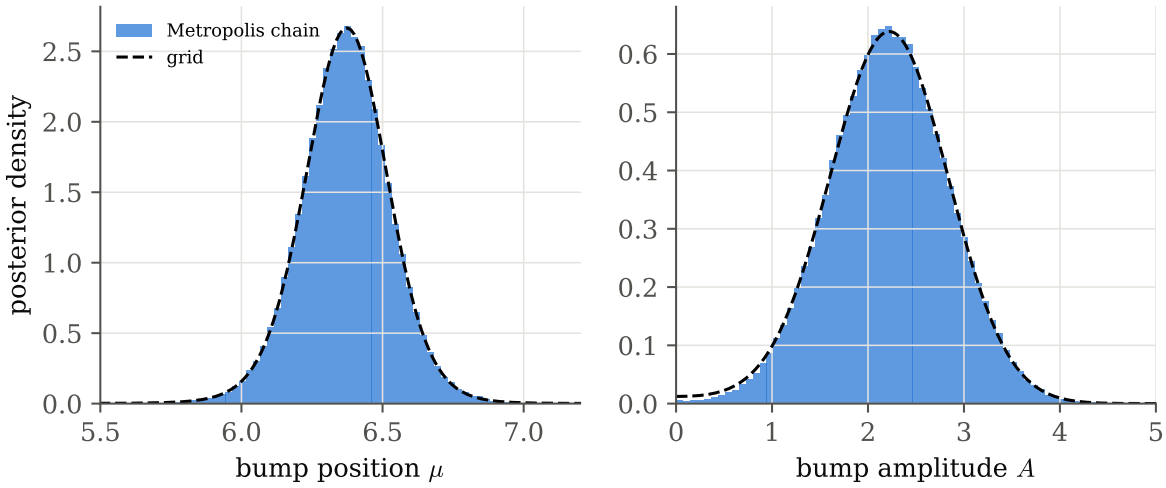


Figure 5.55: The posterior of the bump position and amplitude from a Metropolis chain of 10^6 steps that only ever evaluated ratios of likelihood times prior (histograms), against the grid posterior (dashed). The chain never computed the evidence, yet reproduces the normalised shape.

A chain can miss part of a posterior, and the reason matters. Besides the peak, a bump posterior can have a faint "pedestal": bumps of small amplitude anywhere in $0 < \mu < 10$, which fit about as well as the line. How much probability the pedestal holds depends on how clearly the data show the bump. For our data the grid gives it 0.007269 (for $A < 0.5$). The chain spends 0.003826 of its time there and `emcee` 0.005811; 6 chains with different seeds give between 0.003826 and 0.007526. When the pedestal holds a noticeable share, a chain reaches it only in rare excursions, so its share is mostly luck, set by how many times it happened to wander off

the peak. Because the pedestal is spread over the whole range of μ, those rare visits matter most for the posterior standard deviation of μ: 0.4253 on the grid against 0.1609 to 0.2646 from the chains. Why are the visits rare? A walker that takes small local steps crosses from the peak to a distant low-amplitude region only rarely, so a finite chain can under-represent that region. How long a chain must run, and how to tell whether it has, is the subject of Chapter 9.

How this is used in real research: Verification here, the instrument elsewhere

For the bump posterior the chain can be checked against a grid, so the chain only *verifies* what we already know. In a cosmological analysis there is no grid: the posterior of ΛCDM and its extensions from Planck, BAO and supernova data exists only as a chain from codes such as CosmoMC [37], Cobaya or MontePython, and the chain *is* the result. Convergence is then judged from the chains themselves, for instance by comparing several chains started far apart [38], and the samples are turned into marginal densities and triangle plots with `getdist`. The two-step habit of this book carries over: write the sampler from scratch once to see the mechanism, then use and cross-check the research code.

5.19.4 The evidence needs more than a chain

A chain never computes the evidence, so it cannot simply report it. Wasserman gives a sharp form of the difficulty [Wasserman Example 11.10]. Suppose a density is known up to its constant, $f(x) = c\,g(x)$ with g known, and we can draw n samples $x_1, \dots, x_n$ from f. Can the samples tell us c? Treat c as a parameter. The likelihood of the samples is $\prod_i c\,g(x_i) \propto c^n$, the same function of c whatever the samples are, so a Bayesian analysis learns nothing about c from them. Yet c can be estimated another way. Estimate the density $\hat{f}$ from the samples at any point x; then $\hat{c} = \hat{f}(x)/g(x)$.

The same trick works for the evidence. Bayes' theorem, rearranged, gives an identity for the evidence at *any* parameter value $\boldsymbol{\theta}$:

$$p(\boldsymbol{\theta} \mid \boldsymbol{d}) = \frac{\mathcal{L}(\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})}{Z} \quad \Longrightarrow \quad Z = \frac{\mathcal{L}(\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})}{p(\boldsymbol{\theta} \mid \boldsymbol{d})}. \tag{5.65}$$

The numerator, likelihood times prior, we can compute. The denominator is the normalised posterior density, which the samples estimate. Here $\mathcal{L}(\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ plays the role of the known g, and Z the role of $1/c$. The Savage–Dickey ratio of Section 5.18.1 is this idea for nested models at the point $\psi = \psi_*$. The `MCEvidence` code estimates the posterior density from the nearest-neighbour distances of chain samples [Padilla §7]. Kruschke describes a version that averages over the chain instead of using one point [Kruschke (10.8)]: for any normalised density $h(\boldsymbol{\theta})$, $1/Z = \mathbb{E}_{\text{post}}[h(\boldsymbol{\theta})/(\mathcal{L}(\boldsymbol{\theta})\pi(\boldsymbol{\theta}))]$, which works well when h resembles the posterior and badly when it does not (with h equal to the prior it becomes the harmonic-mean estimator, whose variance can be infinite). All these need a good density estimate in the region that matters, which is again hard in many dimensions. The reliable general-purpose methods are built for the evidence from the start.

A chain can also be made to carry the model index. Give the model label m a prior and let the sampler move in the joint space of m and the parameters of each model, so that the fraction of steps spent at m estimates $\mathrm{P}(m \mid \boldsymbol{d})$, and the ratio of the fractions estimates the posterior odds [Kruschke §10.3.2]. The difficulty is that, when the chain sits at $m = 1$, the parameters that belong only to model 2 have no likelihood to pull them anywhere, so they wander over their whole prior. When the chain proposes a jump to $m = 2$, those parameters are almost certainly somewhere the data exclude, and the jump is rejected. The chain rarely changes model. The remedy is a **pseudo-prior**: while the chain sits in another model, the idle parameters are drawn

from a stand-in density placed where their own model's posterior lives, so that a jump to that model lands somewhere sensible. A pseudo-prior costs a preliminary run in each model, and, because it does not enter the final answer, it only affects efficiency, not the result.

5.19.5 Nested sampling

Skilling's nested sampling turns the D-dimensional evidence integral into a one-dimensional one [39], [HobsonBMC §1.4.1]. How can that be? The evidence only asks how much prior mass has how much likelihood; it does not care where in parameter space that mass sits. So sort all the prior mass by likelihood. For each likelihood level ℓ let

$$X(\ell) = \int_{\mathcal{L}(\boldsymbol{\theta})>\ell} \pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$$

be the prior mass enclosed by the contour $\mathcal{L} = \ell$ [HobsonBMC (1.57)]. It falls from $X = 1$ at $\ell = 0$ to $X = 0$ at the maximum likelihood, and its inverse $\mathcal{L}(X)$ is the likelihood of the contour that encloses prior mass X (Figure 5.56). Then

$$\begin{aligned} Z &\overset{(1)}{=} \int \pi(\boldsymbol{\theta}) \Big[\int_0^\infty \mathbb{1}\big(\mathcal{L}(\boldsymbol{\theta}) > \ell\big)\,\mathrm{d}\ell \Big] \mathrm{d}\boldsymbol{\theta} \overset{(2)}{=} \int_0^\infty \Big[\int_{\mathcal{L}(\boldsymbol{\theta})>\ell} \pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \Big] \mathrm{d}\ell \\ &\overset{(3)}{=} \int_0^\infty X(\ell)\,\mathrm{d}\ell \overset{(4)}{=} \int_0^1 \mathcal{L}(X)\,\mathrm{d}X. \end{aligned} \tag{5.66}$$

(1) Write the positive number $\mathcal{L}(\boldsymbol{\theta})$ as the length of the interval $0 < \ell < \mathcal{L}(\boldsymbol{\theta})$, an integral of the indicator $\mathbb{1}(\cdot)$, which is one when its condition holds and zero otherwise.

(2) Exchange the order of integration: for fixed ℓ the indicator restricts $\boldsymbol{\theta}$ to the region inside the contour $\mathcal{L} = \ell$.

(3) The inner integral is the enclosed prior mass $X(\ell)$.

(4) The area under the decreasing curve $X(\ell)$, plotted against ℓ, is the same region of the plane as the area under its inverse $\mathcal{L}(X)$, plotted against X [HobsonBMC (1.58)].

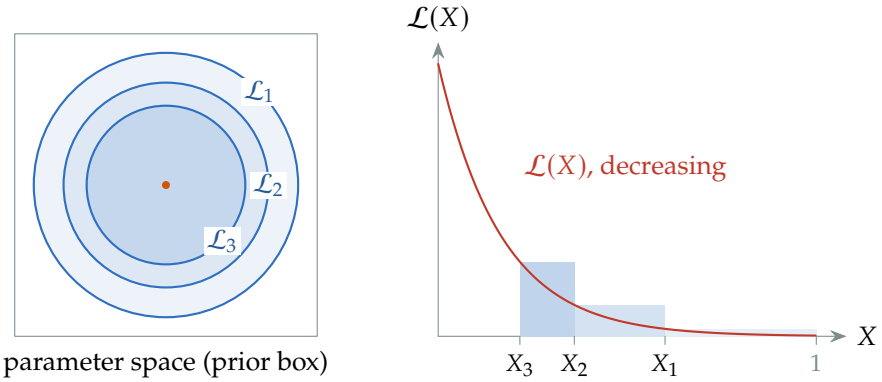


Figure 5.56: Nested sampling in pictures. Left: nested likelihood contours in parameter space; the contour $\mathcal{L}_i$ encloses prior mass X_i, and each step moves inward. Right: the same contours as a decreasing curve $\mathcal{L}(X)$ over the unit interval of prior mass. The evidence is the area under the curve, accumulated strip by strip (shaded), each strip of height $\mathcal{L}_i$ and width $X_{i-1} - X_i$. Here each step shrinks X by the factor 0.6.

A one-dimensional integral of a decreasing function is easy, if we know $\mathcal{L}$ at a sequence of known X values. Nested sampling produces such a sequence with the X values known statistically. Keep N "live" points drawn from the prior. At each step:

1. find the live point with the lowest likelihood, $\mathcal{L}_i$; record it and its estimated enclosed mass X_i;
2. replace it by a new point drawn from the prior subject to $\mathcal{L} > \mathcal{L}_i$ [HobsonBMC (1.61)];
3. add the strip $\mathcal{L}_i\,(X_{i-1} - X_i)$ to the running sum; stop when the live points could not add a significant amount any more.

Why do we know X_i? Before step i the N live points are independent draws from the prior inside the contour $\mathcal{L}_{i-1}$, so their enclosed masses are uniform on $[0, X_{i-1}]$. The discarded point is the one with the largest enclosed mass (the lowest likelihood). The ratio $t = X_i/X_{i-1}$ is therefore the largest of N uniform numbers on $[0,1]$, whose density is Nt^{N-1} [HobsonBMC (1.62)]: the probability that all N lie below t is t^N, and its derivative is Nt^{N-1}. On average

$$\mathbb{E}[\ln t] \overset{(1)}{=} \int_0^1 Nt^{N-1} \ln t \, \mathrm{d}t \overset{(2)}{=} \Big[t^N \ln t\Big]_0^1 - \int_0^1 t^N \, \frac{\mathrm{d}t}{t} \overset{(3)}{=} -\frac{1}{N}, \tag{5.67}$$

(1) The expectation of $\ln t$ over the density Nt^{N-1}.
(2) Integration by parts, with $Nt^{N-1}\mathrm{d}t = \mathrm{d}(t^N)$.
(3) The boundary term vanishes at both ends ($t^N \ln t \to 0$ as $t \to 0$), and $\int_0^1 t^{N-1}\mathrm{d}t = 1/N$.

so $\ln X_i \approx -i/N$: each step shrinks the enclosed prior mass by about $e^{-1/N}$ [HobsonBMC (1.63)]. The prior mass shrinks geometrically, so the algorithm reaches the tiny region $X \sim e^{-H}$ where the posterior lives after about NH steps, a number linear rather than exponential in the information. The uncertainty of the X_i gives an uncertainty of $\ln Z$ of about $\sqrt{H/N}$; it can also be found by redrawing the shrinkage factors t many times from Nt^{N-1} and recomputing the sum [HobsonBMC p. 24].

Example: the bump evidence by nested sampling. We run nested sampling from scratch on the (A, μ) problem of Section 5.16.3, with the line integrated out, so that the result is the Bayes factor B_{10} itself. With $N = 100$ live points the run takes 1211 steps and ends at $\ln X = -12.11$. It returns $\ln B_{10} = 2.066 \pm 0.1805$, the error from redrawing the shrinkage factors, against the exact 2.096 (Figure 5.57). The information is $H = 3.982$, so Skilling's estimate of the error is $\sqrt{H/N} = 0.1996$, and 8 independent runs scatter with standard deviation 0.2064 about a mean of 2.083. The algorithm knows its own error.

Our version is honest but naive in one place. Step 2 draws new points from the whole prior until one beats the current lowest likelihood; late in the run almost every draw fails, and the run spends 90.45 million likelihood evaluations. Research codes draw the replacement point from a short constrained Markov chain or from ellipsoids fitted around the live points, and need of order 10^5 likelihood evaluations for a cosmological evidence [32, §4.3]. With a likelihood that takes a second, that is the difference between a day and several years. Thermodynamic integration, which heats a Markov chain from the prior to the posterior, is the older alternative and typically costs about a hundred times more [32, §4.3], [HobsonBMC §4.4.1]. Two further options are VEGAS, an adaptive multidimensional integrator that refines a sampling density on the fly and is popular in particle physics, and parallel tempering, in which Markov chains at different "temperatures" (likelihood raised to a power below one, so that the landscape is flatter) run side by side and swap states, which helps chains cross between separated peaks [HobsonBMC §4.4.1].

5.19.6 The same question for lists, parameters and models

Bayes' theorem has now been used on three kinds of hypotheses, and the questions never changed. For a short list (Monty's doors, a disease, two urns) the posterior was a set of numbers.

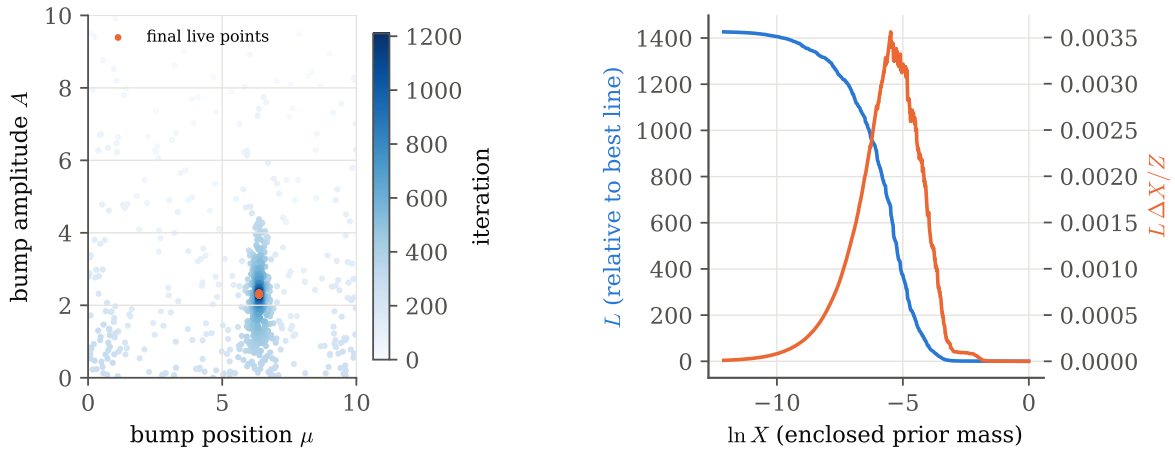


Figure 5.57: Nested sampling of the bump evidence. Left: the discarded points in the plane of bump position and amplitude, shaded by the step at which they were discarded; the early ones are spread over the whole prior box, the late ones close in on the peak, where the final live points sit. Right: the likelihood of the discarded points against the logarithm of the enclosed prior mass (blue), and each point's share of the evidence (orange). The evidence comes from a band around $\ln X \approx -5$, that is, from about e^{-5} of the prior.

For the values of a continuous parameter it was a density. For whole models it was again a set of numbers, and the probability of the data under each model was itself an average over that model's parameters. Each time we asked the two questions of Section 5.6: which hypotheses could have produced what we observed, and how probable the observation was under each. For model comparison the first question lists the models, each with its prior; the second is answered by the evidence.

The difference between the two uses lies in the evidence. For parameter estimation inside a model the evidence cancels, and a sampler that uses only ratios delivers the posterior. For model comparison it does not cancel, so it must be computed: exactly, by a Laplace or Savage–Dickey approximation, or by nested sampling. The Markov chains of Chapter 9 and the Fisher forecasts of Chapter 10 are the two tools the rest of the book uses to answer the second question in many dimensions.

Recap

- A model is a likelihood and a prior. Its evidence $Z = \int \mathcal{L}\,\pi\,\mathrm{d}\boldsymbol{\theta}$ is the probability it gave to the observed data: the likelihood averaged over its prior.
- $\mathrm{P}(M_i \mid \boldsymbol{d}) \propto Z_i\,\mathrm{P}(M_i)$; posterior odds = Bayes factor × prior odds. The comparison says nothing about parameters; model averaging combines the parameter posteriors with the model probabilities as weights.
- No model is rejected on its own: $p(\boldsymbol{d} \mid \text{not } M)$ is undefined without specific alternatives.
- Occam's razor is derived, not assumed: $Z \approx \mathcal{L}_{\max} \times V_{\text{post}}/V_{\text{prior}}$. Better fits are rewarded, wasted prior range is penalised, unconstrained parameters cost nothing. For the bump: a best-fit gain of $e^{7.265}$ became odds of $8.133 : 1$.
- Read Bayes factors on the logarithmic Jeffreys scale ($|\ln B| = 1, 2.5, 5$), as posterior odds, and with their numerical and prior uncertainty.
- The evidence depends on prior widths ($-\ln 10$ per decade of range) and can reverse between two vague priors that give the same posterior; improper priors are forbidden.

Equally informed priors, physically motivated ranges and sensitivity plots are the remedies.

- Lindley: at fixed significance, large samples favour the null. A p-value bounds the odds: at most $1/(e\,p\ln p)$ against the null, 2.5 : 1 at 2σ, 150 : 1 at 3.6σ.
- Nested models: Savage–Dickey $B_{01} = p(\psi_* \mid \boldsymbol{d})/\pi(\psi_*)$ from the bigger model alone; BIC is the Laplace evidence without its prior terms; AIC, WAIC and LOO rank predictions, a different question.
- Estimating a difference and testing whether it is exactly zero are different questions with different priors, and can give opposite-looking answers.
- Grids cost n^D and prior sampling fails exponentially in D; posterior samples give errors independent of D. Ratios of posteriors do not need the evidence (Metropolis, Chapter 9); the evidence needs its own algorithm, such as nested sampling, $Z = \int_0^1 \mathcal{L}(X)\,\mathrm{d}X$ with $\ln X_i \approx -i/N$.

Reading the literature

- **The main line.** Wasserman, *All of Statistics*, Chapter 11: the Bayesian method, conjugate examples, simulation from the posterior, flat and Jeffreys priors, Bayesian testing and its prior-free bound, and the strengths and weaknesses of Bayesian inference [Wasserman pp. 175–189].
- **The gentlest route.** Kruschke, *Doing Bayesian Data Analysis*, Chapters 2 and 5 (credibility, Bayes' rule), 6 (the Beta–Bernoulli model), 10 (model comparison, Occam's razor, prior sensitivity) and 12 (HDIs, estimation against testing).
- **The physicist's tutorial.** Sivia, *Data Analysis: A Bayesian Tutorial*, Chapters 1–3 (plausible reasoning, parameter estimation, the Gaussian approximation, marginalisation), 4 (Mr A and Mr B, how many lines) and 5 (assigning priors).
- **Priors, prediction and checks.** Lambert, *A Student's Guide to Bayesian Statistics*, Chapters 3–7 and 9 (Bayes' rule, likelihood, priors, the denominator, conjugacy) and 10 (predictive checks, AIC to LOO-CV, Bayes factors, sensitivity analysis).
- **Cosmology.** Hobson, Jaffe, Liddle, Mukherjee and Parkinson (eds.), *Bayesian Methods in Cosmology*, Chapter 1 (foundations, priors, nested sampling), Chapter 3 (Monte Carlo parameter estimation) and Chapter 4 (model selection, Savage–Dickey, the Jeffreys scale, the spectral index). Trotta's review "Bayes in the sky" [32] covers evidence, Occam's razor, the Lindley paradox and the p-value calibration in cosmological language. Padilla et al. [40] is a hands-on introduction with code.
- **Original papers.** Kass and Raftery on Bayes factors [33]; Lindley's paradox [34]; Sellke, Bayarri and Berger on calibrating p-values [35]; Skilling on nested sampling [39]; Metropolis et al. on the algorithm of Chapter 9 [36].
- **Pictures.** *Seeing Theory*, chapter "Bayesian Inference", for interactive versions of the updating examples.

Chapter summary

Bayesian inference turns the direction of probability around. Instead of asking what a hypothesis predicts, we start from an outcome already observed and ask which hypotheses might lie behind it, and how probable each is after the fact. Bayes' theorem answered this question first for a short list of hypotheses, then for a continuous parameter, and then for whole models, where the denominator we had been ignoring became the quantity of interest. In the pipeline of the book,

$$\text{model} \to \text{data} \to \text{summary } S \to P(S \mid \theta) \to \boxed{\text{infer physics}},$$

Bayesian inference is the last arrow: the sampling distribution $P(S \mid \theta)$ becomes the likelihood, and the posterior over parameters and models is the inference. Chapter 4 gave the frequentist form of the same arrow.

Bayes' theorem with a list of hypotheses

concept	operational meaning	used later in
prediction vs inference	$\mathrm{P}(E \mid H)$ reads the tree forward; $\mathrm{P}(H \mid E)$ asks which path produced the observed leaf	every chapter
joint table, logical not causal	one table of (hypothesis, outcome) probabilities: columns predict, rows infer; a posterior links hypothesis and evidence logically, it does not say the evidence caused anything	every chapter
Bayes' theorem	product rule in both orders: $\mathrm{P}(H \mid E) = \mathrm{P}(E \mid H)\,\mathrm{P}(H)/\mathrm{P}(E)$	Chapters 8 and 9
information shrinks the sample space	conditioning on E keeps only the region where E holds and re-measures everything inside it	Chapter 8
unit-square picture	priors are widths, likelihoods heights, posterior = surviving rectangle over total surviving area	Chapter 8
reasoning in words	the event; what could have produced it; its probability under each hypothesis; then priors and normalise. Monty Hall: "Monty opened C"; what made him open that door?; $\frac{1}{2}, 1, 0$ for car behind A, B, C; posterior $\frac{1}{3}, \frac{2}{3}, 0$	every chapter
base rates	a tall bar in a thin strip can lose to a short bar in a wide one (rare signals, trigger purity)	Chapters 8 and 12
reallocation of credibility	zero likelihood eliminates, zero prior can never be revived, noisy data only nudge	Chapter 9
odds, likelihood ratio	posterior odds = likelihood ratio × prior odds; independent evidence adds in logarithms	Chapters 8 and 10
sequential updating	yesterday's posterior is today's prior, if the data are independent given the hypothesis	Chapter 9
denominator, prior predictive	$\mathrm{P}(E) = \sum_i \mathrm{P}(E \mid H_i)\,\mathrm{P}(H_i)$: normaliser and forecast; needs every hypothesis	Chapters 6 and 9

Parameter estimation

concept	operational meaning	used later in
posterior density	likelihood × prior on a grid, normalised; the parameter value is the hypothesis, the data the event	Chapters 8 and 9
conjugate priors	Beta–Binomial, Gamma–Poisson, Normal–Normal: posterior in closed form, pseudo-counts and inverse-variance weights	Chapters 7 and 8
posterior summaries	mean, median, mode from three losses; equal-tailed interval; HDI, the shortest set	Chapters 9 and 10
credible vs confidence	probability of θ in a set given the data, against coverage over repetitions	Chapters 8 and 10
Gaussian (Laplace) approximation	peak and curvature: $\sigma^2 = (-\partial^2 \ln p)^{-1}$, covariance $= -H^{-1}$, $\Delta\chi^2$ regions	Chapter 10
marginalising, profiling	integrate nuisance parameters out (volume included) or maximise over them	Chapters 8 to 10
choosing a prior	knowledge, flat (in which variable?), Jeffreys $\propto \sqrt{I}$; improper priors as limits	Chapter 9
forgetting the prior	the likelihood dominates for many data (Bernstein–von Mises); not the same as a chain forgetting its start	Chapter 9
posterior sampling	a histogram of draws is the density; functions of parameters for free	Chapters 6, 9 and 12

Model comparison and the need for sampling

concept	operational meaning	used later in
evidence Z	likelihood averaged over the model's prior; the probability the model gave to the data	Chapters 9 and 12
Bayes factor, posterior odds	$B_{12} = Z_1/Z_2$ multiplies the prior odds; quote $\ln B$ and which model is on top	Chapters 9 and 12
Occam factor	$Z \approx \mathcal{L}_{\max} V_{\text{post}}/V_{\text{prior}}$; wasted prior range is penalised, unconstrained parameters are free	Chapter 10
Jeffreys scale	$\lvert\ln B\rvert = 1, 2.5, 5$: odds 3, 12, 150; apply to posterior odds, with the prior uncertainty	Chapter 12
prior sensitivity	$\ln B$ falls by $\ln 10$ per decade of prior range; vague priors can reverse it; improper priors forbidden	Chapter 12
Lindley, p-value bounds	fixed significance favours the null at large n; $B_{10} \leq -1/(e\, p \ln p)$	Chapters 8 and 12
Savage–Dickey	$B_{01} = p(\psi_* \mid \boldsymbol{d})/\pi(\psi_*)$ from the bigger model's samples	Chapter 9
model averaging	parameter posteriors weighted by model probabilities	Chapter 14
BIC, AIC, WAIC, LOO	crude evidence without prior terms; predictive accuracy, a different question	Chapter 12
estimation vs comparison	"how large is the effect?" against "is it exactly zero?": different priors, different answers	Chapters 8 and 12
curse of dimension	grids cost n^D; prior draws miss the posterior; posterior draws do not care about D	Chapters 6 and 9
ratios, Metropolis	the evidence cancels in posterior ratios; a biased random walk samples the posterior	Chapter 9
nested sampling	$Z = \int_0^1 \mathcal{L}(X)\mathrm{d}X$, prior mass shrinking by $e^{-1/N}$ per step, error $\sqrt{H/N}$	Chapters 9 and 12

Part II

Physics Toolkits

Chapter T1

Cosmology for the Data Analyst

Where this chapter is going

For a given set of six cosmological parameters, CAMB [8] predicts the angular power spectrum C_ℓ of the microwave background, the matter power spectrum $P(k)$ and the relation between distance and redshift. The two spectra describe fluctuations scale by scale, in temperature on the sky and in matter density in space. Our sky, with its particular hot and cold spots and its particular galaxies, is one realisation of the model. Inferring the parameters therefore requires us to understand both how the predicted spectra depend on those parameters and how far a single observed spectrum can depart from the prediction.

Cosmology supplies the first step in the analysis of Chapter 0 by fixing the expected values of the summaries we measure. The companion cosmology notes derive the underlying physics. Here we define each observable, relate the conventions used in those notes to CAMB's output, and follow how the prediction changes as we vary the six parameters (Figure T1.1).

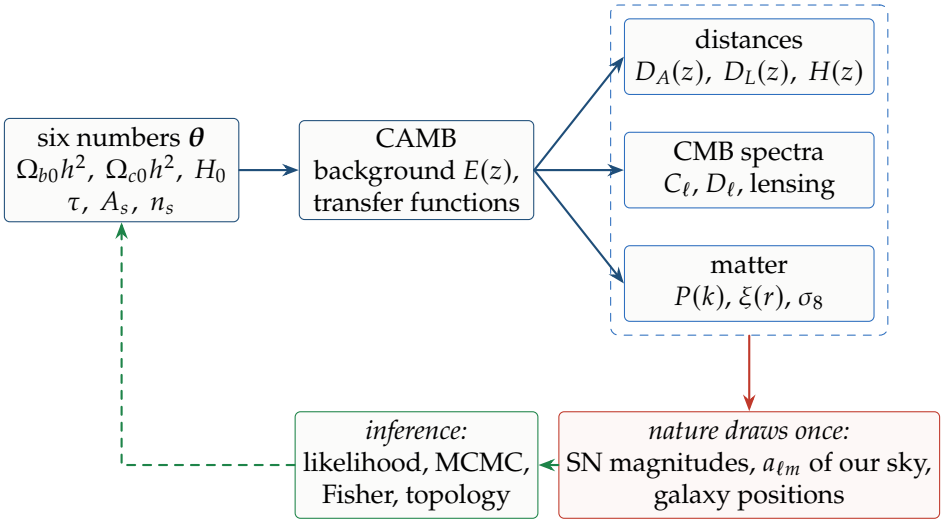


Figure T1.1: The machine and its one output we can see. Six numbers go into CAMB, which returns three families of predicted statistics (blue). Nature does not hand us those statistics: it hands us a single noisy draw from them (red), and the statistics of the later chapters (green) turn that draw back into constraints on the six numbers.

The link between the cosmology and the statistics is how the outputs move when the inputs move. Nudge one parameter and watch C_ℓ respond. That response is a derivative, and the Fisher matrix of Chapter 10 is built from such derivatives. If two parameters produce responses that look alike, the data cannot tell those parameters apart. This is a degeneracy, and the MCMC chains of Chapter 9 show it as a long ridge. On the matter side, the linear $P(k)$ is the

input of the Gaussian random fields of Chapter 7. On small scales gravity makes the real spectrum (computed by the halofit recipe) depart from the linear one. There the field is no longer Gaussian, and the topological summaries of Chapter 12 have something to find that the power spectrum misses.

The microwave light is also polarized, and a field of polarization sticks splits into a part that looks the same in a mirror and a part that does not. Ordinary physics leaves no correlation between the two, so a measured one is a test of whether the laws of physics, or our own detectors, prefer one hand (Sections T1.10 and T1.13).

T1.1 Six numbers in, statistics out

Think of the flat ΛCDM model as a function with six inputs. Two inputs are physical densities: the baryon density $\Omega_{b0}h^2$ and the cold dark matter density $\Omega_{c0}h^2$ [today's density of each component divided by the critical density, times h^2; the product is a physical density that does not depend on H_0]. One input fixes the distances: either the Hubble constant $H_0 = 100\,h$ km s^{-1} Mpc^{-1}, or equivalently the angle $\theta_{\rm MC}$ that the sound horizon at decoupling subtends on the sky. One is the optical depth to reionisation, τ [the expected number of Thomson scatterings a CMB photon undergoes after the Universe is re-ionised; a fraction $e^{-\tau}$ of the photons arrive unscattered]. The last two describe the primordial fluctuations: their amplitude A_s and their spectral tilt n_s. Everything else is held fixed: the radiation density (from the measured CMB temperature), flatness $\Omega_{K0} = 0$, and a single massive neutrino of 0.06 eV.

These numbers appear in three places, and Table T1.1 translates between them: the symbol used in the cosmology notes, the keyword that CAMB's Python interface takes, and the Planck 2018 value with its 68% error [25, Table 1]. The book's fiducial model (`code/camb_fiducial.py`) is this best fit. The $\pm1\sigma$ errors are the step sizes of the first experiment (Section T1.3).

symbol (cosmology notes)	one-line meaning	CAMB keyword	Planck 2018 (68%)
$\Omega_{b0}h^2$	physical baryon density; sets the baryon loading of the acoustic oscillations	`ombh2`	0.02237 ± 0.00015
$\Omega_{c0}h^2$	physical cold dark matter density; sets matter–radiation equality	`omch2`	0.1200 ± 0.0012
$100\,\theta_{\rm MC}$ *or* H_0	angular size of the sound horizon at decoupling, or the expansion rate today	`cosmomc_theta` or `H0`	1.04092 ± 0.00031; $H_0 = 67.36 \pm 0.54$ (derived)
τ	optical depth to reionisation	`tau`	0.0544 ± 0.0073
A_s	amplitude of the primordial curvature spectrum at $k_* = 0.05\,\mathrm{Mpc}^{-1}$	`As`	$\ln(10^{10}A_s) = 3.044 \pm 0.014$
n_s	tilt of that spectrum, $n_s = 1$ being scale invariant	`ns`	0.9649 ± 0.0042

Table T1.1: The six ΛCDM numbers in three notations. Planck samples $\theta_{\rm MC}$ because the data measure that angle almost directly; CAMB is usually called with H_0 instead, and the book's scripts do so (CAMB's `cosmomc_theta` takes $\theta_{\rm MC}$ itself, not $100\,\theta_{\rm MC}$). The error on $\ln(10^{10}A_s)$ is the fractional error on A_s, 1.4%.

The machine's output is a set of averages, not a sky. This is the one idea from cosmology that the statistics cannot do without. Take the CMB. Write the temperature pattern of the sky as a sum of spherical harmonics $Y_{\ell m}$; the coefficient of each one is a number $a_{\ell m}$, and large ℓ

means small angular scales. What does the model predict about these numbers? It predicts a definite value for C_ℓ, the average of $|a_{\ell m}|^2$, computed from the primordial power spectrum and the radiation transfer function. It never predicts any single $a_{\ell m}$. Each $a_{\ell m}$ of our sky is one random draw from a distribution of variance C_ℓ. The matter field works the same way: the model predicts $P(k)$, the variance of each Fourier mode of the matter density, and not the modes themselves. So comparing theory with data always means comparing a predicted *distribution* with one *draw*. A single draw scatters about its mean, and nothing can remove that scatter, because we have only one sky. That scatter is cosmic variance.

Key idea: What the model predicts

A cosmological model predicts the statistics of the fields (their power spectra, correlation functions and distances), never the fields themselves. The observed sky is one realisation. Every estimator in the later chapters is a function of that realisation, and its sampling distribution around the predicted statistic is what the likelihood describes.

T1.2 The objects the statistics uses

A revision of the cosmology a data analyst needs follows the order in which the machine of Section T1.1 works. First comes the smooth background, the expansion of the Universe and the distances that follow from it. Then the primordial seed and how the radiation and matter eras process it. Last come the statistics of the two fields that result, the temperature of the sky and the density of matter. For every object the question comes first, then the equation it starts from and the working equation the book uses, then the size of the thing, then the chapters that use it. The physics between the two equations is worked out in the cosmology notes. Here it is a sentence.

How fast the Universe expands: $H(z)$ and $E(z)$. The question is how the expansion rate H changed over cosmic history, because every distance below is built from it. A homogeneous, isotropic universe has the Friedmann–Lemaître–Robertson–Walker metric, with one function of time in it, the scale factor $a(t)$. Einstein's equations then reduce to the first Friedmann equation

$$H^2 \equiv \left(\frac{\dot a}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{K}{a^2} + \frac{\Lambda}{3},$$

where ρ is the total density of radiation and matter, K the spatial curvature and Λ the cosmological constant. Each component dilutes in its own way as space expands: radiation as a^{-4} (the photons also lose energy), matter as a^{-3}, curvature as a^{-2}, and Λ not at all. Write the redshift as $1+z = a_0/a$ and divide by today's rate H_0. The result is the working equation, in terms of the dimensionless expansion rate $E(z) \equiv H(z)/H_0$:

$$E^2(z) = \Omega_{r0}(1+z)^4 + \Omega_{m0}(1+z)^3 + \Omega_{K0}(1+z)^2 + \Omega_{\Lambda 0}, \qquad \Omega_{K0} = 1 - \Omega_{m0} - \Omega_{r0} - \Omega_{\Lambda 0}.$$

The Ω_{X0} are today's densities as fractions of the critical density $3H_0^2/8\pi G$, with $X = r$ for radiation, m for matter, K for curvature and Λ for the cosmological constant, and the second relation says that the fractions add up to one. For the Planck fit, $\Omega_{m0} \approx 0.315$, $\Omega_{\Lambda 0} \approx 0.685$, $\Omega_{r0} \approx 9 \times 10^{-5}$ and $\Omega_{K0} \approx 0$. The matter term wins over the others at $z \gtrsim 0.3$, and radiation beats matter only for $z \gtrsim 3000$. Today's numbers are $E(1) \approx 1.79$, so the Universe expanded 1.8 times faster at $z = 1$. Of the six parameters of Table T1.1, the data fix not Ω_{m0} itself but the physical density $\Omega_{m0}h^2 = \Omega_{b0}h^2 + \Omega_{c0}h^2$ $(+\Omega_{\nu 0}h^2) \approx 0.143$, which is what the expansion history before the epoch of dark energy depends on. This is the model mean of the supernova and BAO data of Chapters 3 and 5, and the matter density is what the fits of Chapters 9 and 10 constrain.

How far away is a source: D_C, D_M, D_A and D_L. The question is what "distance" means for an object at redshift z, because the same source has a different distance depending on whether we measure it with a ruler or a candle. A light ray travels along a null path, $c\,\mathrm{d}t = a\,\mathrm{d}\chi$ in terms of the comoving radial coordinate χ, and $\mathrm{d}z = -(1+z)H\,\mathrm{d}t$. Integrating along the ray from the source to us gives the comoving distance, with the Hubble distance $D_{H0} \equiv c/H_0$ as the unit:

$$a_0\chi(z) = D_{H0}\int_0^z \frac{\mathrm{d}z'}{E(z')}.$$

Curvature bends the ray's transverse separation, so the comoving transverse distance is $a_0 f_K[\chi(z)]$, where $f_K(\chi) = \chi$ in a flat universe and is a sin or sinh of χ in a closed or open one. A source of physical size s at redshift z subtends the angle $\vartheta = s/D_A$ with the angular-diameter distance, and a source of luminosity L gives the flux $F = L/4\pi D_L^2$ with the luminosity distance:

$$D_A(z) = \frac{a_0 f_K[\chi(z)]}{1+z}, \qquad D_L(z) = (1+z)^2 D_A(z).$$

The first divides by $1+z$ because the source was closer to us when the light left it, and the second multiplies by $(1+z)^2$ because redshift costs each photon energy and stretches the arrival rate. The Hubble distance is $D_{H0} = c/H_0 \approx 4450\,\mathrm{Mpc}$. For the Planck model, a source at $z = 1$ has comoving distance 3.4 Gpc, $D_A = 1.7\,\mathrm{Gpc}$ and $D_L = 6.8\,\mathrm{Gpc}$, and the last-scattering surface at $z \approx 1090$ is 13.9 Gpc away. Note that D_A does not grow forever: it peaks near 1.8 Gpc at $z \approx 1.6$ and then falls, so a rod at high redshift looks larger on the sky. Supernova magnitudes depend on $D_L(z)$, and the BAO ruler of the next paragraph on $D_A(z)$ and $H(z)$ (Chapters 3 and 5).

The acoustic ruler: the sound horizon r_s. The question is how far a sound wave could travel in the early plasma, because that length is a ruler that the CMB calibrates. Before recombination photons and baryons were locked into one fluid. A pressure wave in it moves at the sound speed $c_s = c/\sqrt{3(1+R)}$, which is the radiation value $c/\sqrt{3}$ slowed down by the baryons' inertia. Here $R \equiv 3\rho_b/(4\rho_\gamma)$ is the baryon-to-photon momentum ratio, a number of order 0.6 at the drag epoch. The distance such a wave covers up to conformal time η is

$$r_s(\eta) \equiv \int_0^\eta c_s\,\mathrm{d}\eta' = \int_0^\eta \frac{c\,\mathrm{d}\eta'}{\sqrt{3(1+R)}}.$$

Two epochs matter. At the last scattering of photons ($z_* \approx 1090$) the CMB records the waves. Its angular scale is $\theta_{\mathrm{MC}} = r_s(z_*)/a_0 f_K[\chi_*] \approx 0.0104$ rad, and the acoustic peaks are spaced in ℓ by about $\pi/\theta_{\mathrm{MC}} \approx 300$. At the later drag epoch ($z_{\mathrm{drag}} \approx 1060$) the baryons are released from the photons, the wave stops, and $r_s(z_{\mathrm{drag}}) \approx 147.10\,\mathrm{Mpc}$ is the radius of the shell that galaxies inherit as the BAO scale. The first is the angle in the third row of Table T1.1; the second is the standard ruler of Section T1.4. Both appear in Chapters 5, 7 and 12.

The primordial seed: $\mathcal{P}_\zeta$, A_s and n_s. The question is what the fluctuations looked like before any physics of the radiation or matter eras acted on them. Inflation, in the cosmology notes, stretches quantum fluctuations into a nearly Gaussian random field of comoving curvature perturbation $\zeta(\boldsymbol{x})$. Its Fourier modes are independent, with variance $P_\zeta(k)$. For any spectrum P_X, the dimensionless version

$$\mathcal{P}_X(k) \equiv \frac{k^3 P_X(k)}{2\pi^2}$$

is the variance of the field per logarithmic interval of k, since $\int \mathrm{d}^3k/(2\pi)^3\,P_X = \int \mathrm{d}\ln k\,\mathcal{P}_X$. Inflation predicts a nearly scale-invariant power law. The cosmology notes write it with an

amplitude A_S and a tilt n_S (capital S), and CAMB with A_s and n_s:

$$A_S^2 \equiv \tfrac{4}{25}\,\mathcal{P}_\zeta, \quad A_S^2(k) \propto k^{n_S-1}; \qquad \mathcal{P}_\zeta(k) = A_s\Big(\frac{k}{k_*}\Big)^{n_s-1}, \quad k_* = 0.05\,\mathrm{Mpc}^{-1}.$$

Section T1.2.1 (item 3) converts between them: $A_s = \frac{25}{4}A_S^2(k_*)$. The size is $A_s \approx 2.1\times10^{-9}$, so ζ has an rms of 5×10^{-5} per $\ln k$, and $n_s = 0.965$, slightly less than one, so there is a little more power on large scales than on small. The primordial spectrum is the input of the Gaussian random fields of Chapter 7 and the quantity the parameter fits of Chapter 9 constrain.

From seed to sky: the CMB transfer function and C_ℓ. The question is how the primordial ζ becomes the temperature pattern we see. The photon, baryon, dark matter and neutrino fluids obey the Boltzmann hierarchy, and as the photons stream to us they carry the acoustic oscillations, the gravitational redshifts, and the effect of reionisation. All of this is linear in ζ. So each Fourier mode ζ_k produces a temperature anisotropy at multipole ℓ with a fixed coefficient $g_{T\ell}^{(s)}(k)$, the radiation transfer function [how much anisotropy at multipole ℓ one unit of ζ at wavenumber k produces], and the sum over k of the independent contributions gives

$$C_\ell = \frac{2}{\pi}\int k^2\,\mathrm{d}k\,P_\zeta(k)\,g_{T\ell}^{(s)}(k)^2.$$

This integral is what CAMB evaluates. C_ℓ is the variance of the sky at multipole ℓ, that is at angular scale $\vartheta \approx \pi/\ell$, and the cosmology notes define it through the expansion of the temperature contrast $\Theta = \delta T/T = \sum_{\ell m} a_{\ell m}Y_{\ell m}$:

$$\langle a_{\ell m}a^*_{\ell' m'}\rangle = \delta_{\ell\ell'}\delta_{mm'}\,C_\ell, \qquad C(\vartheta) = \frac{1}{4\pi}\sum_\ell (2\ell+1)\,C_\ell\,P_\ell(\cos\vartheta).$$

The second relation is the correlation of two directions an angle ϑ apart, with P_ℓ the Legendre polynomial. Because the sky is one draw, the estimator is $\widehat{C}_\ell = \frac{1}{2\ell+1}\sum_m |a_{\ell m}|^2$, and it obeys $(2\ell+1)\widehat{C}_\ell/C_\ell \sim \chi^2_{2\ell+1}$, the cosmic-variance law of Chapter 7. The plotted combination is $D_\ell \equiv \ell(\ell+1)C_\ell/2\pi$. For the fiducial model D_ℓ^{TT} peaks at $5732\,\mu\mathrm{K}^2$ at $\ell = 220$, and the fractional cosmic variance at one ℓ is $\sqrt{2/(2\ell+1)}$, about 9% at $\ell = 100$. The spectrum is the input of Chapters 7 to 10.

From seed to structure: the matter transfer function and growth. The question is how the same seed becomes the density field of galaxies. Two things happen, one depending on scale and one not. The first is growth. A small overdensity in pressureless matter grows according to the equation

$$\ddot{D} + 2H\dot{D} - \tfrac{3}{2}H^2\Omega_m(t)\,D = 0, \qquad D_+(a) = \frac{5}{2}\frac{H(a)}{H_0}\,\Omega_{m0}\int_0^a \frac{\mathrm{d}a'}{[a'E(a')]^3},$$

where the damping term $2H\dot{D}$ is Hubble friction and the last term is gravity. The growing solution is the integral on the right, and in a matter-dominated universe it is simply $D_+ \propto a$. The second is the scale dependence. A mode that enters the horizon during radiation domination is held back, because radiation pressure prevents it from clumping, while a mode that enters later grows undisturbed. The **transfer function** packages this,

$$T(k,a) = \frac{\delta(k,a)\,D_+(a_\mathrm{i})}{\delta(k,a_\mathrm{i})\,D_+(a)}, \qquad P_\delta(k,a_0) = P_\delta(k,a_\mathrm{i})\,T^2(k,a_0)\Big[\frac{D_+(a_0)}{D_+(a_\mathrm{i})}\Big]^2,$$

where a_i is an early time before any scale of interest has entered the horizon, and the second relation is the first one squared. The transfer function goes to one for $k \ll k_\mathrm{eq}$ and falls as $(k_\mathrm{eq}/k)^2$ for $k \gg k_\mathrm{eq}$ (up to a logarithm), where $k_\mathrm{eq} \approx 0.017\,h\,\mathrm{Mpc}^{-1}$ is the wavenumber of the horizon at matter–radiation equality. Growth is a size we can read off: the matter fluctuations at $z = 1$ are 0.61 of those today (0.4928 against 0.8111 in the σ_8 below). Used in Chapter 7.

The statistics of the matter field: $P(k)$, $\xi(r)$ and σ_8. The question is how to describe a random field of matter density with statistics we can predict. Write $\delta = (\rho - \bar\rho)/\bar\rho$ for the density contrast, with Fourier modes $\delta_{\boldsymbol{k}}$. Homogeneity makes distinct modes uncorrelated, and the variance of one mode is the matter power spectrum $P_\delta(k)$ (also written P_m). Its real-space partner is the correlation function $\xi(r)$, the excess probability of finding two points at separation r:

$$\langle \delta_{\boldsymbol{k}} \delta_{\boldsymbol{k}'} \rangle = \delta^{(3)}(\boldsymbol{k} + \boldsymbol{k}')\, P_\delta(k), \qquad \xi(r) = \int_0^\infty \frac{\mathrm{d}k}{2\pi^2}\, k^2 P_\delta(k)\, \frac{\sin kr}{kr}.$$

The second equation is the Fourier transform of the first, after the angle average of $e^{i\boldsymbol{k}\cdot\boldsymbol{r}}$ is done, which gives $\sin kr/kr$. P_δ peaks near k_eq at a few $10^4\,(h^{-1}\mathrm{Mpc})^3$, and the two spectra of linear theory and halofit separate at $k \approx 0.15\,h\,\mathrm{Mpc}^{-1}$ at $z = 0$. The single-number summary of the amplitude is the rms fluctuation in a sphere of radius $8\,h^{-1}$Mpc:

$$\sigma_8^2 \equiv \int \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3}\, \left|W_{8h^{-1}\mathrm{Mpc}}(k)\right|^2 P_m(k), \qquad W_R(k) = \frac{3j_1(kR)}{kR}.$$

Here W_R is the Fourier transform of a spherical top-hat of radius R and j_1 the spherical Bessel function, so the integral weights P by how much of each mode a sphere of that size sees. The value is $\sigma_8 = 0.8111$: fluctuations in $8\,h^{-1}$Mpc spheres are about 80% of the mean density, just short of non-linear. Section T1.4 computes both $\xi(r)$ and σ_8 from the CAMB $P(k)$. They feed the Gaussian-field simulations of Chapter 7, the Fisher forecasts of Chapter 10 and the topological summaries of Chapter 12.

Lensing of the CMB. The question is what happens to the pattern on the way to us. Matter between us and the last-scattering surface deflects the photons by a small angle, so the temperature we see in direction $\boldsymbol{n}$ was emitted from a slightly different direction:

$$\hat T[\boldsymbol{n}] = T[\boldsymbol{n} + \boldsymbol{\xi}(\boldsymbol{n})], \qquad \boldsymbol{\xi} = \boldsymbol{\nabla}\psi.$$

The deflection $\boldsymbol{\xi}$ (not the correlation function $\xi(r)$) is the gradient on the sky of a single scalar, the projected lensing potential ψ, which is the matter potential integrated along the line of sight with a geometric weight. The size is a few arcminutes, which is why the effect is a smoothing of the acoustic peaks and a coupling between modes, not a visible distortion. The same ψ builds the shear of galaxy images in Section T1.6, with a source plane at last scattering. Used in Chapters 8 and 12.

The primordial three-point function: f_NL. The question is whether the seed is exactly Gaussian. A Gaussian ζ is fully described by P_ζ and has zero three-point function. Any departure shows up in the bispectrum B_ζ, the Fourier transform of $\langle\zeta\zeta\zeta\rangle$, and the amplitude f_NL is defined by comparing it with the products of two power spectra:

$$\tfrac{6}{5} f_\mathrm{NL} \equiv \frac{B_\zeta(k_1,k_2,k_3)}{P_\zeta(k_1)P_\zeta(k_2) + P_\zeta(k_1)P_\zeta(k_3) + P_\zeta(k_2)P_\zeta(k_3)}.$$

Planck finds $f_\mathrm{NL}^\mathrm{local} = -0.9 \pm 5.1$, consistent with a Gaussian field. A non-zero f_NL is invisible to C_ℓ and $P(k)$ and is searched for with the tools of Chapter 12.

object	working equation	typical size	used in
$E(z)$	$E^2 = \Omega_{r0}(1+z)^4 + \Omega_{m0}(1+z)^3 + \Omega_{K0}(1+z)^2 + \Omega_{\Lambda 0}$	$\Omega_{m0} \approx 0.315$, $E(1) \approx 1.79$	Chapters 3 and 5
distances	$a_0\chi = D_{H0}\int \mathrm{d}z/E$; $D_A = a_0 f_K/(1+z)$; $D_L = (1+z)^2 D_A$	$D_{H0} \approx 4450\,\mathrm{Mpc}$; $D_L(1) \approx 6.8\,\mathrm{Gpc}$	Chapters 3 and 5
r_s	$\int c_s \mathrm{d}\eta = \int c\,\mathrm{d}\eta/\sqrt{3(1+R)}$	$\approx 147.10\,\mathrm{Mpc}$ at drag	Chapters 5, 7 and 12
primordial	$\mathcal{P}_\zeta = A_s(k/k_*)^{n_s-1}$; $A_s = \frac{25}{4}A_S^2$	$A_s \approx 2.1\times10^{-9}$, $n_s = 0.965$	Chapters 7 and 9
C_ℓ, D_ℓ	$C_\ell = \frac{2}{\pi}\int k^2\mathrm{d}k\, P_\zeta g_{T\ell}^2$; $\langle a_{\ell m} a^*_{\ell' m'}\rangle = \delta\delta C_\ell$; $(2\ell+1)\widehat{C}_\ell/C_\ell \sim \chi^2_{2\ell+1}$	D_ℓ peak $5732\,\mu\mathrm{K}^2$ at $\ell = 220$	Chapters 7 to 10
$T(k,a)$, D_+	$P_\delta(k,a_0) = P_\delta(k,a_i)T^2[D_+(a_0)/D_+(a_i)]^2$	$T \to (k_{\mathrm{eq}}/k)^2$ beyond $k_{\mathrm{eq}} \approx 0.017\,h\,\mathrm{Mpc}^{-1}$	Chapter 7
$P(k)$, $\xi(r)$	$\langle\delta_{\boldsymbol{k}}\delta_{\boldsymbol{k}'}\rangle = \delta^{(3)}P_\delta$; $\xi = \int \frac{\mathrm{d}k}{2\pi^2}k^2 P \frac{\sin kr}{kr}$	BAO bump at $\approx 101.5\,h^{-1}\mathrm{Mpc}$	Chapters 7 and 12
σ_8	$\int \frac{\mathrm{d}^3k}{(2\pi)^3}\lvert W_8\rvert^2 P_m$; $W_R = 3j_1(kR)/kR$	0.8111	Chapters 7 and 10
lensing	$\hat{T}[\boldsymbol{n}] = T[\boldsymbol{n} + \boldsymbol{\nabla}\psi]$	a few arcmin	Chapters 8 and 12
f_{NL}	$\frac{6}{5}f_{\mathrm{NL}} = B_\zeta/(PP + \mathrm{perms})$	-0.9 ± 5.1	Chapter 12

Table T1.2: The objects of the section at a glance: the working equation of each, a typical size, and where it is used.

T1.2.1 Conventions to carry forward

The same object is often written differently in the cosmology notes, in CAMB and in this book. There are only a few such differences. Each one, if missed, changes a number by a large factor without any warning.

1. **C_ℓ versus D_ℓ.** The book writes $D_\ell \equiv \ell(\ell+1)C_\ell/2\pi$, the combination that is usually plotted, because it makes the acoustic peaks comparable in height. CAMB's `get_cmb_power_spectra` returns D_ℓ by default and returns the raw C_ℓ only with `raw_cl=True`. The fiducial spectra in `data/camb_fiducial.npz` are raw C_ℓ. Every statistic of the later chapters ($\widehat{C}_\ell$, cosmic variance $2C_\ell^2/(2\ell+1)$, the $\chi^2_{2\ell+1}$ law) is written for the raw C_ℓ.

2. **Units of C_ℓ.** In the cosmology notes $\Theta = \delta T/T$ is dimensionless, so C_ℓ is too. With `CMB_unit="muK"`, CAMB returns $T_0^2\,C_\ell$ in $\mu\mathrm{K}^2$, with $T_0 = 2.7255\,\mathrm{K} = 2.7255\times10^6\,\mu\mathrm{K}$. Noise spectra must be in the same units as the signal before they are added.

3. **The primordial amplitude.** The cosmology notes normalise the scalar amplitude as $A_S^2 \equiv \frac{4}{25}\mathcal{P}_\zeta$. CAMB and Planck use $\mathcal{P}_\zeta(k) = A_s(k/k_*)^{n_s-1}$ with pivot $k_* = 0.05\,\mathrm{Mpc}^{-1}$, so that $A_s = \mathcal{P}_\zeta(k_*) = \frac{25}{4}A_S^2(k_*)$. The Planck number $\ln(10^{10}A_s) = 3.044$ refers to CAMB's A_s.

4. **Fourier conventions and $P(k)$.** Two Fourier conventions appear, and both lead to the same numerical $P(k)$, the one CAMB returns; only the non-linear mapping needs care. The matter spectrum in the cosmology notes uses the symmetric transform (a factor $(2\pi)^{-3/2}$ on each side), with $\langle\delta_{\boldsymbol{k}}\delta_{\boldsymbol{k}'}\rangle = \delta^{(3)}(\boldsymbol{k}+\boldsymbol{k}')P_\delta$. The non-Gaussianity material puts $(2\pi)^3\delta^{(3)}$ there instead. Both give the same $\xi(r)$ relation in the table above, and so the same $P(k)$. The non-linear mapping in the cosmology notes writes the dimensionless variance as $4\pi k^3 P(k)$. That form assumes a P carrying an extra factor $(2\pi)^{-3}$. With CAMB's $P(k)$ the same quantity

is $k^3P(k)/2\pi^2$.

5. **Factors of h.** CAMB's `get_matter_power_spectrum` takes k in $h\,\mathrm{Mpc}^{-1}$ and returns P in $(h^{-1}\mathrm{Mpc})^3$; σ_8 is defined in h^{-1}Mpc. Derived lengths, such as the drag-epoch sound horizon, come back in plain Mpc. To compare one with $\xi(r)$ plotted against r in h^{-1}Mpc, multiply it by $h = H_0/100$.

6. **Lensing potential.** The cosmology notes call the projected potential ψ; CAMB and Planck call the same object ϕ. CAMB's `lens_potential` output is $[\ell(\ell+1)]^2 C_\ell^{\phi\phi}/2\pi$, not $C_\ell^{\phi\phi}$. The fiducial CMB spectra of the book are lensed.

7. **Real and complex $a_{\ell m}$.** The cosmology notes write the one-point density of $a_{\ell m}$ for a real Gaussian variable of variance C_ℓ. In fact, for $m \neq 0$ the $a_{\ell m}$ are complex, with $a_{\ell,-m} = (-1)^m a_{\ell m}^*$, and their real and imaginary parts each have variance $C_\ell/2$. The count does not change: there are still $2\ell+1$ independent real numbers per ℓ, so the $\chi^2_{2\ell+1}$ law still holds. Chapter 7 does the counting.

Watch out: A factor that looks like a physics result

Mixing A_S^2 with CAMB's A_s moves the amplitude by 25/4; mixing C_ℓ with D_ℓ moves it by $\ell(\ell+1)/2\pi$, about 10^5 at $\ell = 1000$. Neither error crashes a script. Both give a spectrum with a sensible shape and a wrong height. Checking that D_ℓ^{TT} of the fiducial model peaks at 5732 $\mu\mathrm{K}^2$ at $\ell = 220$ catches both.

T1.3 Experiment 1: how each parameter moves the CMB spectra

The inference chapters need one thing from the machine above all. If we turn one input knob a little, how does the output move? For the CMB the output is the whole curve C_ℓ, so the answer is a curve for each parameter. We find it the most direct way. Write θ_i for the i-th of the six parameters of Table T1.1 ($i = 1, \ldots, 6$; this θ_i is any parameter, not the angle θ_{MC}), and δ_i for its Planck 1σ error.

1. Start at the Planck best fit.
2. Run CAMB at $\theta_i + \delta_i$ and at $\theta_i - \delta_i$, with the other five parameters held at the best fit.
3. Form the symmetric fractional response

$$R_i(\ell) \equiv \frac{C_\ell(\theta_i+\delta_i) - C_\ell(\theta_i-\delta_i)}{2\,C_\ell(\text{fiducial})}, \qquad \frac{\partial C_\ell}{\partial \theta_i} \approx \frac{C_\ell\, R_i(\ell)}{\delta_i}. \tag{T1.1}$$

The second relation follows from the first. A derivative is approximately the change in C_ℓ divided by the change in θ_i, here $[C_\ell(\theta_i+\delta_i) - C_\ell(\theta_i-\delta_i)]/2\delta_i$; by the definition of R_i, the numerator equals $2\,C_\ell\,R_i(\ell)$, and the 2 cancels. This second relation is why the experiment matters. The derivatives $\partial C_\ell/\partial\theta_i$ are exactly what enters the CMB Fisher matrix of Chapter 10, and `code/chT1/01_cl_derivatives.py` computes them by this central difference. The step $\delta_i = 1\sigma$ also gives the size of R_i a plain meaning: a response of 1% is what a change of one standard deviation in that parameter does to the spectrum.

Listing T1.1: Planck 1σ steps, one CAMB call, and the $\pm\delta$ loop (`code/chT1/01_cl_derivatives.py`)

```
STEP = dict(ombh2=0.00015, omch2=0.0012, H0=0.54, tau=0.0073,
            As=0.014 * FIDUCIAL["As"], ns=0.0042)
LABEL = dict(ombh2=r"$\Omega_b h^2$", omch2=r"$\Omega_c h^2$", H0=r"$H_0$",
             tau=r"$\tau$", As=r"$A_s$", ns=r"$n_s$")
```

```
34
35
36 def spectra(params):
37     """Lensed TT, EE in muK^2 (raw C_ell) for ell = 0..LMAX."""
38     import camb
39     p = camb.set_params(**params, lmax=LMAX + 300, lens_potential_accuracy=1)
40     r = camb.get_results(p)
41     tot = r.get_cmb_power_spectra(p, CMB_unit="muK", raw_cl=True)["total"][: LMAX + 1]
42     return tot[:, 0], tot[:, 1]
43
44
45 def compute():
46     tt0, ee0 = spectra(FIDUCIAL)
47     out = dict(tt0=tt0, ee0=ee0)
48     for name, d in STEP.items():
49         for sgn, tag in ((+1, "p"), (-1, "m")):
50             par = dict(FIDUCIAL)
51             par[name] = FIDUCIAL[name] + sgn * d
52             out[f"tt_{name}_{tag}"], out[f"ee_{name}_{tag}"] = spectra(par)
```

The steps in lines 30–31 are the Planck errors of Table T1.1; A_s is stepped by 1.4% of itself because Planck quotes the error on $\ln(10^{10}A_s)$. The function `spectra` (lines 36–42) makes one CAMB call and keeps the lensed TT and EE spectra as raw C_ℓ in $\mu\mathrm{K}^2$ up to $\ell = 2500$. The loop in lines 48–52 moves one parameter up and down while the other five stay at the fiducial values. That makes thirteen CAMB runs in all: one at the fiducial model, plus two for each of the six parameters. The results are cached in `data/chT1/cl_derivs.npz`.

Figure T1.2 shows the six responses. Each can be read physically in one sentence:

- A_s: a flat +1.41% at every ℓ, because the primordial amplitude multiplies the whole spectrum.
- τ: a flat −1.46% for $\ell \gtrsim 30$, the $e^{-2\tau}$ screening of the anisotropies by the reionised electrons, with no effect on the largest scales of TT and a large rise (46%) in the low-ℓ EE reionisation bump, which those same electrons create (see the cosmology notes).
- n_s: a straight line in $\ln\ell$, rotating the spectrum about $\ell \approx 542$, since a tilt adds power on one side of the pivot and removes it on the other.
- $\Omega_{b0}h^2$ and $\Omega_{c0}h^2$: wiggles in phase with the acoustic peaks, largest 1.1% for $\Omega_{c0}h^2$, because the baryon loading and the epoch of matter–radiation equality set the relative heights of the peaks (see the cosmology notes).
- H_0, at fixed physical densities: a slow drift at low ℓ from the late-time decay of the potentials, and oscillations at high ℓ because a changed distance to last scattering slides every peak sideways in ℓ (see the cosmology notes).

Two of these curves are nearly the same curve with opposite sign. Above $\ell \approx 30$ the A_s response is flat and positive, and the τ response is flat and negative. How alike are they? Treat each response as a list of numbers, one per ℓ, and compute the cosine of the angle between the two lists, giving each ℓ the weight $2\ell + 1$ (the number of $a_{\ell m}$ at that ℓ). A cosine of +1 or −1 means the same shape up to a factor. Over $30 \le \ell \le 2500$ it is −0.999. The reason is visible in the list above: at high ℓ the spectrum is proportional to A_s times the screening factor $e^{-2\tau}$, so it depends on the two only through $A_s e^{-2\tau}$. Raise A_s and raise τ together so that $A_s e^{-2\tau}$ stays fixed, and the high-ℓ spectrum does not change. Figure T1.3 puts $+R_{A_s}$ on top of $-R_\tau$ to make this visible.

This is what a degeneracy looks like before any statistics is done: two columns of derivatives that are almost parallel. In Chapter 10 it makes the Fisher matrix nearly singular along one

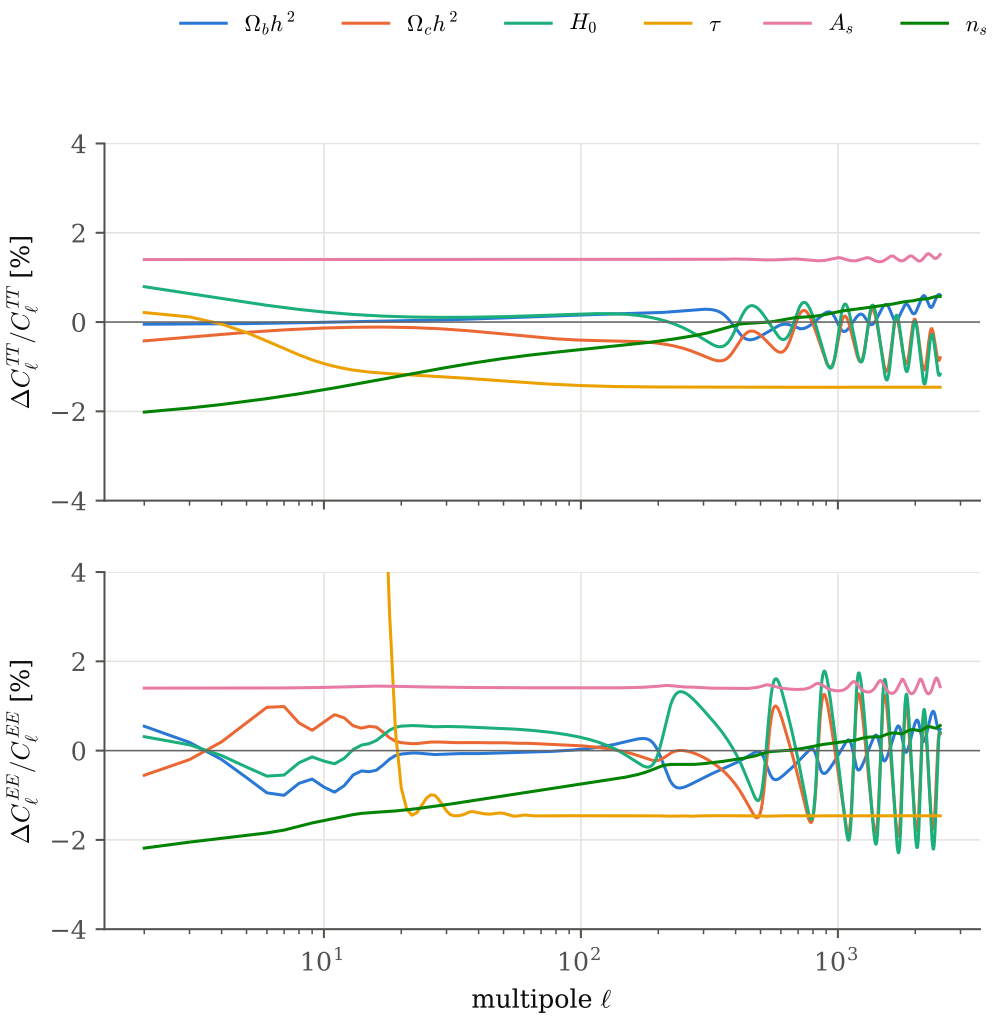


Figure T1.2: Fractional change of the lensed TT (top) and EE (bottom) spectra when each ΛCDM parameter moves by its Planck 1σ error, Eq. (T1.1). A flat line is a pure change of amplitude, a slanted line is a tilt, and a wiggle means the acoustic peaks change height or position. The τ curve in EE leaves the plot at low ℓ (it reaches 46%).

direction. As a result, the marginalised errors on A_s and τ (each parameter's error when the other is left free) are far larger than the conditional ones (its error when the other is held fixed). In Chapter 9 the same degeneracy appears as a long, thin ridge in the MCMC samples of $(\ln 10^{10} A_s, \tau)$. The right panel shows how it is broken in practice. Large-angle polarisation measures τ on its own, and A_s then follows from the high-ℓ amplitude.

How this is used in real research: Verification and research

Here the finite differences only *verify* what the physics already tells us (a flat A_s response, the $e^{-2\tau}$ screening, the tilt about a pivot). In a Fisher forecast the same derivatives are the research instrument, and their accuracy has to be tested: the step must be small enough that C_ℓ is linear in θ_i over $\pm\delta_i$, and large enough that CAMB's numerical noise does not dominate the difference. The standard check is to halve and double δ_i and confirm that the derivative does not move; Chapter 10 does this.

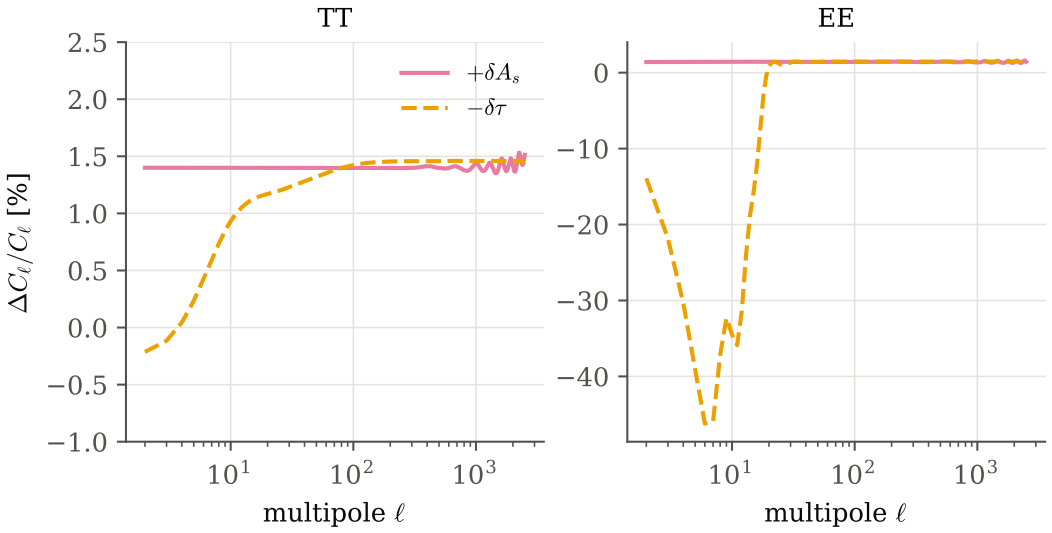


Figure T1.3: The A_s–τ degeneracy. The solid curve is the response to $+1\sigma$ in A_s, the dashed curve minus the response to $+1\sigma$ in τ. In TT (left) they coincide for $\ell \gtrsim 30$, so high-ℓ temperature data measure only the combination $A_s e^{-2\tau}$. In EE (right) they separate completely at $\ell \lesssim 20$, where the reionisation bump responds to τ alone.

T1.4 Experiment 2: the matter power spectrum and the BAO bump

The second experiment turns to the matter field, the three-dimensional counterpart of the CMB sky. CAMB returns the matter power spectrum $P(k)$ at any redshift, either in linear theory or with the halofit correction [41] [a fit to N-body simulations that maps the linear $P(k)$ to the non-linear one]. From the linear $P(k)$ we compute two things ourselves. The first is the correlation function $\xi(r)$, from the Fourier integral in the $P(k)$ row of the last table in Section T1.2. The second is σ_8, from the top-hat integral in the σ_8 row. CAMB also reports σ_8, so the second is a check on our code.

Listing T1.2: $\xi(r)$ and σ_R from the linear $P(k)$ (`code/chT1/02_pk_xi.py`)

```
48 def xi_of_r(k, pk, r, s=1.0):
49     """xi(r) as in the cosmology notes, on a dense log-k grid with Gaussian damping."""
50     kk = np.logspace(-4, np.log10(50), 20000)
51     pp = np.exp(np.interp(np.log(kk), np.log(k), np.log(pk))) * np.exp(-(kk * s) ** 2)
52     kr = np.outer(r, kk)
53     integrand = kk ** 3 * pp * np.sinc(kr / np.pi) / (2 * np.pi ** 2)  # d ln k form
54     return np.trapezoid(integrand, np.log(kk), axis=1)
55
56
57 def tophat_sigma(k, pk, R):
58     """sigma_R from the linear P(k) and a spherical top hat of radius R (Mpc/h)."""
59     x = k * R
60     W = 3 * (np.sin(x) - x * np.cos(x)) / x ** 3
61     return np.sqrt(np.trapezoid(k ** 3 * pk * W ** 2 / (2 * np.pi ** 2), np.log(k)))
```

`xi_of_r` evaluates $\xi(r) = \int \frac{\mathrm{d}k}{2\pi^2} k^2 P(k) \frac{\sin kr}{kr}$ as an integral over $\ln k$. Since $\mathrm{d}k = k\,\mathrm{d}\ln k$, the factor $k^2\,\mathrm{d}k$ becomes $k^3\,\mathrm{d}\ln k$; that is the k^3 in line 53. (`np.sinc(x/np.pi)` is $\sin x/x$.) The factor $e^{-k^2 s^2}$ with $s = 1\,h^{-1}$Mpc in line 51 is there only so the integral converges: it damps the fast-oscillating tail of the integrand. It acts on scales a hundred times smaller than the BAO peak and does not move the peak. `tophat_sigma` is the σ_R integral with the window $W_R(k) = 3(\sin kR - kR\cos kR)/(kR)^3 = 3j_1(kR)/(kR)$. At $R = 8\,h^{-1}$Mpc it gives $\sigma_8 = 0.8112$, against CAMB's 0.8111, and 0.4928 at $z = 1$, where structure has had less time to grow.

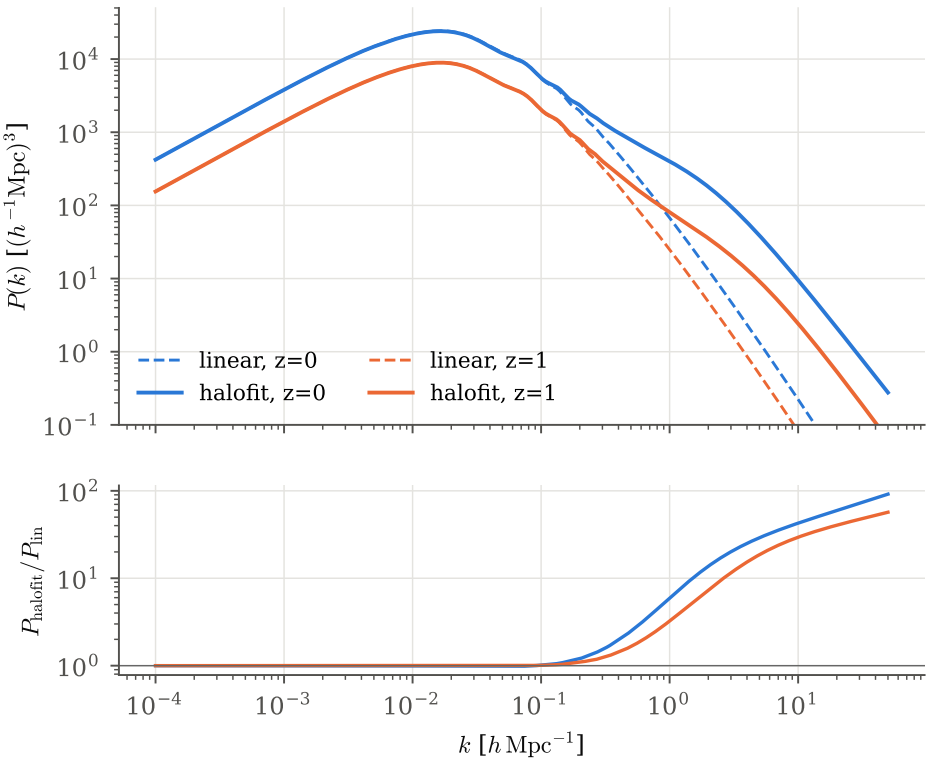


Figure T1.4: Matter power spectrum of the fiducial model at $z = 0$ (blue) and $z = 1$ (orange), linear theory dashed and halofit solid. The lower panel is the ratio halofit/linear: the two agree on large scales and differ by 10% at $k \approx 0.15\,h\,\mathrm{Mpc}^{-1}$ at $z = 0$, and only at $k \approx 0.21\,h\,\mathrm{Mpc}^{-1}$ at $z = 1$.

Figure T1.4 shows two pieces of physics.

The first is the turnover at $k \approx 0.017\,h\,\mathrm{Mpc}^{-1}$. It marks the size of the horizon at matter–radiation equality. Modes with larger k (smaller scales) entered the horizon earlier, during radiation domination, and grew little there. So the transfer function falls roughly as $(k_{\mathrm{eq}}/k)^2$ (up to a logarithm), where k_{eq} is the turnover wavenumber. Since $P(k) \propto k^{n_s} T^2$ and $n_s \approx 1$, beyond the turnover $P(k)$ falls roughly as $k \cdot k^{-4} = k^{-3}$ (see the cosmology notes).

The second is where halofit leaves linear theory. At $z = 0$ the two differ by 10% at $k \approx 0.15\,h\,\mathrm{Mpc}^{-1}$ and by a factor 5.9 at $k = 1\,h\,\mathrm{Mpc}^{-1}$. The reason is gravitational collapse into haloes, which piles up power on small scales. For the statistics this boundary matters more than any number on the plot. On large scales the matter field is close to Gaussian, and $P(k)$ carries all of its information; this is the regime of the Gaussian random fields of Chapter 7. On small scales the field is non-Gaussian, and $P(k)$ no longer tells the whole story. There, summaries that see the shape of the field, the topological summaries of Chapter 12, have something extra to find.

Figure T1.5 shows the same information in real space, and there the baryon acoustic oscillations are much easier to see. In $P(k)$ they are small wiggles on a steep curve. In $\xi(r)$ they collapse into a single bump near $r \approx 101.5\,h^{-1}\mathrm{Mpc}$. Where does the bump come from? Each initial overdensity launched a spherical sound wave into the photon–baryon plasma. The wave travelled until the baryons were released from the photons at the drag epoch, and then stopped. Its radius at that moment is the sound horizon $r_s(z_{\mathrm{drag}})$ of the BAO row, and it leaves an excess of pairs of points at that separation (see the cosmology notes). CAMB gives $r_s(z_{\mathrm{drag}}) = 147.10\,\mathrm{Mpc}$, that is $99.1\,h^{-1}\mathrm{Mpc}$. The peak of $r^2\xi$ sits a couple of percent higher because it rides on the sloping broad-band part of ξ.

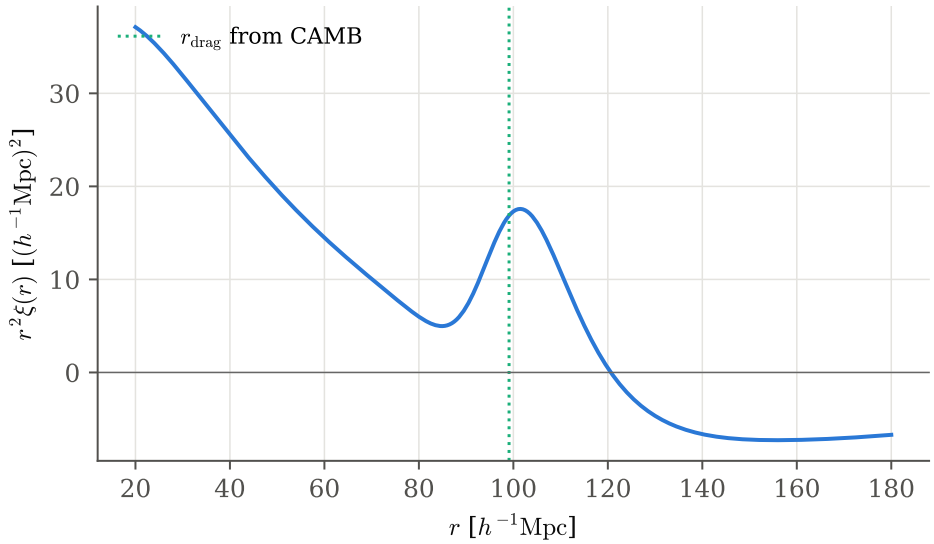


Figure T1.5: The linear correlation function at $z = 0$, multiplied by r^2 so that the bump is visible. The peak at $r \approx 101.5\,h^{-1}$Mpc sits close to the drag-epoch sound horizon computed by CAMB (dotted line, $99.1\,h^{-1}$Mpc).

Why does this matter? The CMB fixes r_s precisely, so the bump is a ruler of known length. Its observed position in a galaxy survey at redshift z therefore measures $D_A(z)$ (from its angular size) and $H(z)$ (from its extent along the line of sight). This is how BAO data enter the parameter fits mentioned in Chapter 5. The bump was first detected in the galaxy correlation function by [42]. In Chapter 7 it is a feature to recover from a simulated field with noisy estimates of $\xi(r)$. In Chapter 12 it is one of the structures that a summary beyond the two-point function must respect.

Recap

- Six numbers ($\Omega_{b0}h^2, \Omega_{c0}h^2, H_0$ or $\theta_{\rm MC}, \tau, A_s, n_s$) go into CAMB; out come $E(z)$ and distances, C_ℓ, $P(k)$, $\xi(r)$, σ_8.
- The model predicts statistics, $C_\ell = \langle|a_{\ell m}|^2\rangle$ and $P(k)$, never the $a_{\ell m}$ or $\delta_{\boldsymbol{k}}$ themselves. Our sky is one draw.
- Conventions: raw C_ℓ (not D_ℓ) in all statistics; μK^2 means a factor T_0^2; CAMB's $A_s = \frac{25}{4}A_S^2$ of the cosmology notes; $P(k)$ in $(h^{-1}\mathrm{Mpc})^3$; CAMB's ϕ is the notes' ψ.
- Nudging each parameter by 1σ gives $\partial C_\ell/\partial\theta_i$, the Fisher derivatives; A_s and τ respond identically above $\ell \approx 30$ (cosine -0.999), and low-ℓ EE breaks the tie.
- Linear and halofit $P(k)$ separate at $k \approx 0.15\,h\,\mathrm{Mpc}^{-1}$ today; $\xi(r)$ has the BAO bump near $r_s(z_{\rm drag}) \approx 99.1\,h^{-1}$Mpc; our σ_8 matches CAMB's to better than 0.1%.

T1.5 Weak lensing: the matter field written on galaxy shapes

Light from a distant galaxy crosses billions of light years of clumpy matter on its way to us. Every lump it passes bends it a little. What arrives is an image that is slightly displaced, slightly enlarged and slightly stretched in one direction, typically by about one per cent. Nothing in the image says which part of it is the stretch: a galaxy that looks elongated may simply be elongated. Yet the stretch is the cleanest probe of matter we have, because gravity bends light

whatever the matter is made of, dark or luminous, and no assumption about how galaxies trace mass enters. This is **weak gravitational lensing**, and **cosmic shear** is its version with the large-scale structure as the lens [43, 44].

For the data analyst, weak lensing is the situation of Section T1.1 in a sharper form. There the model predicted the variance C_ℓ of the CMB sky and the sky was one draw. Here the model predicts the angular power spectrum $C_\ell^{\kappa\kappa}$ of a lensing field κ (defined below), the sky gives one draw of κ, and on top of that we do not see κ itself. We see the shapes of galaxies, and each shape is mostly the galaxy's own.

The question. What we observe is a catalogue: for every galaxy, a position on the sky and an **ellipticity** [a complex number $\varepsilon = \varepsilon_1 + i\varepsilon_2$ whose size is how elongated the image is and whose phase is twice the angle of its long axis]. What could have produced the ellipticity of one particular galaxy? There are two candidates, and we take them one at a time.

1. *The galaxy's own shape.* Galaxies are not round. Their intrinsic ellipticities $\varepsilon^{(0)}$ scatter with an rms $\sigma_e \approx 0.27$ per component [45], and their long axes point in random directions. So the intrinsic part has mean zero, but it is large.
2. *The shear of the matter in front of it.* The lensing stretch γ is about 0.01, thirty times smaller. But it is *coherent*: two galaxies that lie close together on the sky sent their light past the same lumps, so they are stretched in nearly the same way.

Only now do we compute anything. In the weak regime the two simply add,

$$\varepsilon \approx \varepsilon^{(0)} + \gamma. \tag{T1.2}$$

The reason is that ε is defined from the shape of the image, and a shear of one per cent changes that shape by one per cent: to first order in small quantities it moves the ellipticity by γ itself, whatever the galaxy's own shape was. Section T1.6.1 works this out for a round source; with the other common definition of ε the shift is 2γ instead (convention 3 of Section T1.6.2). A single galaxy tells us almost nothing, since the shear is buried under a noise thirty times larger. But the intrinsic parts of different galaxies are independent, and the shears of neighbours are correlated. Two moves turn this into a measurement.

Average galaxies that share one shear. Take N galaxies close enough together that they all feel the same γ, and average their ellipticities:

$$\bar\varepsilon \equiv \frac{1}{N}\sum_{a=1}^{N}\varepsilon_a \overset{(1)}{=} \gamma + \frac{1}{N}\sum_{a=1}^{N}\varepsilon_a^{(0)}, \qquad \mathbb{E}\,\bar\varepsilon \overset{(2)}{=} \gamma, \qquad \mathrm{Var}\,\bar\varepsilon_1 \overset{(3)}{=} \frac{\sigma_e^2}{N}. \tag{T1.3}$$

(1) Insert (T1.2) for each galaxy; the shear is the same in every term, so its average is itself.
(2) The intrinsic ellipticities point in random directions, so each has mean zero.
(3) The variance of a mean of N independent readings of variance σ_e^2, (3.20), applied to one real component; the same holds for $\bar\varepsilon_2$.

So the noise on the shear estimate is $0.27/\sqrt{N}$ per component. How large must N be? A signal-to-noise ratio of one on a single shear of 0.01 needs $0.27/\sqrt{N} = 0.01$, that is $N = (0.27/0.01)^2 \approx 730$ galaxies, all sharing that one shear. A map with many independent shear values, each measured several times better than that, needs millions of galaxies.

Correlate pairs. The shear changes across the sky, so averaging is limited to small patches. Correlation uses every pair instead. For two galaxies $a \neq b$, multiply the ellipticity of one by

the complex conjugate of the other and expand:

$$\begin{aligned}\left\langle \varepsilon_a \varepsilon_b^* \right\rangle &\overset{(1)}{=} \left\langle (\varepsilon_a^{(0)} + \gamma_a)(\varepsilon_b^{(0)} + \gamma_b)^* \right\rangle \\ &\overset{(2)}{=} \left\langle \varepsilon_a^{(0)} \varepsilon_b^{(0)*} \right\rangle + \left\langle \varepsilon_a^{(0)} \gamma_b^* \right\rangle + \left\langle \gamma_a \varepsilon_b^{(0)*} \right\rangle + \left\langle \gamma_a \gamma_b^* \right\rangle \overset{(3)}{=} \left\langle \gamma_a \gamma_b^* \right\rangle . \end{aligned} \tag{T1.4}$$

(1) Equation (T1.2) for each galaxy.
(2) Multiply out the brackets; the average of a sum is the sum of the averages.
(3) The intrinsic shapes of two different galaxies are independent, each with mean zero, so the first term is the product of two zero means. In the two middle terms an intrinsic shape is independent of the shear in front of the other galaxy, and the same reasoning makes them vanish. Only the correlation of the shears survives.

Galaxy a paired with itself would add $\langle \left|\varepsilon_a^{(0)}\right|^2 \rangle$, which does not vanish; this is why the self-pairs are left out, and why shape noise reappears as an added term in the power spectrum (Section T1.6). The information is in the statistics of many shapes, never in one shape. This is why every weak-lensing number in the later chapters is a two-point function, a peak count or a topological summary of a field built from millions of galaxies, and why its noise term is set by how many galaxies there are.

Figure T1.6 shows what the two parts of the lensing distortion do, exaggerated twenty-five-fold. The **convergence** κ [the projected matter density along the line of sight, in units of a critical density] enlarges an image without changing its shape. The **shear** $\gamma = \gamma_1 + i\gamma_2$ [the tidal part of the lensing distortion, which stretches the image along one axis and squeezes it along the perpendicular one] turns a circle into an ellipse.

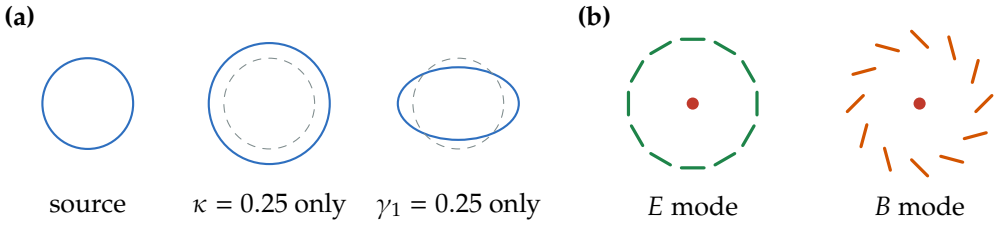


Figure T1.6: (a) A round source (left), the same source magnified by convergence alone (middle), and stretched by shear alone (right); the dashed circle is the source for comparison. Real values are near 0.01, not 0.25. (b) Each stick is the long axis of a sheared image. Around a mass concentration (red dot) lensing lines the images up tangentially, the E pattern. Rotating every stick by 45° gives the B pattern, a swirl, which lensing by a scalar potential cannot make.

The sticks in panel (b) have no arrowheads, and that is the key property of shear. An ellipse rotated by 180° is the same ellipse. What kind of number can describe such a stick?

A stick first. Let the long axis of an ellipse make an angle ϕ with the first reference axis. If we turn the reference axes by an angle α, the same stick is now seen at $\phi - \alpha$. A complex number that records the direction must be a function $f(\phi)$, and the simplest functions of an angle are the pure tones $e^{im\phi}$ with a whole number m. Which m will do? The stick at $\phi + \pi$ is the same stick, so we need $e^{im(\phi+\pi)} = e^{im\phi}$, that is $e^{im\pi} = 1$, and m must be even. The choice $m = 1$ fails: $e^{i(\phi+\pi)} = -e^{i\phi}$, so the same stick would get two opposite values. That is the description of an arrow, a vector, which returns to itself only after a full turn of 2π. The choice $m = 0$ is allowed but records no direction at all. The smallest useful choice is $m = 2$, so a stretch of size $|\gamma|$ along the angle ϕ is written $\gamma = |\gamma|\, e^{2i\phi}$, the form in the table of Section T1.6. When the axes turn by α,

$$\gamma = |\gamma|\, e^{2i\phi} \;\longrightarrow\; |\gamma|\, e^{2i(\phi-\alpha)} = e^{-2i\alpha}\gamma . \tag{T1.5}$$

A quantity that picks up the phase $e^{-is\alpha}$ under a turn of the axes is said to have **spin** s. Shear has spin 2. The CMB polarization $Q + iU$ is also a headless stick and changes in exactly the same way (see the cosmology notes). Every tool built for CMB polarization (the split into E and B modes, spin-2 spherical harmonics, `healpy`'s polarization transforms) uses only the rule (T1.5), so it applies to shear without change.

Where the two angles live. Every lensing formula below is written in two angles, θ_1 and θ_2. They locate a point on the sky the way x and y locate a point on a sheet of graph paper held at arm's length. The patch we analyse is a few degrees across at most, so we may treat it as flat: a plane facing us, with its origin on one chosen line of sight (in Figure T1.7, the one through a lump of matter), and θ_1, θ_2 the two angles, measured from that line, to a point on it. Now follow the light of one galaxy. Without the lump, the galaxy would be seen in the direction $\boldsymbol{\theta}_S$ where it really is. The light that actually reaches us left the galaxy in a slightly different direction, was bent as it passed the lump, and arrives along a straight final leg. We place every object where its light appears to come from, so we extend that last leg backwards, and it meets the sky plane at a different point, $\boldsymbol{\theta}$. The catalogue records $\boldsymbol{\theta}$; the true $\boldsymbol{\theta}_S$ is never seen. The difference is the deflection, $\boldsymbol{\theta}_S = \boldsymbol{\theta} - \hat{\boldsymbol{\alpha}}(\boldsymbol{\theta})$, the second row of the table in Section T1.6.

The deflection by itself changes nothing we can measure. If every point of a galaxy moved by the same $\hat{\boldsymbol{\alpha}}$, the image would simply be the galaxy, displaced, and we never know where it would have been. What we can measure is how the deflection *changes* across one image: points on the side nearer the lump are pulled more than points on the far side, so the image is enlarged (κ) and stretched (γ). Panel (b) of Figure T1.7 shows the outcome for the lump of the figure: the image is pushed outward and stretched along the circle around the lump, the tangential pattern of Figure T1.6(b).

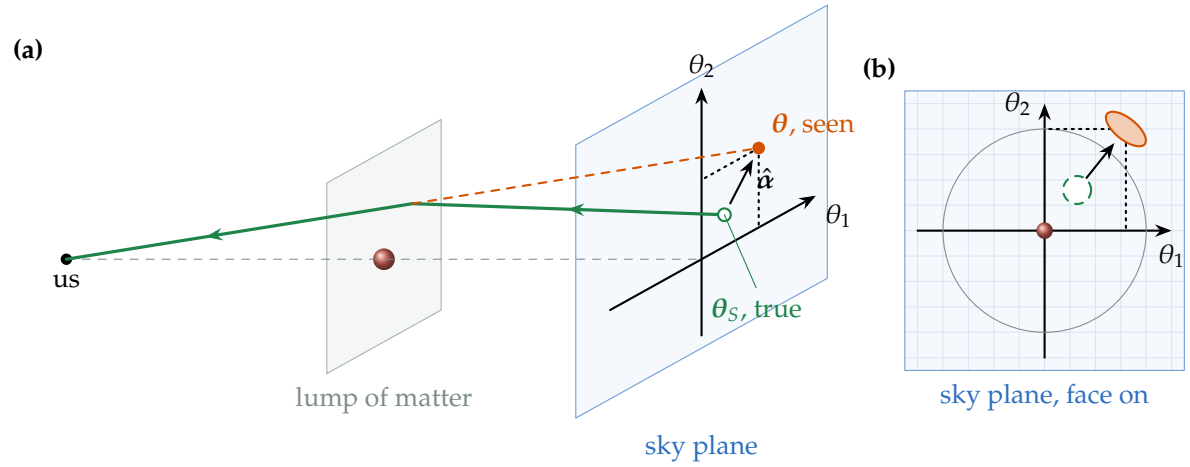


Figure T1.7: What the two angles θ_1, θ_2 are. (a) A patch of sky, treated as a flat plane facing us, with its origin on the line of sight through a lump of matter (dashed grey). A galaxy sits at the true position $\boldsymbol{\theta}_S$ (open green circle). Its light (solid green) is bent where it passes the lump, so it reaches us along a final leg that, extended backwards (dashed orange), meets the sky plane at $\boldsymbol{\theta}$ (orange dot): this is where the galaxy is seen and catalogued, and the dotted lines read off its two coordinates. The black arrow, from the true position to the seen one, is the deflection $\hat{\boldsymbol{\alpha}} = \boldsymbol{\theta} - \boldsymbol{\theta}_S$. (b) The same patch face on. The round galaxy (dashed green) is seen further from the lump and stretched along the grey circle around it, tangentially. The deflection is exaggerated enormously; the lens drawn halfway is for clarity, any lump between us and the galaxy acts the same way.

The pattern around a lump. Does the formula of the table, $\gamma_1 = \frac{1}{2}(\psi_{,11} - \psi_{,22})$, $\gamma_2 = \psi_{,12}$, really give the tangential sticks of panel (b)? Take the simplest lens, a point mass at the origin, whose potential on the sky is $\psi = A \ln \theta$ with $A > 0$ and $\theta = \sqrt{\theta_1^2 + \theta_2^2}$. On the first axis, at $(\theta_1, 0)$

with $\theta_1 > 0$,

$$\psi_{,11} \overset{(1)}{=} A\,\frac{\theta_2^2-\theta_1^2}{\theta^4} = -\frac{A}{\theta_1^2}, \qquad \psi_{,22} \overset{(1)}{=} A\,\frac{\theta_1^2-\theta_2^2}{\theta^4} = +\frac{A}{\theta_1^2}, \qquad \psi_{,12} \overset{(1)}{=} -\frac{2A\theta_1\theta_2}{\theta^4} = 0,$$
$$\gamma_1 \overset{(2)}{=} -\frac{A}{\theta_1^2} < 0, \qquad \gamma_2 \overset{(2)}{=} 0. \tag{T1.6}$$

(1) Differentiate $A \ln\theta = \frac{A}{2}\ln(\theta_1^2+\theta_2^2)$ twice, then set $\theta_2 = 0$.
(2) The shear components of the table.

In panel (a) a positive γ_1 stretches the circle along the first axis, so a negative γ_1 stretches it along the second: on the first axis the stick stands vertical, at right angles to the line to the mass, which is tangential. Every other point follows from the rule (T1.5): turning the axes so that the first one points at the new position leaves the same calculation. In complex form the result is $\gamma = -(A/\theta^2)e^{2i\varphi} = (A/\theta^2)e^{2i(\varphi+\pi/2)}$ at polar angle φ, a stick at angle $\varphi + 90°$, always tangential.

Why lensing makes no B. The E and B rows of the second table of Section T1.6 ($\Delta = \nabla^2$ is the angular Laplacian) split a shear field into the part whose sticks line up tangentially or radially about each point (rotation 0) and the part whose sticks sit at 45° (the swirl). Insert the lensing shear into the B row:

$$\Delta B \overset{(1)}{=} -2\partial_1\partial_2\,\tfrac{1}{2}(\psi_{,11}-\psi_{,22}) + (\partial_1^2-\partial_2^2)\,\psi_{,12} \overset{(2)}{=} -\psi_{,1112}+\psi_{,1222}+\psi_{,1112}-\psi_{,1222} = 0,$$
$$\Delta E \overset{(3)}{=} \tfrac{1}{2}(\psi_{,1111}-2\psi_{,1122}+\psi_{,2222}) + 2\psi_{,1122} = \tfrac{1}{2}\Delta(\psi_{,11}+\psi_{,22}) \overset{(4)}{=} \Delta\kappa. \tag{T1.7}$$

(1) The B row of the table with $\gamma_1 = \frac{1}{2}(\psi_{,11}-\psi_{,22})$ and $\gamma_2 = \psi_{,12}$.
(2) Partial derivatives can be taken in any order, so the four fourth derivatives cancel in pairs.
(3) The E row with the same substitution, multiplied out.
(4) $\kappa = \frac{1}{2}\nabla^2\psi$, from the first table.

A B pattern needs a curl, and a shear built from the second derivatives of one scalar potential has none: lensing makes only E, and its E is κ. A B mode in the data therefore measures what went wrong, not what the cosmos did: residual errors in the correction for the telescope's point-spread function, intrinsic alignments, or approximations in the theory, and the astrophysical ones among these are at the per-cent level of the E signal [44, §3.6], so a larger B points to an error in the processing. Measuring B and finding it consistent with zero is the standard *null test* of a weak-lensing analysis, the lensing counterpart of the hypothesis tests of Chapter 4.

T1.6 The objects a weak-lensing analysis uses

A weak-lensing analysis is a chain of objects, each defined from the one before (Figure T1.8). Light from a distant galaxy crosses the matter between us and it, and the matter bends the ray. That bending is described first at one point of the sky, by a single scalar potential and its derivatives. Next the three-dimensional matter field is projected along the line of sight, which gives the statistics of the sky that a model predicts. Last come the galaxy shapes, the noise they carry and the nuisances that stand between them and the model. For each object the question comes first, then the equation it starts from and the working equation of the book, then a size and the chapters that use it. The physics between the two equations is in the cosmology notes.

The lens map at one point. The question is how the image of a source moves and deforms when matter lies in front of it. Work on a small patch of sky, treated as flat. The observed direction is $\boldsymbol{\theta} = (\theta_1, \theta_2)$, and $\boldsymbol{\theta}_S$ is the direction the source would have without lensing

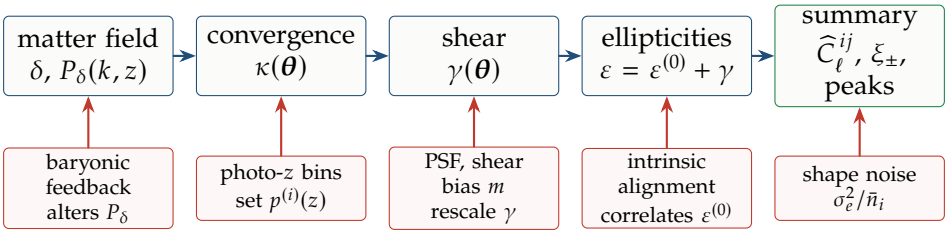


Figure T1.8: The weak-lensing chain. The matter field is projected into a convergence map, whose tidal part is the shear, which is added to the intrinsic shapes of the galaxies; the summary statistics are computed from those shapes. Each red box is a nuisance or noise source that enters at the object above it and has to be modelled and marginalised, or, for shape noise, added to the covariance.

(Figure T1.7). A weak gravitational field perturbs the metric by the Newtonian potential Φ (equal to the second potential Ψ when anisotropic stress is negligible), and a light ray in it is bent at each point by the transverse gradient of Φ. Adding up the bends along the line of sight, each weighted by the geometry of the line of sight, gives one scalar, the **lensing potential**:

$$\psi(\boldsymbol{\theta},\chi) = \frac{2}{c^2}\int_0^\chi \mathrm{d}\chi'\,\frac{f_K(\chi-\chi')}{f_K(\chi)f_K(\chi')}\,\Phi\big[f_K(\chi')\boldsymbol{\theta},\chi'\big].$$

Here χ is comoving distance and $f_K(\chi)$ the curvature function ($f_K=\chi$ in a flat universe). Every lensing quantity is a derivative of ψ. The first derivative is the deflection, and it maps the observed position back to the true one:

$$\hat{\boldsymbol{\alpha}} = \boldsymbol{\nabla}\psi, \qquad \boldsymbol{\theta}_S = \boldsymbol{\theta} - \boldsymbol{\nabla}\psi(\boldsymbol{\theta}).$$

The second derivative says how a small image is deformed. A comma denotes an angular derivative, $\psi_{,ab} \equiv \partial^2\psi/\partial\theta_a\partial\theta_b$, and the Jacobian of the map from image to source is the distortion matrix, whose entries are named after what they do:

$$\mathcal{A}_{ab} \equiv \frac{\partial\theta_{S,a}}{\partial\theta_b} = \delta_{ab} - \psi_{,ab} = \begin{pmatrix} 1-\kappa-\gamma_1 & -\gamma_2 \\ -\gamma_2 & 1-\kappa+\gamma_1 \end{pmatrix},$$
$$\kappa = \tfrac{1}{2}\nabla^2\psi, \qquad \gamma_1 = \tfrac{1}{2}(\psi_{,11}-\psi_{,22}), \qquad \gamma_2 = \psi_{,12}.$$

Here ∇^2 is the angular Laplacian. The trace part is the **convergence** κ, the projected density (for a thin lens, $\kappa = \Sigma/\Sigma_{\rm crit}$), which magnifies the image isotropically. The traceless part is the **shear** $\gamma \equiv \gamma_1 + i\gamma_2 = |\gamma|\,e^{2i\phi}$, a tidal stretch with direction ϕ and a spin of two. A galaxy shape responds not to γ but to the reduced shear $\gamma/(1-\kappa) \approx \gamma$ for $\kappa \ll 1$, because κ rescales the image and leaves only the ratio visible. All of κ, γ are of order 10^{-2} for cosmic shear, which is why each galaxy is a weak and noisy measurement and a survey needs millions. The deflection itself is a few arcminutes. The potential and its derivatives are used in Chapters 7 and 12 and in Chapter T3.

object	physical meaning	key equation (cosmology notes)	used in
lensing potential ψ	the Newtonian potential projected along the line of sight with a geometric weight; every lensing quantity is a derivative of it	$\psi(\boldsymbol{\theta},\chi) = \frac{2}{c^2}\int_0^{\chi} \mathrm{d}\chi' \, \frac{f_K(\chi-\chi')}{f_K(\chi) f_K(\chi')}\, \Phi[f_K(\chi')\boldsymbol{\theta},\chi']$	Ch. 12, T3
deflection	how far the image is moved from where the source really is	$\hat{\boldsymbol{\alpha}} = \boldsymbol{\nabla}\psi, \quad \boldsymbol{\theta}_S = \boldsymbol{\theta} - \boldsymbol{\nabla}\psi(\boldsymbol{\theta})$	Ch. 12, T3
distortion matrix	how a small image is mapped back to its source	$\mathcal{A}_{ab} \equiv \frac{\partial\theta_{S,a}}{\partial\theta_b} = \delta_{ab} - \psi_{,ab} = \begin{pmatrix} 1-\kappa-\gamma_1 & -\gamma_2 \\ -\gamma_2 & 1-\kappa+\gamma_1 \end{pmatrix}$	Ch. 12
convergence κ	projected density; isotropic magnification of the image	$\kappa = \frac{1}{2}\nabla^2\psi$ (thin lens: $\kappa = \Sigma/\Sigma_{\mathrm{crit}}$)	Ch. 7, 12
shear γ	tidal stretch of the image; spin-2	$\gamma_1 = \frac{1}{2}(\psi_{,11} - \psi_{,22}), \quad \gamma_2 = \psi_{,12}, \quad \gamma \equiv \gamma_1 + i\gamma_2 = \lvert\gamma\rvert\, e^{2i\phi}$	Ch. 12
reduced shear	what a galaxy shape actually responds to, since κ rescales the image	$\frac{\gamma}{1-\kappa} \approx \gamma$ for $\kappa \ll 1$	Ch. 12

Three remarks on this table. The cosmology notes write the potential with angular derivatives as $\tilde{\psi}$, and the thin-lens and CMB versions as ψ; for the data analyst all are the same object, and we write ψ. The off-diagonal entries of $\mathcal{A}$ are $-\gamma_2$, because $\mathcal{A}_{12} = -\psi_{,12}$ and $\gamma_2 = \psi_{,12}$; a matrix printed with $+\gamma_2$ uses the opposite sign of γ_2. And the reduced shear has no letter in the table on purpose. The literature calls it g, but $g(\chi)$ is already the lensing efficiency in the next table; we use the letter only in Section T1.6.1, where the reduced shear is derived.

From the matter field to the statistics of the sky. The question is how the three-dimensional matter field becomes a map and a power spectrum on the sky. The sources have a redshift distribution $p_z(z)$, normalised to one, and $p_\chi(\chi)$ is the same distribution in comoving distance, $p_\chi\,\mathrm{d}\chi = p_z\,\mathrm{d}z$. The lensing potential above is for one source distance. Averaging over the source sample, with χ_H the comoving horizon, the weight of matter at distance χ is the **lensing efficiency**

$$g(\chi) \equiv \int_{\chi}^{\chi_H} p_\chi(\chi') \, \frac{f_K(\chi'-\chi)}{f_K(\chi')}\, \mathrm{d}\chi'.$$

It is zero at the observer, where there is no lever arm, and zero behind the sources, which cannot be lensed by matter behind them, so it peaks midway. Poisson's equation turns the potential into the matter density contrast δ, and the second derivative of ψ gives the convergence as a weighted projection of δ, with $a(\chi)$ the scale factor at distance χ:

$$\kappa(\boldsymbol{\theta}) = \frac{3}{2}\frac{H_0^2}{c^2}\Omega_{m0} \int_0^{\chi_H} g(\chi) f_K(\chi)\, \frac{\delta[f_K(\chi)\boldsymbol{\theta},\chi]}{a(\chi)}\, \mathrm{d}\chi.$$

The prediction the data are compared with is the variance of κ per angular mode, the convergence power spectrum, with $P_\delta(k,\chi)$ the matter power spectrum of Section T1.2 (non-linear in practice) and the index i labelling a redshift bin:

$$P_\kappa(\ell) = \frac{9H_0^4}{4c^4}\Omega_{m0}^2 \int_0^{\chi_H} \left[\frac{g(\chi)}{a(\chi)}\right]^2 P_\delta\!\left(\frac{\ell}{f_K(\chi)}, \chi\right) \mathrm{d}\chi, \qquad g^2 \to g_i g_j \text{ for bins.}$$

This is the Limber approximation, and it is built below from the independent-slices picture. Since Ω_{m0} and the amplitude of P_δ (σ_8) multiply it, cosmic shear measures them together. The shear has the same spectrum, $P_\gamma = P_\kappa$, and surveys often measure it in real space through the correlations $\xi_\pm(\vartheta) = \int_0^\infty \frac{\mathrm{d}\ell}{2\pi}\,\ell P_\kappa(\ell)\,J_{0,4}(\ell\vartheta)$ (J_0 for +, J_4 for −, with J_n the Bessel functions). To go from the shear map back to a mass map one uses the Kaiser–Squires inversion [46], in the two-dimensional Fourier transforms $\hat\gamma(\boldsymbol{\ell})$ and $\hat\kappa(\boldsymbol{\ell})$ with $\boldsymbol{\ell} = (\ell_1, \ell_2)$ the angular wavevector:

$$\hat\kappa(\boldsymbol{\ell}) = \tfrac{1}{\pi}\hat{\mathcal{D}}^*(\boldsymbol{\ell})\hat\gamma(\boldsymbol{\ell}), \qquad \hat{\mathcal{D}}(\boldsymbol{\ell}) = \pi\frac{\ell_1^2 - \ell_2^2 + 2i\ell_1\ell_2}{\ell^2}.$$

A spin-2 field splits into a curl-free part E and a curl part B by $\Delta E = (\partial_1^2 - \partial_2^2)\gamma_1 + 2\partial_1\partial_2\gamma_2$ and $\Delta B = -2\partial_1\partial_2\gamma_1 + (\partial_1^2 - \partial_2^2)\gamma_2$, where Δ is the angular Laplacian. Lensing gives $E = \kappa$ and $B = 0$, so B is a null test. The CMB is lensed in the same way, with a single source plane at last scattering, χ_*, so that $g(\chi) = f_K(\chi_* - \chi)/f_K(\chi_*)$ and the remapping is $\hat T[\boldsymbol{n}] = T[\boldsymbol{n} + \boldsymbol{\xi}(\boldsymbol{n})]$ with $\boldsymbol{\xi} = \boldsymbol{\nabla}\psi$ (Section T1.2). Typical sizes: $g \sim 1$ at its peak, a convergence rms of about 1%, and ℓ from tens to a few thousand. Used in Chapters T3, 7, 8, 10 and 12.

object	physical meaning	key equation (cosmology notes)	used in
lensing efficiency	how strongly matter at distance χ lenses the source sample: zero at the observer and behind the sources	$g(\chi) \equiv \int_\chi^{\chi_H} p_\chi(\chi')\,\frac{f_K(\chi'-\chi)}{f_K(\chi')}\,\mathrm{d}\chi'$	Ch. 10, 12
convergence field	the matter contrast projected with the lensing weight	$\kappa(\boldsymbol{\theta}) = \frac{3}{2}\frac{H_0^2}{c^2}\Omega_{m0}\int_0^{\chi_H} g(\chi) f_K(\chi)\,\frac{\delta[f_K(\chi)\boldsymbol{\theta},\chi]}{a(\chi)}\,\mathrm{d}\chi$	Ch. 7, 12
Limber $C_\ell^{\kappa\kappa}$	the variance of κ per angular mode, the prediction the data are compared with	$P_\kappa(\ell) = \frac{9H_0^4}{4c^4}\Omega_{m0}^2\int_0^{\chi_H}\left[\frac{g(\chi)}{a(\chi)}\right]^2 P_\delta\Big(\frac{\ell}{f_K(\chi)},\chi\Big)\mathrm{d}\chi$; bins i,j: $g^2 \to g_i g_j$; full sky: $C_\ell^{\kappa\kappa} = \frac{1}{4}\ell^2(\ell+1)^2 C_\ell^{\psi\psi} \to P_\kappa(\ell)$	Ch. 7, 10, 12
shear two-point	shear has the same spectrum as κ; its real-space correlations are what surveys often measure	$P_\gamma(\ell) = P_\kappa(\ell)$, $\xi_\pm(\vartheta) = \int_0^\infty \frac{\mathrm{d}\ell}{2\pi}\,\ell P_\kappa(\ell)\,J_{0,4}(\ell\vartheta)$ (J_0 for +, J_4 for −)	Ch. 12
Kaiser–Squires	a mass map from the shear field [46], up to a constant sheet	$\hat\kappa(\boldsymbol{\ell}) = \frac{1}{\pi}\hat{\mathcal{D}}^*(\boldsymbol{\ell})\hat\gamma(\boldsymbol{\ell})$, $\hat{\mathcal{D}}(\boldsymbol{\ell}) = \pi\frac{\ell_1^2 - \ell_2^2 + 2i\ell_1\ell_2}{\ell^2}$	Ch. 12
E and B modes	curl-free and curl parts of a spin-2 field; lensing gives $E = \kappa$, $B = 0$	$\Delta E = (\partial_1^2 - \partial_2^2)\gamma_1 + 2\partial_1\partial_2\gamma_2$, $\Delta B = -2\partial_1\partial_2\gamma_1 + (\partial_1^2 - \partial_2^2)\gamma_2$	Ch. 12, T3
CMB lensing	the CMB is lensed too, with a single source plane at last scattering, χ_*	$\hat T[\boldsymbol{n}] = T[\boldsymbol{n} + \boldsymbol{\xi}(\boldsymbol{n})]$, $\boldsymbol{\xi} = \boldsymbol{\nabla}\psi$, $g(\chi) = f_K(\chi_* - \chi)/f_K(\chi_*)$	Ch. 8, T3

The Limber row is the one the later chapters lean on, so it is worth reading in words. The convergence spectrum at multipole ℓ is a sum over distances χ of the matter power at the wavenumber $k = \ell/f_K(\chi)$ that projects onto that angle, weighted by g^2/a^2. Two pieces of it need a reason: why that wavenumber, and why the weight is squared. The full calculation is in the cosmology notes; the two ideas behind it are short.

One angle picks one wavenumber at each distance. A wave in the matter of wavelength $2\pi/k$, at comoving distance χ, spans an angle $2\pi/[k f_K(\chi)]$ on the sky, since a transverse comoving

length L at that distance subtends $L/f_K(\chi)$. A pattern on the sky with multipole ℓ repeats every $2\pi/\ell$ radians (the flat-sky form of (7.43)). The two match when $2\pi/\ell = 2\pi/[k f_K(\chi)]$, that is $k = \ell/f_K(\chi)$. The same ℓ therefore probes large wavenumbers in nearby matter and small wavenumbers in distant matter, and a single multipole mixes a whole range of k.

A sum of independent slices. Cut the line of sight into slices of thickness $\Delta\chi$, thick compared with the wavelengths of interest, and let δ_s be the density contrast averaged through slice s. The convergence row of the table is then a weighted sum, $\kappa = \sum_s w_s \delta_s \Delta\chi$ with $w_s = \frac{3}{2}(H_0/c)^2 \Omega_{m0}\, g f_K/a$ evaluated at slice s. Matter in two thick slices is nearly uncorrelated, because the waves that could connect them are averaged away inside each slice; assuming exactly zero correlation is the **Limber approximation**. For a sum of independent terms the variances add, each multiplied by the square of its weight, and the same holds mode by mode for the power:

$$P_\kappa(\ell) \overset{(1)}{=} \sum_s w_s^2\, \Delta\chi^2\, P_s(\ell) \overset{(2)}{=} \sum_s w_s^2\, \Delta\chi^2\, \frac{P_\delta\big(\ell/f_K, \chi_s\big)}{f_K^2\, \Delta\chi} \overset{(3)}{=} \frac{9H_0^4}{4c^4}\, \Omega_{m0}^2 \int \left[\frac{g}{a}\right]^2 P_\delta\Big(\frac{\ell}{f_K}, \chi\Big) \mathrm{d}\chi. \quad \text{(T1.8)}$$

(1) $\mathrm{Var}(\sum_s c_s X_s) = \sum_s c_s^2 \,\mathrm{Var}\, X_s$ for independent X_s; $P_s(\ell)$ is the angular power of slice s.
(2) The angular power of one thick slice: the three-dimensional power at the matching wavenumber, divided by $\Delta\chi$ because the averaging through the slice keeps only one wave in $\Delta\chi$ along the line of sight, and by f_K^2 to turn transverse comoving area into solid angle (see the cosmology notes).
(3) Insert w_s; the factor f_K^2 cancels, and the sum over thin slices becomes an integral.

This is the Limber row. The weight is squared because a power is the average of a squared amplitude, and the slices enter one at a time because they are independent.

Three things now follow. The prefactor carries Ω_{m0}^2, and in linear theory P_δ is proportional to σ_8^2, so for a fixed shape of P_δ

$$C_\ell^{\kappa\kappa} \propto \Omega_{m0}^2\, \sigma_8^2, \quad \text{(T1.9)}$$

and the two parameters enter together (Section T1.7 finds out how far this holds once the shape of P_δ is allowed to change). Each multipole mixes many wavenumbers, so a sharp feature in P_δ is smoothed away in $C_\ell^{\kappa\kappa}$. And the weight g is broad, so lensing measures a smeared, projected density and not a three-dimensional map. Splitting the sources into redshift bins (**tomography**) recovers some of the lost depth, through the cross spectra C_ℓ^{ij} with weights $g_i g_j$.

The last row is the CMB version. Lensing by the same structures remaps the CMB by deflections of a few arcminutes (see the cosmology notes). How can a map of the potential be recovered from a single lensed CMB sky, with no galaxy shapes at all? *What does lensing do to the statistics of the CMB that the unlensed sky could never do?*

Unlensed modes are independent. Without lensing the CMB is statistically the same in every direction, and its Fourier modes on a flat patch are uncorrelated: $\langle T_\ell T_{\ell'}\rangle = 0$ unless $\boldsymbol{\ell}' = -\boldsymbol{\ell}$ (Section T1.1; the flat-sky version of the independence of the $a_{\ell m}$). *Lensing mixes them.* Hold one lensing potential fixed and expand the remapping of the table to first order in the small deflection:

$$\begin{aligned} \hat{T}(\boldsymbol{n}) &\overset{(1)}{=} T(\boldsymbol{n}) + \boldsymbol{\nabla}\psi \cdot \boldsymbol{\nabla} T, \\ \hat{T}_{\boldsymbol{\ell}} &\overset{(2)}{=} T_{\boldsymbol{\ell}} - \int \frac{\mathrm{d}^2\boldsymbol{\ell}_1}{(2\pi)^2} \left[(\boldsymbol{\ell} - \boldsymbol{\ell}_1)\cdot\boldsymbol{\ell}_1\right] \psi_{\boldsymbol{\ell}-\boldsymbol{\ell}_1} T_{\boldsymbol{\ell}_1}, \\ \big\langle \hat{T}_{\boldsymbol{\ell}} \hat{T}_{\boldsymbol{\ell}'} \big\rangle_{\mathrm{CMB}} &\overset{(3)}{=} \left[(\boldsymbol{L}\cdot\boldsymbol{\ell})\, C_\ell + (\boldsymbol{L}\cdot\boldsymbol{\ell}')\, C_{\ell'}\right] \psi_{\boldsymbol{L}}, \qquad \boldsymbol{L} \equiv \boldsymbol{\ell} + \boldsymbol{\ell}' \neq \boldsymbol{0}. \end{aligned} \quad \text{(T1.10)}$$

(1) Taylor expansion of $T(\boldsymbol{n} + \boldsymbol{\xi})$ with $\boldsymbol{\xi} = \boldsymbol{\nabla}\psi$, kept to first order.

(2) Fourier transform: each gradient becomes i times its wavevector, and the transform of a product is the convolution of the transforms, so the potential at wavevector $\boldsymbol{\ell} - \boldsymbol{\ell}_1$ moves CMB power from mode $\boldsymbol{\ell}_1$ into mode $\boldsymbol{\ell}$.
(3) Average over the unlensed CMB with ψ held fixed. The unlensed term alone gives zero for $\boldsymbol{L} \neq \boldsymbol{0}$; in each cross term only the pairing allowed by independence survives, giving the two terms with the unlensed spectrum C_ℓ; the term of second order in ψ is dropped.

So with lensing, two CMB modes whose wavevectors add up to $\boldsymbol{L}$ are correlated, by an amount proportional to $\psi_{\boldsymbol{L}}$ times a weight built from the known CMB spectrum. Multiplying pairs of observed modes, dividing by the weight and averaging over all pairs with the same $\boldsymbol{L}$ gives an estimate of $\psi_{\boldsymbol{L}}$: this is the **quadratic estimator** [47], which uses temperature and polarization in the same way. Its noise comes from the CMB itself, since each product $\hat{T}_{\boldsymbol{\ell}}\hat{T}_{\boldsymbol{\ell}'}$ also contains the chance product of two unlensed modes. How large that noise is, and what else it adds to the power spectrum of the reconstructed map, is worked out in Section T1.9.

Planck reconstructed the lensing potential over the full sky in this way. Its total signal-to-noise ratio, summed over all L, is 40 (the "40σ" of the paper), the same kind of number as the signal-to-noise ratio of a galaxy survey computed in Section T1.8. From it alone Planck measured $\sigma_8\Omega_m^{0.25} = 0.589 \pm 0.020$ [48]. The kernel in the last row, $g(\chi)f_K(\chi) = f_K(\chi)f_K(\chi_* - \chi)/f_K(\chi_*)$, is largest midway to the last-scattering surface in distance; per unit redshift it peaks near $z \approx 2$ (Figure T1.11), far behind the galaxy samples. Section T1.7 finds that deeper samples hold a combination $\sigma_8\Omega_m^{\alpha}$ with a smaller α fixed, and the CMB, the deepest source of all, has $\alpha = 0.25$. A different combination of the same two parameters is what is used to break the degeneracy of the next section.

The data, their noise and the nuisances. The question is what the survey actually records and how it differs from the model's κ. The measured ellipticity of a galaxy, ε, is its own intrinsic shape $\varepsilon^{(0)}$ plus the shear. Since galaxies point in random directions, $\langle\varepsilon^{(0)}\rangle = 0$, and the average over many gives the reduced shear:

$$\varepsilon \approx \varepsilon^{(0)} + \gamma, \qquad \langle\varepsilon\rangle = \frac{\gamma}{1-\kappa} \quad \text{(Section T1.6.1)}.$$

The scatter of $\varepsilon^{(0)}$, with σ_e the rms of one component, acts as white noise added to the convergence spectrum, $\bar{n}_i$ being the number of galaxies per steradian in bin i:

$$N_i = \frac{\sigma_e^2}{\bar{n}_i}, \qquad \langle\widehat{C}_\ell^{ij}\rangle = C_\ell^{ij} + \delta_{ij}N_i.$$

This is the **shape noise**, with $\sigma_e \approx 0.27$. Redshifts of the sources are photometric ones $z_{\rm ph}$ [estimated from a few broad-band colours], with a scatter $p(z_{\rm ph}\,|\,z)$ about the true z, so the true distribution of each tomographic bin has tails: $p^{(i)}(z) \propto p_z(z)\int_{\text{bin }i} p(z_{\rm ph}\,|\,z)\,\mathrm{d}z_{\rm ph}$, and a shift Δz_i of each is a nuisance. The combination of the two parameters that cosmic shear measures best is $S_8 \equiv \sigma_8(\Omega_{m0}/0.3)^{0.5}$, about 0.8. Three more effects are modelled or marginalised. *Intrinsic alignment*: nearby galaxies are stretched by the same tidal field when they form, so $\varepsilon^{(0)}$ is not purely random, and an amplitude $A_{\rm IA}$ is added to C_ℓ^{ij}. *Baryonic feedback*: gas expelled from haloes by supernovae and active nuclei lowers P_δ at $k \gtrsim 1\,h\,\mathrm{Mpc}^{-1}$, which looks like a lower S_8 and is handled by scale cuts or a feedback parameter. *Shear bias*: an imperfect shape measurement rescales the shear, $\varepsilon \to (1+m)\gamma + \ldots$, with m calibrated on image simulations. Used in Chapters 9, 10 and 12.

object	physical meaning	key equation	used in
ellipticity	one galaxy's measured shape: its own shape plus the shear	$\varepsilon \approx \varepsilon^{(0)} + \gamma, \quad \langle\varepsilon^{(0)}\rangle = 0 \Rightarrow \langle\varepsilon\rangle = \dfrac{\gamma}{1-\kappa}$ (Section T1.6.1)	Ch. 12
shape noise	the intrinsic shapes act as white noise added to the convergence spectrum	$N_i = \dfrac{\sigma_e^2}{\bar{n}_i}, \quad \langle\widehat{C}_\ell^{ij}\rangle = C_\ell^{ij} + \delta_{ij}N_i$	Ch. 10, 12
photo-z bins	sources sorted by estimated redshift; each bin's true-z distribution has tails	$p^{(i)}(z) \propto p_z(z)\int_{\text{bin } i} p(z_{\text{ph}} \mid z)\,\mathrm{d}z_{\text{ph}}$; nuisance: a shift Δz_i of each	Ch. 10, 12
S_8	the amplitude combination cosmic shear measures best	$S_8 \equiv \sigma_8\,(\Omega_{m0}/0.3)^{0.5}$	Ch. 9–12
intrinsic alignment	galaxies near each other are stretched by the same tidal field when they form, so $\varepsilon^{(0)}$ is not purely random	adds terms to C_ℓ^{ij} that mimic or cancel part of the lensing signal; an amplitude A_{IA} is marginalised	Ch. 10, 12
baryonic feedback	gas blown out of haloes by supernovae and active nuclei lowers P_δ at $k \gtrsim 1\,h\,\text{Mpc}^{-1}$	lowers $C_\ell^{\kappa\kappa}$ at high ℓ, like a lower S_8; handled by scale cuts or a feedback parameter	Ch. 10, 12
shear bias	imperfect shape measurement rescales the shear	$\varepsilon \to (1+m)\gamma + \ldots$, with m calibrated on image simulations	Ch. 12

Why is shape noise flat in ℓ, and why is it $\sigma_e^2/\bar{n}$? Here σ_e is the rms of one real component of $\varepsilon^{(0)}$, the first symbol of the data table and the 0.27 of Section T1.5, and $\bar{n}$ is the number of galaxies per steradian. Think of a pixel of solid angle A. It holds on average $\bar{n}A$ galaxies; the actual count scatters about that from pixel to pixel (Poisson scatter), but when $\bar{n}A$ is large this barely changes the average noise, and we ignore it. Then

$$\mathrm{Var}\big(\bar{\varepsilon}_1 \text{ in one pixel}\big) \overset{(1)}{=} \frac{\sigma_e^2}{\bar{n}A}, \qquad N \overset{(2)}{=} \mathrm{Var}\big(\bar{\varepsilon}_1 \text{ in one pixel}\big) \times A \overset{(3)}{=} \frac{\sigma_e^2}{\bar{n}}. \tag{T1.11}$$

(1) The variance of a mean, (3.20), for $\bar{n}A$ independent galaxies; the shear is the same for all of them and adds no scatter.

(2) Different pixels hold different galaxies, so the noise in one pixel is independent of the noise in every other one. A field of independent pixels has the same power at every ℓ, which is white noise, and its power is the pixel variance times the pixel area, (8.16): the discrete harmonic sum spreads the variance of each pixel evenly over all the modes the pixels can carry.

(3) The pixel area cancels, as it must: the noise power cannot depend on how we chose to pixelise the sky.

This is the same white-noise reasoning that gives the instrumental noise of a CMB map in Chapter 8. Is this the noise on κ or on γ? Both. The Kaiser–Squires row turns $\hat{\gamma}$ into $\hat{\kappa}$ with the factor $\hat{\mathcal{D}}^*/\pi$, and $\left|\hat{\mathcal{D}}\right|^2 = \pi^2$, since $\left|\ell_1^2 - \ell_2^2 + 2i\ell_1\ell_2\right| = \left|(\ell_1 + i\ell_2)^2\right| = \ell^2$. So the inversion keeps the size of every mode. The two shear components carry noise $\sigma_e^2/\bar{n}$ each, $2\sigma_e^2/\bar{n} = \sigma_\varepsilon^2/\bar{n}$ in all, and this splits equally into the E part, which is κ, and the B part; the κ noise power is $\sigma_e^2/\bar{n} = \sigma_\varepsilon^2/(2\bar{n})$, the relation of convention 2 of Section T1.6.2.

The size, for the survey of Section T1.8: 1.24 galaxies per arcmin^2 per bin (a fifth of 6.2), and

one steradian holds $(180\times 60/\pi)^2 = (10800/\pi)^2 = 1.18\times 10^7$ arcmin^2, so

$$\bar{n}_i = 1.24\times 1.18\times 10^7\,\mathrm{sr}^{-1} = 1.47\times 10^7\,\mathrm{sr}^{-1}, \qquad N_i = \frac{0.27^2}{1.47\times 10^7} = \frac{0.073}{1.47\times 10^7} \approx 4.965\times 10^{-9},$$

which the signal of the nearest bin exceeds only up to $\ell \approx 21$ (Figure T1.12).

T1.6.1 Reduced shear: what a galaxy shape responds to

The data table says that the mean ellipticity is $\gamma/(1-\kappa)$, while (T1.2) says γ. Which is it, and why should the convergence enter the shape at all, when panel (a) of Figure T1.6 shows it only enlarging the image? The answer is in the distortion matrix of the first table. Pull the factor $1-\kappa$ out of it:

$$\mathcal{A} \overset{(1)}{=} (1-\kappa)\begin{pmatrix} 1-\dfrac{\gamma_1}{1-\kappa} & -\dfrac{\gamma_2}{1-\kappa} \\ -\dfrac{\gamma_2}{1-\kappa} & 1+\dfrac{\gamma_1}{1-\kappa}\end{pmatrix} \overset{(2)}{=} (1-\kappa)\begin{pmatrix} 1-g_1 & -g_2 \\ -g_2 & 1+g_1\end{pmatrix}, \qquad g \equiv g_1 + i g_2 \equiv \frac{\gamma}{1-\kappa}. \tag{T1.12}$$

(1) Divide every entry of $\mathcal{A}$ by $1-\kappa$ and multiply the whole matrix by it.
(2) Give the ratios a name. The literature's letter for the **reduced shear** is g; in this subsection only, g without an argument is the reduced shear, while $g(\chi)$ elsewhere remains the lensing efficiency.

The simplest source first. Take a round source of radius r, and turn the axes so that the stretch lies along the first one, $g = g_1 > 0$ (any other direction follows by the rule (T1.5)). $\mathcal{A}$ maps the image back to the source, so the image is the set of directions $\boldsymbol{\theta}$ with $|\mathcal{A}\boldsymbol{\theta}| = r$:

$$(1-\kappa)^2\left[(1-g)^2\theta_1^2 + (1+g)^2\theta_2^2\right] = r^2 \overset{(1)}{=} \text{an ellipse with } a = \frac{r}{(1-\kappa)(1-g)},\ b = \frac{r}{(1-\kappa)(1+g)},$$
$$\varepsilon = \frac{a-b}{a+b} \overset{(2)}{=} \frac{(1+g)-(1-g)}{(1+g)+(1-g)} = g. \tag{T1.13}$$

(1) $|\mathcal{A}\boldsymbol{\theta}|^2$ with the diagonal matrix of (T1.12); the semi-axes are read off where $\theta_2 = 0$ and where $\theta_1 = 0$.
(2) The ellipticity definition of convention 3 of Section T1.6.2; multiply the numerator and the denominator by $(1-\kappa)(1-g)(1+g)/r$, and κ cancels.

So κ only rescales the size, as in panel (a), and the shape responds to g, not to γ. For a source that is itself elliptical, the same calculation kept to first order in small quantities gives $\varepsilon \approx \varepsilon^{(0)} + g$, and averaging with $\langle\varepsilon^{(0)}\rangle = 0$ gives $\langle\varepsilon\rangle = g$, the table's entry. How different are g and γ? Expanding $1/(1-\kappa) = 1+\kappa+\kappa^2+\ldots$ gives $g = \gamma(1+\kappa+\ldots)$. With $\kappa \approx 0.01$ the two differ by about one per cent of a one-per-cent effect, which is why the rest of the part writes γ; analyses at per-cent precision model g.

T1.6.2 Conventions to carry forward

As with the CMB conventions of Section T1.2.1, each of these changes a number by a fixed factor without any warning.

1. **CAMB outputs.** With a lensing source window (`SplinedSourceWindow` from `camb.sources`, called with `source_type="lensing"`) CAMB computes the Limber integral above for any $p_z(z)$, and `get_source_cls_dict(raw_cl=True)` returns $C_\ell^{\kappa\kappa}$ under the key `"W1xW1"` (and

C_ℓ^{ij} under `"WixWj"`). Without `raw_cl=True` it returns $\ell(\ell+1)C_\ell/2\pi$, which at $\ell = 1000$ is the raw C_ℓ times $1000 \times 1001/2\pi = 1.59 \times 10^5$. The key `"PxP"` is the CMB lensing potential, $C_\ell^{\phi\phi}$ in raw form, and CAMB's ϕ is the cosmology notes' ψ (Section T1.2.1, item 6). The factor between the two spectra comes from $\kappa = \frac{1}{2}\nabla^2\psi$ of the first table. On the sphere the Laplacian acting on a spherical harmonic gives $\nabla^2 Y_{\ell m} = -\ell(\ell+1)Y_{\ell m}$ (by (7.39), since $\boldsymbol{L}^2 = -\nabla^2$), so expanding both fields in harmonics gives $\kappa_{\ell m} = -\frac{1}{2}\ell(\ell+1)\psi_{\ell m}$. A power is the mean square of an amplitude, so the factor is squared: $C_\ell^{\kappa\kappa} = \frac{1}{4}[\ell(\ell+1)]^2 C_\ell^{\phi\phi}$.

2. **Shape-noise conventions.** Some papers quote the rms of one component, σ_e; others the rms of the whole complex ellipticity, $\sigma_\varepsilon^2 = \langle|\varepsilon^{(0)}|^2\rangle = 2\sigma_e^2$. The noise power is $\sigma_e^2/\bar{n} = \sigma_\varepsilon^2/(2\bar{n})$, as written in [44, eq. 52].

3. **Ellipticity definitions.** For $\varepsilon = \frac{a-b}{a+b}e^{2i\phi}$ (axes $a \geq b$, position angle ϕ) the response to shear is $\varepsilon \approx \varepsilon^{(0)} + \gamma$. Some codes use $\frac{a^2-b^2}{a^2+b^2}e^{2i\phi}$ instead, which responds as 2γ [44, §3.5].

4. **Flat versus full sky.** $P_\gamma = P_\kappa$ holds on a flat patch. On the sphere the E-mode shear spectrum is $\frac{(\ell+2)(\ell-1)}{\ell(\ell+1)}C_\ell^{\kappa\kappa}$. The reason: κ takes two derivatives of ψ in the combination of the Laplacian, which multiplies $\psi_{\ell m}$ by $\frac{1}{2}\ell(\ell+1)$ (item 1), while the shear takes two derivatives that each raise the spin by one, and on spin-weighted harmonics these multiply by $\sqrt{\ell(\ell+1)}$ and then $\sqrt{(\ell-1)(\ell+2)}$ (see the cosmology notes). The shear amplitude is therefore $\frac{1}{2}\sqrt{(\ell+2)(\ell+1)\ell(\ell-1)}\,\psi_{\ell m}$, and the ratio of the two powers is $(\ell+2)(\ell+1)\ell(\ell-1)/[\ell(\ell+1)]^2 = (\ell+2)(\ell-1)/[\ell(\ell+1)]$. It tends to one as ℓ grows: 0.982 at $\ell = 10$, 0.9998 at $\ell = 100$. Our Limber integral uses $k = (\ell+\frac{1}{2})/\chi$ rather than ℓ/χ. On the sphere the role of ℓ^2 in the flat-sky Laplacian is played by $\ell(\ell+1)$, so the effective flat wavenumber is $\sqrt{\ell(\ell+1)} = \ell + \frac{1}{2} - \frac{1}{8\ell} + \ldots$, and $\ell + \frac{1}{2}$ is the closer match; it is also where the spherical Bessel functions of the exact projection peak. The change in k is a fraction $1/(2\ell)$: 5% at $\ell = 10$, 0.5% at $\ell = 100$, and the spectrum moves by a fraction of the same order, so again only low ℓ is affected.

5. **The sign of γ_2.** It differs between survey catalogues and `healpy`'s polarization convention. The safe check is to build a shear map from a known κ map, invert it, and confirm that $+\kappa$ comes back in E and nothing in B.

Watch out: A factor of two that looks like a result

Mixing σ_e with σ_ε doubles or halves the noise power; mixing the two ellipticity definitions doubles or halves the shear, and so quadruples or quarters $C_\ell^{\kappa\kappa}$ inferred from shapes. Neither error breaks a script. Both shift S_8 by far more than its quoted error.

T1.7 Experiment 1: what cosmic shear measures, S_8

Suppose a survey measures a convergence spectrum that is somewhat higher than our fiducial prediction. What could have produced it? Before any calculation, there are two candidates, and the Limber row of Section T1.6 tells us how each one acts.

1. *More matter*, a larger Ω_{m0}. There are more lenses, and the prefactor Ω_{m0}^2 in the Limber integral grows. With h and the baryon density fixed, a larger Ω_{m0} also changes the shape of P_δ and how fast structure grows.

2. *The same matter, more clumped*, a larger σ_8. The prefactor is unchanged, but P_δ is larger at every k, by σ_8^2 in linear theory and by more on non-linear scales, where clumping feeds on itself.

Both raise $C_\ell^{\kappa\kappa}$. How probable is the observed spectrum under each candidate? If both predict nearly the same spectrum, the observation is about equally probable under both. The data then cannot prefer one over the other, and the posterior keeps both, exactly as in the inference of Section 1.7. So lensing data alone can only fix some combination of the two. The experiment finds that combination, and where the two candidates stop predicting the same thing.

T1.7.1 What we vary

Recall the two ingredients of the Limber row: $P_\delta(k, \chi)$ is the matter power spectrum at wavenumber k and distance χ, and $g(\chi)$ the lensing efficiency of the second table, a weighted count of how much source light lies behind χ. We change only Ω_m and σ_8, and choose five points of the (Ω_m, σ_8) plane, each with a job. Three lie on the line of constant $S_8 = 0.832$, the Planck value [25]. For a chosen Ω_m the definition of S_8 in the data table fixes σ_8:

$$\sigma_8 = S_8\left(\frac{0.3}{\Omega_m}\right)^{0.5}, \qquad \text{e.g.} \quad \Omega_m = 0.25: \quad \sigma_8 = 0.832 \times \sqrt{0.3/0.25} = 0.832 \times 1.095 = 0.911.$$

Two lie across the line, at the fiducial Ω_m with $S_8 = 0.76$ (close to what the lensing surveys find) and $S_8 = 0.90$.

point	Ω_m	σ_8	S_8	role
fiducial	0.3153	0.811	0.832	reference (Planck)
along, low Ω_m	0.25	0.911	0.832	same S_8, less matter, more clumped
along, high Ω_m	0.40	0.720	0.832	same S_8, more matter, less clumped
across, low	0.3153	0.741	0.76	S_8 lowered by 9%
across, high	0.3153	0.878	0.90	S_8 raised by 8%

What should we see? Across the line Ω_m is fixed, so in linear theory $C_\ell \propto \sigma_8^2$ by (T1.9), and a 9% change of S_8 should move the spectrum by about $(0.76/0.832)^2 - 1 = -17\%$ and $(0.90/0.832)^2 - 1 = +17\%$, a little more where non-linear growth adds to it. Along the line the answer is not obvious. If only the prefactor of (T1.9) mattered, a constant spectrum would need $\sigma_8\Omega_m$ fixed, and along the S_8 line, where $\sigma_8\Omega_m^{0.5}$ is fixed, the spectrum would still change by the factor $(0.40/0.25)$ in Ω_m between the two ends. But a larger Ω_m at fixed σ_8 also reshapes P_δ. If S_8 is the right combination, the three spectra on the line nearly coincide.

T1.7.2 What we hold fixed

Everything else stays at the fiducial values of Table T1.1, and all five models use one source sample:

held fixed	value
Hubble parameter h	0.6736
baryon density $\Omega_{b0}h^2$	0.02237
spectral index n_s	0.9649
neutrino mass $\sum m_\nu$	0.06 eV (one massive neutrino)
geometry	flat, $f_K(\chi) = \chi$
source distribution	$p_z \propto z^2 e^{-(z/0.5)^{1.5}}$, median redshift 0.71
non-linear P_δ	halofit [41]

The source distribution is a Stage-III-like sample. Its factor z^2 makes it start at zero, because nearby space holds little volume and so few galaxies; the factor $e^{-(z/0.5)^{1.5}}$ cuts it off where galaxies become too faint for the survey to measure their shapes.

T1.7.3 What we compute, and how

Listing T1.3: Setting (Ω_m, σ_8) and asking CAMB for $C_\ell^{\kappa\kappa}$ (`code/chT1/03_kappa_s8.py`)

```
48 def camb_params(om, s8, windows=True, as_ref=FIDUCIAL["As"]):
49     """CAMB parameters for flat LCDM at (Omega_m, sigma_8), other parameters fiducial.
50     Omega_m includes the massive neutrino; sigma_8 is set by rescaling A_s
51     (linear sigma_8 is proportional to sqrt(A_s))."""
52     import camb
53     from camb.sources import SplinedSourceWindow
54     par = dict(FIDUCIAL)
55     p0 = camb.set_params(**par)
56     par["omch2"] = om * H**2 - par["ombh2"] - p0.omnuh2
57     # first a cheap linear run to find sigma_8 at the reference A_s
58     plin = camb.set_params(**{**par, "As": as_ref}, WantTransfer=True)
59     plin.set_matter_power(redshifts=[0.0], kmax=2.0)
60     s8_ref = camb.get_results(plin).get_sigma8_0()
61     par["As"] = as_ref * (s8 / s8_ref) ** 2
62     p = camb.set_params(**par, lmax=LMAX + 200, lens_potential_accuracy=1,
63                         NonLinear=camb.model.NonLinear_both, halofit_version="takahashi")
64     if windows:
65         p.SourceWindows = [SplinedSourceWindow(z=ZGRID[1:], W=nz[1:], source_type="lensing")
66                            for nz in NZS]
67     return p, par
68
69
70 def kappa_spectra(om, s8):
71     """Raw C_ell^{kappa kappa} (ell = 0..LMAX) of the three source samples, as rows."""
72     import camb
73     p, _ = camb_params(om, s8)
74     cls = camb.get_results(p).get_source_cls_dict(raw_cl=True, lmax=LMAX)
75     return np.array([cls[f"W{i}xW{i}"] for i in (1, 2, 3)])
```

CAMB is not given Ω_m and σ_8 directly, so the function translates. Line 56 finds the cold-dark-matter density from the total matter density at fixed h, keeping the massive neutrino inside Ω_m:

$$\Omega_{c0}h^2 = \Omega_{m0}h^2 - \Omega_{b0}h^2 - \Omega_{\nu0}h^2, \qquad \Omega_{\nu0}h^2 = \frac{\sum m_\nu}{93.14\,\mathrm{eV}} = \frac{0.06}{93.14} = 6.4\times10^{-4},$$

so the fiducial $\Omega_m = 0.3153$ gives $0.3153\times0.6736^2 - 0.02237 - 0.00064 = 0.1200$, the fiducial $\Omega_{c0}h^2$. CAMB also takes A_s, not σ_8, as input. By its definition σ_8^2 is an integral of the linear power,

$$\sigma_8^2 = \int \frac{\mathrm{d}k}{k}\,\Delta^2(k)\,W^2(kR), \qquad R = 8\,h^{-1}\mathrm{Mpc},$$

with $\Delta^2 = k^3P_\delta/2\pi^2$ and W the top-hat window (7.32). The linear $\Delta^2(k)$ is A_s times factors (the tilt, the transfer function, the growth) that do not depend on A_s, so σ_8^2 is proportional to A_s, and the ratio of two values of σ_8^2 equals the ratio of their A_s. Lines 58–61 therefore do one cheap linear run at a reference A_s to find $\sigma_{8,\mathrm{ref}}$, and then set $A_s = A_{s,\mathrm{ref}}(\sigma_8/\sigma_{8,\mathrm{ref}})^2$; no iteration is needed, because the relation is exact. Lines 62–66 switch on the halofit non-linear spectrum and attach the source distributions as lensing windows. Line 74 asks for the raw spectra (convention 1 of Section T1.6.2).

A check by hand. Before trusting the library call we check it. The function `limber_by_hand` in the same script evaluates the Limber row of Section T1.6 directly, in four moves. It works in redshift rather than distance, using $\mathrm{d}\chi = (c/H)\,\mathrm{d}z$ with CAMB's $H(z)$. It builds the lensing efficiency on the redshift grid: since $p_\chi\,\mathrm{d}\chi' = p_z\,\mathrm{d}z'$ and $f_K = \chi$, the integral of the second table becomes $g(\chi_i) = \int_{z_i} p_z(z')\,[1 - \chi_i/\chi(z')]\,\mathrm{d}z'$, done as a trapezoid sum for every grid point. It

evaluates CAMB's non-linear P_δ at $k = (\ell + \frac{1}{2})/\chi$ (convention 4) through CAMB's interpolator. And it multiplies by the prefactor $\frac{9}{4}(H_0/c)^4\Omega_m^2$ and sums $[g/a]^2 P_\delta\,(\mathrm{d}\chi/\mathrm{d}z)$ over the grid with the trapezoid rule. Over $100 \le \ell \le 2000$ the two agree to a median of 0.1% (the dashed black curve in Figure T1.9).

Why should they differ at all? CAMB samples the redshifts and interpolates P_δ in its own way, and our sum uses a grid of 700 redshifts and drops wavenumbers above $60\,\mathrm{Mpc}^{-1}$ (at $\ell = 2000$ only the matter within 33 Mpc of us); both effects are far below one per cent. The choice of k matters more: had one code used ℓ/χ and the other $(\ell + \frac{1}{2})/\chi$, k would differ by a fraction $1/(2\ell)$, 0.5% at $\ell = 100$, and with P_δ falling roughly as $k^{-1.5}$ there the spectra would differ by close to 1%. The agreement is ten times better, so both use the same choice. And at low multipoles, where the Limber approximation itself is poorest, a library may compute the exact projection instead, which is why the comparison starts at $\ell = 100$. The lesson: a library result is trusted only after a short independent version reproduces it.

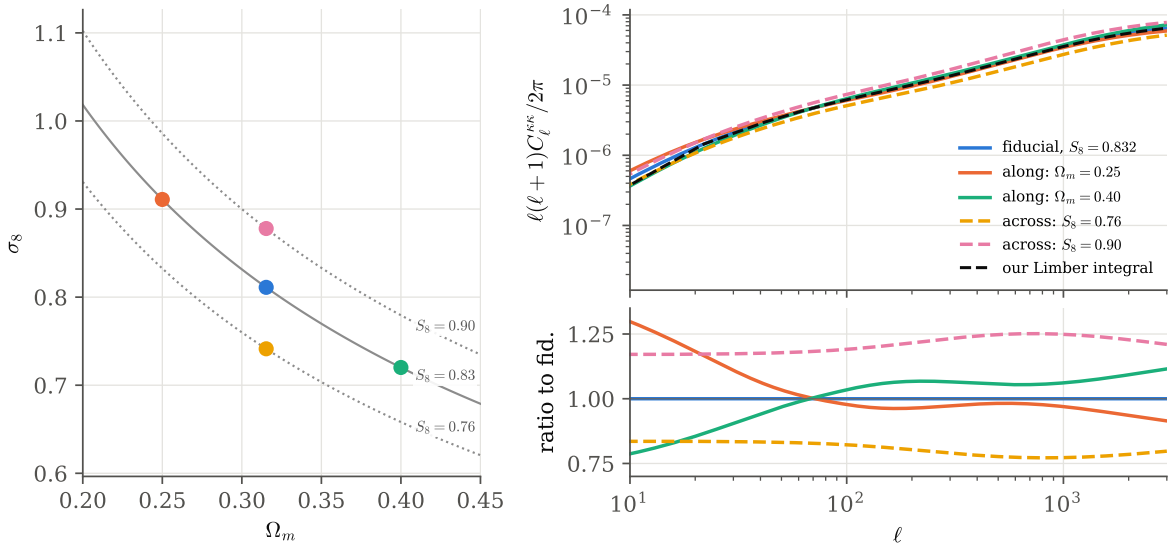


Figure T1.9: Left: the five models in the (Ω_m, σ_8) plane, with lines of constant S_8. Right: their convergence spectra (top) and the ratio to the fiducial one (bottom). The three solid curves, on the fiducial S_8 line, stay close together for $\ell \gtrsim 100$, even though Ω_m changes by a factor 1.6 along it. The two dashed curves, off the line by about 9% in S_8, are shifted by more than 20% at every ℓ. The dashed black curve is our own Limber integral for the fiducial model.

T1.7.4 The answer

Figure T1.9 is the answer. Along the line of constant S_8 the spectra differ by a median of −4% and +7% over $100 \le \ell \le 2000$ (at most 9%), although Ω_m runs from 0.25 to 0.40 and σ_8 from 0.911 to 0.720. Across the line, a change of about 9% in S_8 moves the spectrum by −22% and +24%. So the data pin down S_8 and leave the position along the line almost free: for $\ell \gtrsim 100$ the observation is about equally probable anywhere on the line, and the two candidates of the opening cannot be told apart there. Two things break that tie. A move across the line changes the spectrum by more than 20%, so the posterior is narrow in that direction. And at the low multipoles, $\ell \lesssim 50$, the three solid curves separate.

Why at low ℓ? The matter spectrum has a turnover, a peak at the wavenumber of the horizon size at matter–radiation equality, as described with the matter spectrum in Section T1.4. That wavenumber is $k_{\mathrm{eq}} \approx 0.073\,\Omega_{m0}h^2\,\mathrm{Mpc}^{-1}$, which is $0.073\,\Omega_{m0}h$ in the units $h\,\mathrm{Mpc}^{-1}$ in which σ_8 is defined (see the cosmology notes). At fixed h it moves in proportion to Ω_m, by the factor 1.6

between the ends of the line, and so the shape of P_δ around the turnover differs along the line even where its height does not. For the fiducial model $k_{\rm eq} \approx 0.073{\times}0.315{\times}0.674^2 \approx 0.010\,{\rm Mpc}^{-1}$; matter near the peak of the lensing weight, at $\chi \approx 1500$–$2000\,{\rm Mpc}$, projects it to $\ell = k\chi \approx 15$–20. Only the lowest multipoles see the turnover, and they carry few modes.

T1.7.5 Which exponent?

Is the exponent 0.5 in S_8 special? We can ask the spectrum which exponent it prefers. Change Ω_m by ±5% at fixed σ_8, then σ_8 by ±5% at fixed Ω_m, and form the logarithmic slopes. At $\ell = 500$ they are $\partial \ln C_\ell / \partial \ln \sigma_8 = 2.8$ and $\partial \ln C_\ell / \partial \ln \Omega_m = 1.6$.

The first slope can be checked against what we know. In linear theory $C_\ell \propto \sigma_8^2$ at fixed Ω_m, by (T1.9), so $\partial \ln C_\ell / \partial \ln \sigma_8 = 2$ exactly. Non-linear power grows faster than σ_8^2, because on small scales the extra clumping feeds on itself, and the measured 2.8 says how much faster: at $\ell = 500$ the spectrum responds about 40% more strongly than in linear theory. The same slope explains the across-line shifts of Figure T1.9: $(0.76/0.832)^{2.8} - 1 = -22\%$ and $(0.90/0.832)^{2.8} - 1 = +25\%$, close to the measured −22% and +24%, where the linear expectation of Section T1.7.1 was ±17%.

Which combination leaves the spectrum unchanged? Write $a \equiv \partial \ln C_\ell / \partial \ln \sigma_8$ and $b \equiv \partial \ln C_\ell / \partial \ln \Omega_m$, and ask for the small changes of the two parameters that keep C_ℓ fixed:

$$0 = {\rm d} \ln C_\ell \overset{(1)}{=} a\, {\rm d} \ln \sigma_8 + b\, {\rm d} \ln \Omega_m \quad \Longrightarrow \quad {\rm d} \ln \sigma_8 \overset{(2)}{=} -\frac{b}{a}\, {\rm d} \ln \Omega_m$$
$$\Longrightarrow \quad \ln \sigma_8 + \frac{b}{a} \ln \Omega_m \overset{(3)}{=} {\rm const} \quad \Longleftrightarrow \quad \sigma_8 \Omega_m^{b/a} = {\rm const}. \tag{T1.14}$$

(1) The change of a function of two variables is the sum of the changes along each (the chain rule), with $\ln \sigma_8$ and $\ln \Omega_m$ as the variables.
(2) Solve for ${\rm d} \ln \sigma_8$.
(3) Integrate, treating the ratio b/a as constant over the small region of interest; then take the exponential.

The exponent is the ratio of the two slopes,

$$\alpha(\ell) \equiv \frac{\partial \ln C_\ell^{\kappa\kappa} / \partial \ln \Omega_m}{\partial \ln C_\ell^{\kappa\kappa} / \partial \ln \sigma_8}. \tag{T1.15}$$

With the numbers above, $\alpha = 1.6/2.8 \approx 0.57$ at $\ell = 500$, which the unrounded slopes give as 0.55. For the fiducial sample $\alpha = 0.59$ at $\ell = 200$, 0.55 at $\ell = 500$ and 0.62 at $\ell = 2000$.

Figure T1.10 shows $\alpha(\ell)$ for three source samples, with median redshifts 0.49, 0.71 and 0.99. At $\ell = 500$ it is 0.58, 0.55 and 0.53. So the exponent 0.5 is a convention close to the truth, not a law: the best exponent depends on the scales and the redshifts a survey uses. This is why the surveys quote their own. HSC finds its tightest combination at $\alpha = 0.45$ and reports both $S_8(\alpha{=}0.45) = 0.800^{+0.029}_{-0.028}$ and the standard $S_8 = 0.780^{+0.030}_{-0.033}$ [49], and CMB lensing, with its deeper kernel, measures $\sigma_8 \Omega_m^{0.25}$ (Section T1.6).

T1.7.6 What the surveys find

The question the surveys asked is simple to state. The CMB shows the Universe at $z \approx 1100$, and the ΛCDM model extrapolates from it how clumpy matter should be today. Lensing measures today's clumpiness directly. *Do the two agree?* Each experiment measures something different:

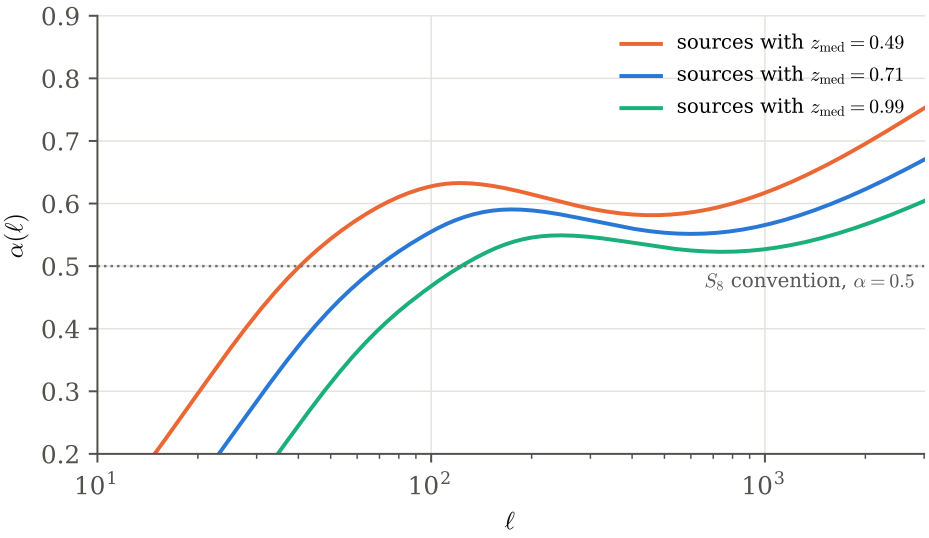


Figure T1.10: The exponent α of the combination $\sigma_8\Omega_m^\alpha$ that leaves $C_\ell^{\kappa\kappa}$ unchanged, multipole by multipole, Eq. (T1.15), for three source samples of increasing depth. Over the range where surveys have most of their signal, $100 \lesssim \ell \lesssim 1000$, it sits between 0.5 and 0.6, lower for deeper samples. It falls at low ℓ, where the shape of the matter spectrum near its turnover enters.

- *KiDS-1000* [45]: about 1000 deg^2 (777 deg^2 effective), about 21 million galaxies in five photometric bins out to $z \approx 1.2$; its headline number comes from a two-point statistic of the shear in all fifteen bin pairs.
- *DES Year 3* [50]: about 4100 deg^2, about 100 million galaxies in four bins, from the real-space correlations $\xi_\pm$.
- *HSC Year 1* [49]: only 137 deg^2 but the deepest images, four bins out to $z \approx 1.5$, from tomographic power spectra C_ℓ^{ij}.
- *Planck* [25]: the CMB temperature and polarization spectra over the full sky, with S_8 today computed from the fitted ΛCDM parameters.

experiment	published S_8	symmetrised error	distance from Planck
KiDS-1000	$0.759^{+0.024}_{-0.021}$	0.0225	2.8σ
DES Year 3	$0.759^{+0.025}_{-0.023}$	0.024	2.7σ
HSC Year 1	$0.780^{+0.030}_{-0.033}$	0.0315	1.5σ
Planck (CMB)	0.832 ± 0.013	0.013	—
this part	0.832; 0.76, 0.90	—	—

The distances in the last column are computed here, in one line of arithmetic. In the last row 0.832 is the solid model of Figure T1.9 and 0.76, 0.90 the dashed ones. If the two measurements are independent and their errors Gaussian, their difference has a variance equal to the sum of the two variances, the rule behind (3.20), so for KiDS-1000

$$\frac{0.832 - 0.759}{\sqrt{0.013^2 + 0.0225^2}} = \frac{0.073}{0.026} = 2.8.$$

The "about 3σ" that is often quoted is this number, rounded. It is only approximate: the errors are not symmetric (using the upper error, 0.024, which is the one facing Planck, gives 2.7σ), and the posteriors are not exactly Gaussian.

The sign of the shift is the one Figure T1.9 prepares. The lensing values sit near $S_8 = 0.76$, the lower dashed model, and the CMB value is the solid one: the surveys see a spectrum about 20% below what Planck's ΛCDM predicts. Whether this is new physics, a nuisance not fully modelled (the red boxes of Figure T1.8), or a fluctuation is one of the open questions of the field, and it is a question about statistics: about the sampling distribution of $\widehat{C}_\ell$, its covariance, and how much the nuisance parameters are allowed to move. The lesson of the comparison: a 9% difference in one number is a 20% difference in the spectrum, and whether it is significant depends entirely on errors that the later chapters learn to compute.

For the statistics chapters, the long valley of Figure T1.9 is a degeneracy of the kind met in Section T1.3: two parameters whose derivative curves are nearly proportional. In the Fisher matrix of Chapter 10 it makes the (Ω_m, σ_8) error ellipse long and thin, along $\sigma_8\Omega_m^\alpha = \text{const}$. In the MCMC of Chapter 9 it is a curved banana-shaped ridge. Chapter 12 forecasts the error on S_8 from simulated maps and asks whether summaries beyond $C_\ell^{\kappa\kappa}$ shorten the banana.

T1.8 Experiment 2: signal and shape noise in a Stage-III survey

The galaxy shapes carry a lensing signal of about 0.01 under an intrinsic scatter of 0.27. On which angular scales does the signal still win, and how many σ is the whole detection worth? In the language of Section T1.6: C_ℓ^{ij} is the convergence spectrum between redshift bin i and bin j, $N_i = \sigma_e^2/\bar{n}_i$ the shape-noise power of bin i from (T1.11), and the measured spectrum of bin i is $C_\ell^{ii} + N_i$. We want to know where C_ℓ^{ii} exceeds N_i.

T1.8.1 The survey

We take survey numbers close to KiDS-1000 [45]:

quantity	this experiment	KiDS-1000
area	$1000\,\text{deg}^2$	$\approx 1000\,\text{deg}^2$ ($777\,\text{deg}^2$ effective)
sky fraction f_{sky}	0.0242	—
galaxy density $\bar{n}$	6.2 arcmin^{-2}, all bins	6.2 arcmin^{-2}
shape scatter σ_e (per component)	0.27	0.25–0.27
photo-z scatter	Gaussian, $0.05(1+z)$	—
redshift bins	5, equal numbers	5

The sky fraction f_{sky} is the observed area divided by the area of the whole sky. The whole sky is 4π steradians, and one steradian is $(180/\pi)^2$ square degrees, so

$$f_{\text{sky}} = \frac{1000\,\text{deg}^2}{4\pi(180/\pi)^2\,\text{deg}^2} = \frac{1000}{41\,253} = 0.0242.$$

The source sample is the one of Section T1.7.2, with the fiducial cosmology.

T1.8.2 The binning

We cut the sample into five bins by photometric redshift, with equal numbers of galaxies in each. Equal numbers make every bin equally noisy, since N_i depends only on $\bar{n}_i$, so no bin is wasted on a handful of galaxies. A photometric redshift is estimated from a few broad colours, and its error grows with redshift because the spectral features it relies on are stretched by the factor $1 + z$; the simplest model of that is a Gaussian scatter of width $0.05(1 + z)$ about

the true redshift, without the outliers that real catalogues have. The bin edges come out at $z_{\rm ph} = 0.43, \ldots, 1.05$. Because of the scatter, a galaxy placed in one bin may truly belong to a neighbour, so the true-redshift distributions of the bins overlap.

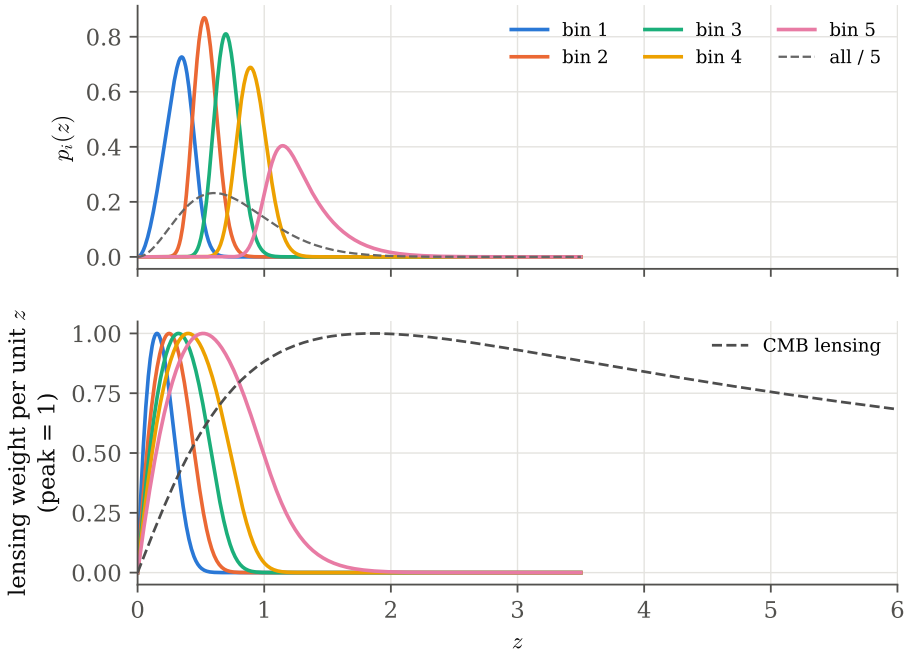


Figure T1.11: Top: the true-redshift distributions of the five photometric bins; photo-z scatter makes them overlap. Bottom: the lensing weight of each bin per unit redshift, $\propto g_i(\chi) f_K(\chi)/a$ times $\mathrm{d}\chi/\mathrm{d}z$, normalised to peak at one. Each weight is broad and peaks at about half the distance of its sources, so the five bins see heavily overlapping slices of matter. The CMB weight (dashed) peaks near $z \approx 2$.

Figure T1.11 shows the price of photometric redshifts and of projection together. The bins themselves overlap, and their lensing weights overlap even more, since every bin is lensed by all the matter in front of it. Tomography therefore does not give five independent maps. It gives five correlated maps whose full 5×5 matrix of spectra, auto and cross, carries the information, and the covariance of the analysis must include all fifteen of them.

T1.8.3 The noise and the error bars

Listing T1.4: The noise power of each bin (`code/chT1/04_kappa_noise.py`)

```
113     bins, chi, c0 = d["bins"], d["chi"], d["c0"]
114     frac = trapezoid(bins, Z, axis=1)                    # fraction of galaxies per
        bin
115     n_i = N_EFF * frac * ARCMIN2_PER_SR                  # per steradian
116     noise = SIGMA_E**2 / n_i
```

Line 114 counts the fraction of galaxies in each bin (a fifth, by construction). Line 115 converts the number density from arcmin^{-2} to sr^{-1}: one steradian holds $(10800/\pi)^2 = 1.18 \times 10^7$ arcmin^2, so 1.24 arcmin^{-2} becomes 1.47×10^7 sr^{-1}. Line 116 is the noise row of the table, $N_i = \sigma_e^2/\bar{n}_i \approx 4.965 \times 10^{-9}$, worked out in Section T1.6.

How large are the error bars on a measured band power? Start from the cosmic variance of

one multipole and make three changes:

$$\operatorname{Var}\widehat{C}_\ell \overset{(1)}{=} \frac{2C_\ell^2}{\nu} \overset{(2)}{=} \frac{2\,(C_\ell+N)^2}{(2\ell+1)\,f_{\rm sky}}, \qquad \operatorname{Var}\Big(\frac{1}{\Delta\ell}\sum_{\ell'\in\text{band}}\widehat{C}_{\ell'}\Big) \overset{(3)}{=} \frac{2\,(C_\ell+N)^2}{(2\ell+1)\,\Delta\ell\,f_{\rm sky}},$$

$$\Delta C_\ell \overset{(4)}{=} \sqrt{\frac{2}{(2\ell+1)\,\Delta\ell\,f_{\rm sky}}}\,(C_\ell+N). \tag{T1.16}$$

(1) The cosmic variance (7.77), written for ν independent modes; on the full sky $\nu = 2\ell+1$.
(2) Two changes. A survey that sees a fraction $f_{\rm sky}$ of the sky has only about $(2\ell+1)f_{\rm sky}$ independent modes per multipole. And what is measured is the spectrum of signal plus noise, whose variance per mode is $C_\ell + N$, so the scatter of the estimate is set by the sum, as in (7.81); the noise is part of what is measured.
(3) Averaging the $\Delta\ell$ independent multipoles of a band divides the variance by $\Delta\ell$, (7.79), the variance of a mean.
(4) The square root.

This is the Gaussian band-power error used in the literature [44, eq. 52].

What does this mean for inference? The observed band power is the model prediction plus one draw of noise with the variance of (T1.16). Read as a function of the parameters (Ω_m, σ_8), the probability of the data asks how plausible the noise draw is that turns the prediction $C_\ell(\Omega_m, \sigma_8)$ into what was observed: a Gaussian centred on the model, with the width of (T1.16). This is the likelihood from a noise model, (0.9), in its band-power form (Section 7.10.3). The S_8 degeneracy of Section T1.7 is a statement about this likelihood: moving along the line leaves the centre almost where it was, so the same noise draw is needed everywhere on the line, and the likelihood does not change.

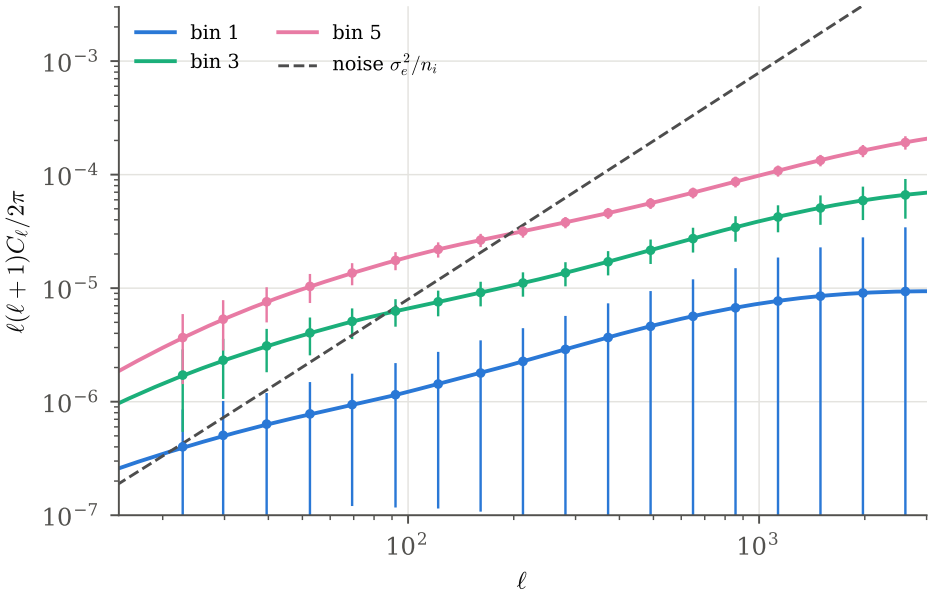


Figure T1.12: Convergence spectra of the nearest (1), middle (3) and farthest (5) bins, with Gaussian error bars on logarithmic bands of ℓ for a 1000 deg^2 survey, Eq. (T1.16). The dashed line is the shape noise, the same for every bin because the bins hold equal numbers of galaxies; it rises as ℓ^2 only because of the factor $\ell(\ell+1)$ on the axis. Beyond the crossing, the error bars are set mostly by the noise.

T1.8.4 Where the signal wins, and how many σ

Figure T1.12 gives the answer to the first question. The signal equals the noise at $\ell \approx 21$ in the nearest bin, $\ell \approx 87$ in the middle one and $\ell \approx 194$ in the farthest, which is lensed by the most matter. Beyond those multipoles each bin is noise-dominated. The measurement does not stop there, since there are many modes at high ℓ, but each additional mode adds little.

The signal-to-noise ratio of one bin. Ask how well the data fix the height of the signal. Multiply the predicted spectrum by an amplitude A, so that the measured spectrum has mean $AC_\ell + N$, and define the signal-to-noise ratio as $1/\sigma(A)$ at $A = 1$: the number of standard deviations by which the signal is distinguished from no signal. The Fisher information for A comes from (10.31), with $(2\ell+1)f_{\rm sky}$ modes per multipole:

$$\left(\frac{S}{N}\right)^2 \overset{(1)}{=} \sum_\ell \frac{(2\ell+1)f_{\rm sky}}{2} \frac{\left[\partial(AC_\ell+N)/\partial A\right]^2}{(C_\ell+N)^2} \overset{(2)}{=} \sum_\ell \frac{(2\ell+1)f_{\rm sky}}{2} \frac{C_\ell^2}{(C_\ell+N)^2}. \tag{T1.17}$$

(1) Equation (10.31) for one parameter, with the full-sky count $2\ell+1$ replaced by $(2\ell+1)f_{\rm sky}$, and the inverse error squared equal to the Fisher information.
(2) The derivative of $AC_\ell + N$ with respect to A is C_ℓ.

Each mode contributes the factor $\frac{1}{2}[C_\ell/(C_\ell+N)]^2$. Where the signal dominates, $C_\ell \gg N$, that factor is $\frac{1}{2}$, the most a mode can give; where the noise dominates it falls as $\frac{1}{2}(C_\ell/N)^2$, and since N is flat while C_ℓ falls, it drops quickly beyond the crossing. So the signal-to-noise ratio essentially counts the modes below the crossing multipole.

Five bins at once. With tomography the spectrum at each ℓ is a 5×5 matrix $\mathbf{C}_\ell$ with entries $C_\ell^{ij} + \delta_{ij}N_i$, and $\mathbf{S}_\ell$ is its signal part. The same calculation with the matrix form of the Fisher information gives

$$\left(\frac{S}{N}\right)^2 = \sum_\ell \frac{(2\ell+1)f_{\rm sky}}{2} \,\mathrm{Tr}\left[\mathbf{C}_\ell^{-1}\mathbf{S}_\ell\,\mathbf{C}_\ell^{-1}\mathbf{S}_\ell\right], \tag{T1.18}$$

which for one bin reduces to (T1.17). The fifteen distinct spectra are the entries of $\mathbf{S}_\ell$, and the inverse $\mathbf{C}_\ell^{-1}$ accounts for the correlations between bins that Figure T1.11 shows, so that shared information is not counted twice. Summing over $10 \le \ell \le 2000$, the total signal-to-noise ratio is 48. A single unbinned sample, with five times the galaxies and so a fifth of the noise, gives 42. The ratio is $48/42 \approx 1.14$: tomography adds about 14% to the detection. What it buys is the redshift dependence, which separates the growth of structure from the geometry and helps tell intrinsic alignments from lensing.

How well is the amplitude measured? If σ_8 were the only unknown, the Fisher information on $\ln\sigma_8$ is (T1.18) with each $\mathbf{S}_\ell$ replaced by the derivative $\partial\mathbf{C}_\ell/\partial\ln\sigma_8$ (the general form is (10.20)), and with a single parameter the error is $\sigma(\ln\sigma_8) = 1/\sqrt{F}$. The derivative comes from two more CAMB runs at $\sigma_8(1\pm\epsilon)$ with $\epsilon = 0.02$, as a central difference (Section 10.5):

$$\frac{\partial\mathbf{C}_\ell}{\partial\ln\sigma_8} \approx \frac{\mathbf{C}_\ell^+ - \mathbf{C}_\ell^-}{\ln(1+\epsilon) - \ln(1-\epsilon)} \approx \frac{\mathbf{C}_\ell^+ - \mathbf{C}_\ell^-}{2\epsilon}.$$

The result is $\sigma(\ln\sigma_8) = 0.8\%$. It can be checked against the signal-to-noise ratio. If the derivative were a fixed multiple s of the signal, $\partial\mathbf{C}_\ell/\partial\ln\sigma_8 = s\,\mathbf{S}_\ell$, then $F = s^2(S/N)^2$ and $\sigma(\ln\sigma_8) = 1/[s\,(S/N)]$. In linear theory $s = 2$, giving $1/(2\times48) = 1.0\%$. The computed 0.8% corresponds to $s \approx 1/(0.008\times48) \approx 2.6$, between the linear 2 and the 2.8 found at $\ell = 500$ in Section T1.7.5: the non-linear boost again.

At fixed Ω_m this is also the error on S_8. KiDS-1000 achieves about 3%. The gap between the two numbers is the content of the later chapters. The first cost can be sized here. Ω_m is free and

degenerate with σ_8 (Section T1.7). For two parameters with Fisher matrix entries $F_{\sigma\sigma}$, $F_{\Omega\Omega}$ and $F_{\sigma\Omega}$,

$$\sigma^2_{\text{marg}}(\ln\sigma_8) \overset{(1)}{=} \left(F^{-1}\right)_{\sigma\sigma} \overset{(2)}{=} \frac{F_{\Omega\Omega}}{F_{\sigma\sigma}F_{\Omega\Omega} - F^2_{\sigma\Omega}} \overset{(3)}{=} \frac{1}{F_{\sigma\sigma}}\,\frac{1}{1-r^2}, \qquad r \equiv \frac{F_{\sigma\Omega}}{\sqrt{F_{\sigma\sigma}F_{\Omega\Omega}}}. \tag{T1.19}$$

(1) The marginal error is read from the inverse of the Fisher matrix (Section 10.4.3).
(2) The inverse of a 2×2 matrix.
(3) Divide the numerator and the denominator by $F_{\sigma\sigma}F_{\Omega\Omega}$; r is the cosine of the angle between the two derivative curves, weighted as in the Fisher sum.

So freeing Ω_m multiplies the error on $\ln\sigma_8$ by $1/\sqrt{1-r^2}$: by 2.3 for $r = 0.9$ and by 7 for $r = 0.99$, and the near-proportional derivative curves of Section T1.7 put r close to one. The combination S_8 is chosen to lie along the well-measured direction, so its error grows much less, which is why the surveys quote it. The rest of the gap comes from the other costs. Nuisance parameters for photo-z shifts, shear bias, intrinsic alignment and baryons are marginalised. The covariance has non-Gaussian and super-sample terms that Eq. (T1.16) ignores. And the smallest scales are cut because baryonic feedback is uncertain there. Chapter 10 computes each of these costs in the Fisher matrix, and Chapter 12 redoes the forecast with simulated maps.

How this is used in real research: What the full analysis does

Our numbers come from a Gaussian, nuisance-free forecast with all other parameters fixed. The published analyses measure tomographic $\xi_\pm(\vartheta)$, band powers or related statistics from tens of millions of shapes [45, 49, 50]. They take the covariance from analytic models checked against hundreds to thousands of ray-traced N-body simulations, such as the full-sky lensing maps of [51], and they sample the posterior of cosmological and nuisance parameters by MCMC. The same simulations are what make summaries beyond two-point statistics usable: peak counts [52, 53], Kaiser–Squires mass maps [54], lognormal fields as a fast model of the skewed κ distribution [55], and persistent homology of κ maps [56]. All of these return in Chapter 12.

The lesson of the two galaxy-shape experiments for the rest of the book is in two parts. The signal is a projected, smoothed field whose two-point statistic is degenerate along S_8. The noise is white, large and known. So the statistical problem is to extract as much as possible from a noisy, non-Gaussian field with a small number of summaries. Chapter 12 takes up that problem with topological summaries.

T1.9 Experiment 3: the spectrum of a reconstructed lensing map, and its two biases

The quadratic estimator of Section T1.6 turns a CMB temperature map into a map $\hat\psi$ of the lensing potential. Planck's 40σ is a statement about the power spectrum of that map. So take $\hat\psi$, square its Fourier modes and average them in rings of L, exactly as one does for any map. The answer is larger than $C_L^{\psi\psi}$, and at most multipoles it is many times larger. Nothing went wrong in the estimator. What could have added the extra power? There are two candidates, and we take them one at a time.

1. *Chance pairs.* Each product $\hat T_{\boldsymbol\ell_1}\hat T_{\boldsymbol\ell_2}$ in the estimator contains the chance product of two unlensed modes, and of two noise values. On average such products vanish, so $\hat\psi$ is unbiased. But the power spectrum squares $\hat\psi$, and the square of a chance fluctuation does

not average to zero. This is the noise bias of Section 8.3 once more, one level up: there the noise of a map added to its spectrum; here the noise of the reconstruction, which comes mostly from the CMB itself, adds to the spectrum of $\hat{\psi}$. It would be there with no lensing at all, that is at zeroth order in $C^{\psi\psi}$, and it is called $N_L^{(0)}$. The general rule is worth stating. When the quantity we measure is a fixed number, noise scatters the estimate about it, and averaging many readings removes the scatter. When the quantity is itself a variance, as a power spectrum is, the noise brings a variance of its own, and that adds to the answer: averaging makes it more precise, not smaller. What is scatter for a deterministic quantity becomes a bias for a statistical one.

2. *Lensing, used twice.* The spectrum of $\hat{\psi}$ is built from four CMB modes, two in each copy of $\hat{\psi}$, and all four were deflected by the same potential. The pairing we want correlates the two modes inside each copy through $\psi_{\boldsymbol{L}}$. But a mode of the first copy and a mode of the second are also correlated, through the lensing at the sum of their wavevectors, which is some other $\boldsymbol{L}'$. Those correlations add power proportional to $C^{\psi\psi}$, at first order, and the term is called $N_L^{(1)}$.

The experiment derives both terms and checks each against simulated skies.

T1.9.1 The estimator and its best weights

On a flat patch the Fourier modes of a statistically homogeneous field X obey $\langle X_{\boldsymbol{\ell}} X^*_{\boldsymbol{\ell}'}\rangle = (2\pi)^2\delta^2(\boldsymbol{\ell}-\boldsymbol{\ell}')\,C_\ell$, the flat-sky form of (T1.39), and $X^*_{\boldsymbol{\ell}} = X_{-\boldsymbol{\ell}}$ because X is real. Write the weight of (T1.10) as

$$f(\boldsymbol{\ell}_1,\boldsymbol{\ell}_2) \equiv (\boldsymbol{L}\cdot\boldsymbol{\ell}_1)\,C_{\ell_1} + (\boldsymbol{L}\cdot\boldsymbol{\ell}_2)\,C_{\ell_2}, \qquad \boldsymbol{L} = \boldsymbol{\ell}_1 + \boldsymbol{\ell}_2,$$

so that $\langle \hat{T}_{\boldsymbol{\ell}_1}\hat{T}_{\boldsymbol{\ell}_2}\rangle_{\rm CMB} = f(\boldsymbol{\ell}_1,\boldsymbol{\ell}_2)\,\psi_{\boldsymbol{L}}$, with the unlensed spectrum C_ℓ. The observed map is $\hat{T}$ plus noise, and its total spectrum, lensed CMB plus noise, is $C_\ell^{\rm tot}$. A general quadratic estimator multiplies every pair of observed modes that adds up to $\boldsymbol{L}$ by a weight F, symmetric in its two arguments, and sums:

$$\hat{\psi}_{\boldsymbol{L}} = A_L \int \frac{\mathrm{d}^2\boldsymbol{\ell}_1}{(2\pi)^2}\,F(\boldsymbol{\ell}_1,\boldsymbol{\ell}_2)\,\hat{T}_{\boldsymbol{\ell}_1}\hat{T}_{\boldsymbol{\ell}_2}, \qquad \boldsymbol{\ell}_2 = \boldsymbol{L}-\boldsymbol{\ell}_1. \tag{T1.20}$$

What fixes A_L? The mean over CMB and noise at fixed ψ:

$$\langle\hat{\psi}_{\boldsymbol{L}}\rangle \overset{(1)}{=} A_L \int \frac{\mathrm{d}^2\boldsymbol{\ell}_1}{(2\pi)^2}\,F(\boldsymbol{\ell}_1,\boldsymbol{\ell}_2)\,f(\boldsymbol{\ell}_1,\boldsymbol{\ell}_2)\,\psi_{\boldsymbol{L}} \overset{(2)}{=} \psi_{\boldsymbol{L}} \quad \text{if} \quad \frac{1}{A_L} = \int \frac{\mathrm{d}^2\boldsymbol{\ell}_1}{(2\pi)^2}\,F\,f. \tag{T1.21}$$

(1) The average acts only on the product of modes, which is (T1.10); the noise is independent of the CMB and has mean zero, so it adds nothing for $\boldsymbol{L}\neq\boldsymbol{0}$.
(2) Choose A_L so that the integral times A_L is one.

How noisy is it? Switch lensing off and let the map be Gaussian with spectrum $C^{\rm tot}$. The power spectrum of the estimate is defined like any other, $\langle\hat{\psi}_{\boldsymbol{L}}\hat{\psi}^*_{\boldsymbol{L}'}\rangle = (2\pi)^2\delta^2(\boldsymbol{L}-\boldsymbol{L}')\,C_L^{\hat{\psi}\hat{\psi}}$, and the four-point average is a sum over pairings by Wick's theorem (7.8):

$$\begin{aligned}
\langle\hat{\psi}_{\boldsymbol{L}}\hat{\psi}^*_{\boldsymbol{L}'}\rangle &\overset{(1)}{=} A_L^2 \int \frac{\mathrm{d}^2\boldsymbol{\ell}_1}{(2\pi)^2}\frac{\mathrm{d}^2\boldsymbol{\ell}_1'}{(2\pi)^2}\,F(\boldsymbol{\ell}_1,\boldsymbol{\ell}_2)F(\boldsymbol{\ell}_1',\boldsymbol{\ell}_2')\,\big\langle\hat{T}_{\boldsymbol{\ell}_1}\hat{T}_{\boldsymbol{\ell}_2}\hat{T}^*_{\boldsymbol{\ell}_1'}\hat{T}^*_{\boldsymbol{\ell}_2'}\big\rangle \\
&\overset{(2)}{=} A_L^2 \int \frac{\mathrm{d}^2\boldsymbol{\ell}_1}{(2\pi)^2}\frac{\mathrm{d}^2\boldsymbol{\ell}_1'}{(2\pi)^2}\,F(\boldsymbol{\ell}_1,\boldsymbol{\ell}_2)F(\boldsymbol{\ell}_1',\boldsymbol{\ell}_2')\,(2\pi)^4\,C^{\rm tot}_{\ell_1}C^{\rm tot}_{\ell_2} \\
&\qquad\times\big[\delta^2(\boldsymbol{\ell}_1-\boldsymbol{\ell}_1')\delta^2(\boldsymbol{\ell}_2-\boldsymbol{\ell}_2') + \delta^2(\boldsymbol{\ell}_1-\boldsymbol{\ell}_2')\delta^2(\boldsymbol{\ell}_2-\boldsymbol{\ell}_1')\big] \\
&\overset{(3)}{=} (2\pi)^2\delta^2(\boldsymbol{L}-\boldsymbol{L}')\,2A_L^2\int\frac{\mathrm{d}^2\boldsymbol{\ell}_1}{(2\pi)^2}\,F(\boldsymbol{\ell}_1,\boldsymbol{\ell}_2)^2\,C^{\rm tot}_{\ell_1}C^{\rm tot}_{\ell_2} \equiv (2\pi)^2\delta^2(\boldsymbol{L}-\boldsymbol{L}')\,N_L^{(0)}.
\end{aligned} \tag{T1.22}$$

(1) Insert (T1.20) twice, the second time conjugated and with $\boldsymbol{\ell}_1' + \boldsymbol{\ell}_2' = \boldsymbol{L}'$; A_L and F are real.
(2) Of the three pairings, the one inside each copy, $\langle \hat{T}_{\boldsymbol{\ell}_1} \hat{T}_{\boldsymbol{\ell}_2} \rangle$, vanishes for $\boldsymbol{L} \neq \boldsymbol{0}$ without lensing. The other two pair a mode of the first copy with a mode of the second, each pair giving $(2\pi)^2 \delta^2 \, C^{\rm tot}$.
(3) The first delta function sets $\boldsymbol{\ell}_1' = \boldsymbol{\ell}_1$, and then $\boldsymbol{\ell}_2 - \boldsymbol{\ell}_2' = \boldsymbol{L} - \boldsymbol{L}'$ leaves the delta function in $\boldsymbol{L}$. The second sets $\boldsymbol{\ell}_1' = \boldsymbol{\ell}_2$ and gives the same with $F(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2) F(\boldsymbol{\ell}_2, \boldsymbol{\ell}_1) = F(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2)^2$, since F is symmetric: twice the same integral.

Which weights are best? Write the normalisation integral as an integral of a product and apply the Cauchy–Schwarz inequality, $\left(\int ab\right)^2 \leq \int a^2 \int b^2$:

$$\frac{1}{A_L^2} = \left(\int \frac{\mathrm{d}^2 \boldsymbol{\ell}_1}{(2\pi)^2} \, F \sqrt{2 C_{\ell_1}^{\rm tot} C_{\ell_2}^{\rm tot}} \, \frac{f}{\sqrt{2 C_{\ell_1}^{\rm tot} C_{\ell_2}^{\rm tot}}} \right)^2 \leq \int \frac{\mathrm{d}^2 \boldsymbol{\ell}_1}{(2\pi)^2} \, 2F^2 C_{\ell_1}^{\rm tot} C_{\ell_2}^{\rm tot} \int \frac{\mathrm{d}^2 \boldsymbol{\ell}_1}{(2\pi)^2} \, \frac{f^2}{2 C_{\ell_1}^{\rm tot} C_{\ell_2}^{\rm tot}}. \tag{T1.23}$$

Multiply both sides by A_L^2 and use (T1.22): $1 \leq N_L^{(0)} \int f^2 / (2 C^{\rm tot} C^{\rm tot})$, so no choice of F can push the noise below $1 / \int f^2 / (2 C^{\rm tot} C^{\rm tot})$. Equality holds when the two factors are proportional, $F \propto f / (2 C_{\ell_1}^{\rm tot} C_{\ell_2}^{\rm tot})$, and since a constant factor in F is undone by A_L we may take

$$F(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2) = \frac{f(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2)}{2 C_{\ell_1}^{\rm tot} C_{\ell_2}^{\rm tot}}, \qquad \frac{1}{A_L} = \int \frac{\mathrm{d}^2 \boldsymbol{\ell}_1}{(2\pi)^2} \, \frac{f(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2)^2}{2 C_{\ell_1}^{\rm tot} C_{\ell_2}^{\rm tot}}, \qquad N_L^{(0)} = A_L. \tag{T1.24}$$

This is the estimator of Hu and Okamoto [57]. The weight is the expected signal of the pair, f, divided by the variance of the pair, $C_{\ell_1}^{\rm tot} C_{\ell_2}^{\rm tot}$, the same rule that weights each reading of an average by its inverse variance. And the noise of the estimate equals its normalisation, so computing A_L computes $N_L^{(0)}$ for free. Check: with (T1.24), $2A_L^2 \int F^2 C^{\rm tot} C^{\rm tot} = 2A_L^2 \int f^2 / (4 C^{\rm tot} C^{\rm tot}) = A_L^2 / A_L = A_L$.

T1.9.2 The four-point function of a lensed sky

With lensing on, the four-point average in (T1.22) is no longer a Gaussian one, and we need it. One observation makes it easy. *Hold ψ fixed.* The first-order lensed map $T + \boldsymbol{\nabla}\psi \cdot \boldsymbol{\nabla} T$ is then a linear function of the Gaussian unlensed T, and adding Gaussian noise keeps it Gaussian. So for a given ψ the observed map is Gaussian, Wick's theorem applies, and all we need is its covariance at fixed ψ:

$$\langle \hat{T}_{\boldsymbol{a}} \hat{T}_{\boldsymbol{b}} \rangle_\psi = (2\pi)^2 \delta^2(\boldsymbol{a} + \boldsymbol{b}) \, C_a^{\rm tot} + f(\boldsymbol{a}, \boldsymbol{b}) \, \psi_{\boldsymbol{a}+\boldsymbol{b}}, \tag{T1.25}$$

the diagonal part plus the off-diagonal part of (T1.10). (The diagonal part depends weakly on ψ too; averaged over ψ it is the lensed spectrum, which is why $C^{\rm tot}$ is the spectrum of the lensed map.) Now average the four-point function first over the CMB and noise at fixed ψ, then over ψ, which is Gaussian with $\langle \psi_{\boldsymbol{L}} \psi_{\boldsymbol{L}'} \rangle = (2\pi)^2 \delta^2(\boldsymbol{L} + \boldsymbol{L}') \, C_L^{\psi\psi}$:

$$\begin{aligned}
\langle \hat{T}_{\boldsymbol{a}} \hat{T}_{\boldsymbol{b}} \hat{T}_{\boldsymbol{c}} \hat{T}_{\boldsymbol{d}} \rangle &\overset{(1)}{=} \Big\langle \langle \hat{T}_{\boldsymbol{a}} \hat{T}_{\boldsymbol{b}} \rangle_\psi \langle \hat{T}_{\boldsymbol{c}} \hat{T}_{\boldsymbol{d}} \rangle_\psi + \langle \hat{T}_{\boldsymbol{a}} \hat{T}_{\boldsymbol{c}} \rangle_\psi \langle \hat{T}_{\boldsymbol{b}} \hat{T}_{\boldsymbol{d}} \rangle_\psi + \langle \hat{T}_{\boldsymbol{a}} \hat{T}_{\boldsymbol{d}} \rangle_\psi \langle \hat{T}_{\boldsymbol{b}} \hat{T}_{\boldsymbol{c}} \rangle_\psi \Big\rangle_\psi \\
&\overset{(2)}{=} \big[\text{the three Gaussian pairings with } C^{\rm tot}\big] \\
&\quad + (2\pi)^2 \delta^2(\boldsymbol{a} + \boldsymbol{b} + \boldsymbol{c} + \boldsymbol{d}) \Big[f(\boldsymbol{a}, \boldsymbol{b}) f(\boldsymbol{c}, \boldsymbol{d}) \, C_{|\boldsymbol{a}+\boldsymbol{b}|}^{\psi\psi} \\
&\qquad + f(\boldsymbol{a}, \boldsymbol{c}) f(\boldsymbol{b}, \boldsymbol{d}) \, C_{|\boldsymbol{a}+\boldsymbol{c}|}^{\psi\psi} \\
&\qquad + f(\boldsymbol{a}, \boldsymbol{d}) f(\boldsymbol{b}, \boldsymbol{c}) \, C_{|\boldsymbol{a}+\boldsymbol{d}|}^{\psi\psi} \Big].
\end{aligned} \tag{T1.26}$$

(1) Wick's theorem (7.8) for the Gaussian map at fixed ψ, then the average over ψ.
(2) Insert (T1.25) into each product. Diagonal times diagonal gives the Gaussian pairings. Diagonal times off-diagonal is linear in ψ and averages to zero. Off-diagonal times off-diagonal gives $f\,f\,\langle\psi_{\boldsymbol{a}+\boldsymbol{b}}\psi_{\boldsymbol{c}+\boldsymbol{d}}\rangle$, and the average of the two potentials is $(2\pi)^2\delta^2(\boldsymbol{a}+\boldsymbol{b}+\boldsymbol{c}+\boldsymbol{d})\,C^{\psi\psi}_{|\boldsymbol{a}+\boldsymbol{b}|}$.

So lensing adds to the four-point function one term per pairing, each the product of the two pair correlations of (T1.10) times the spectrum of the potential at the pair's total wavevector. This is the connected four-point function, the trispectrum, of the lensed CMB.

The spectrum of the estimate. Insert (T1.26) into the first line of (T1.22), written with $\hat{T}^*_{\boldsymbol{\ell}'} = \hat{T}_{-\boldsymbol{\ell}'}$, so that the four wavevectors are $\boldsymbol{a}=\boldsymbol{\ell}_1$, $\boldsymbol{b}=\boldsymbol{\ell}_2$, $\boldsymbol{c}=-\boldsymbol{\ell}'_1$, $\boldsymbol{d}=-\boldsymbol{\ell}'_2$. Each pairing gives one term:

$$C_L^{\hat\psi\hat\psi} = C_L^{\psi\psi} + N_L^{(0)} + N_L^{(1)}, \tag{T1.27}$$

$$N_L^{(1)} = 2A_L^2 \int \frac{\mathrm{d}^2\boldsymbol{\ell}_1}{(2\pi)^2}\frac{\mathrm{d}^2\boldsymbol{\ell}'_1}{(2\pi)^2}\,F(\boldsymbol{\ell}_1,\boldsymbol{\ell}_2)F(\boldsymbol{\ell}'_1,\boldsymbol{\ell}'_2)\,f(\boldsymbol{\ell}_1,-\boldsymbol{\ell}'_1)\,f(\boldsymbol{\ell}_2,-\boldsymbol{\ell}'_2)\,C^{\psi\psi}_{|\boldsymbol{\ell}_1-\boldsymbol{\ell}'_1|}. \tag{T1.28}$$

- The lensing term of the pairing inside each copy, $(\boldsymbol{ab})(\boldsymbol{cd})$, is $f(\boldsymbol{\ell}_1,\boldsymbol{\ell}_2)f(-\boldsymbol{\ell}'_1,-\boldsymbol{\ell}'_2)\,C_L^{\psi\psi}$. Since f is even under a change of sign of both arguments, the two integrals each give $1/A_L$ by (T1.21), and the term is $C_L^{\psi\psi}$: the signal.
- The Gaussian pairings give (T1.22), $N_L^{(0)}$, with the lensed $C^{\rm tot}$.
- The lensing terms of the two pairings across the copies give $N_L^{(1)}$. The first pairs $\boldsymbol{\ell}_1$ with $-\boldsymbol{\ell}'_1$ through the potential at $\boldsymbol{\ell}_1-\boldsymbol{\ell}'_1$; the second pairs $\boldsymbol{\ell}_1$ with $-\boldsymbol{\ell}'_2$ and equals the first after the relabelling $\boldsymbol{\ell}'_1\leftrightarrow\boldsymbol{\ell}'_2$, which leaves the symmetric F unchanged: hence the factor 2.

$N^{(1)}$ is the bias of [58]. Its potential is evaluated at $\boldsymbol{\ell}_1-\boldsymbol{\ell}'_1$, which ranges over all wavevectors, so the measured spectrum at one L responds to the potential at all others. It cannot be computed without assuming a $C^{\psi\psi}$, and in practice it is computed with a fiducial one. An error in that assumption changes $N^{(1)}$ in proportion, so a 10% error in a bias of a few per cent of the signal moves the result by a few tenths of a per cent.

First order, and the real sky. Everything above is first order in the deflection. How small is it? The rms deflection is about 2.5′, which is 7×10^{-4} rad, so for a mode at $\ell=2000$ the shift is about 1.5 radians of phase: not small at all. The expansion $T+\boldsymbol{\nabla}\psi\cdot\boldsymbol{\nabla}T$ describes such a mode badly, and a simulation built on it would put far too much power at high ℓ. The real remapping $T(\boldsymbol{n}+\boldsymbol{\nabla}\psi)$ is gentler than its expansion, because most of the deflection comes from large-scale potential modes that move whole regions of the map together, and a common shift $\boldsymbol{d}$ only multiplies each product $\hat{T}_{\boldsymbol{\ell}_1}\hat{T}_{\boldsymbol{\ell}_2}$ by $e^{i\boldsymbol{L}\cdot\boldsymbol{d}}$, which is close to one for $L\,d\ll1$, whatever $\boldsymbol{\ell}_1$ and $\boldsymbol{\ell}_2$ are. Analyses keep the structure derived here and absorb most of the higher orders by one change: the weight f is built with the lensed spectrum instead of the unlensed one [59]. Our simulations do the same, and they lens by remapping, not by the expansion; how well the first-order formulas then hold is something they measure.

The mean field. On our periodic patches the mean of $\hat\psi_{\boldsymbol{L}}$ is zero for $\boldsymbol{L}\neq\boldsymbol{0}$, by (T1.21) averaged over ψ. On a real sky it is not. A mask, or noise that is deeper in some places than in others, breaks the statistical homogeneity that made unlensed modes independent, and it correlates pairs of modes just as lensing does. The estimator cannot tell the two apart, and it returns a **mean field** $\langle\hat\psi_{\boldsymbol{L}}\rangle\neq0$ even on unlensed skies. It is estimated as the average of $\hat\psi$ over simulations with the real mask and noise, and subtracted from $\hat\psi$ before squaring [48, 59].

T1.9.3 What we compute, and how

The script `code/chT1/11_lens_n0n1.py` works on periodic flat patches of $17^\circ \times 17^\circ$ (512^2 pixels of 2′), temperature only. Each sky is a Gaussian unlensed T with CAMB's spectrum and an independent Gaussian ψ with CAMB's $C^{\psi\psi}$, both drawn on a grid twice as fine (1′ pixels). The temperature is remapped, $T(\boldsymbol{n} + \boldsymbol{\nabla}\psi)$, by cubic-spline interpolation on the fine grid, then cut back to the 2′ grid by keeping its Fourier modes, and white noise of 5 μK-arcmin is added. The estimator uses the multipoles $30 \le \ell \le 2500$. The computation runs in five steps.

1. *The total spectrum.* $C_\ell^{\rm tot}$ is measured as the ring average of $\left|\hat{T}_{\boldsymbol{\ell}}\right|^2$ over the same 400 lensed skies whose reconstructions step 4 analyses: the spectrum of the maps themselves, as on real data. (Each sky has its own random stream, so it can be drawn a second time.) This matters more than it seems. At high L, $N^{(0)}$ is a hundred times the signal, so an error of half a per cent in $C^{\rm tot}$, and hence in $N^{(0)}$, would be as large as the whole of $N^{(1)}$. Measuring $C^{\rm tot}$ on an independent set of skies leaves exactly such an error. Measured on the same skies, the chance excess or deficit of their spectrum enters $\hat{C}^{\hat\psi\hat\psi}$ and A_L alike, and cancels to first order.
2. *The weights and A_L.* The weight f uses the lensed CMB spectrum, $C_\ell^{\rm tot}$ minus the noise. Products in Fourier space are convolutions, so every integral of the form $\int \mathrm{d}^2\ell_1\, g(\boldsymbol{\ell}_1)h(\boldsymbol{L}-\boldsymbol{\ell}_1)$ is one product of two maps, computed with fast Fourier transforms. The estimator becomes $\hat\psi_{\boldsymbol{L}} = -iA_L\, \boldsymbol{L}\cdot\mathcal{F}\left[\boldsymbol{\nabla}(C\bar{T})\,\bar{T}\right]$ with the filtered map $\bar{T} = \hat{T}/C^{\rm tot}$, and $1/A_L$ of (T1.24) splits the same way into six products after expanding f^2. This gives $A_L = N_L^{(0)}$ exactly on the grid.
3. *$N^{(0)}$ by simulation.* The estimator applied to 400 Gaussian skies with spectrum $C^{\rm tot}$ and no lensing. By (T1.22) their mean spectrum is $N_L^{(0)}$.
4. *The raw spectrum.* The estimator applied to 400 lensed skies, with the spectrum of each estimate, its cross-spectrum with the input ψ and the spectrum of the input itself.
5. *$N^{(1)}$ by simulation, and by the integral.* Lensing two independent CMB skies a and b with one shared ψ isolates the cross-copy pairings. In the estimate $\hat\psi[a,b]$, which takes one mode of each pair from a and the other from b, the pairing inside a copy vanishes, because a and b are independent, and only the pairing that joins a to a and b to b survives; with symmetric weights it contributes half of $N^{(0)} + N^{(1)}$, and the swapped estimate $\hat\psi[b,a]$, which is equal, the other half. Lensing b with an independent potential removes the $N^{(1)}$ half, since then $\langle\psi\psi'\rangle = 0$. So $N^{(1)}$ is the mean of $2\left|\hat\psi[a,b]\right|^2$ with a shared potential minus the same with independent potentials, over 400 pairs. The same a and b enter both terms, so their chance products cancel in the difference and a few hundred pairs suffice [48]. Independently, (T1.28) is evaluated as a four-dimensional sum, with $\boldsymbol{\ell}_1$ on a polar grid and $\boldsymbol{\ell}_1 - \boldsymbol{\ell}_1'$ on a polar grid that is finer near zero, where $C^{\psi\psi}$ is largest.

T1.9.4 The answer

Figure T1.13 shows what the estimator delivers on one patch: the structures larger than about a degree are recovered, and finer detail is buried in noise. Figure T1.14 takes the spectrum apart, and every term of (T1.27) is where the derivation put it.

The estimator is unbiased. Averaged over the bins, the cross-spectrum of $\hat\psi$ with the input potential is 0.993 ± 0.002 times the input spectrum. So A_L of (T1.24), with the lensed weights, normalises the estimator to better than one per cent, although the skies were lensed to all orders and the derivation kept only the first. The small shortfall is what the lensed weights leave of the higher orders.

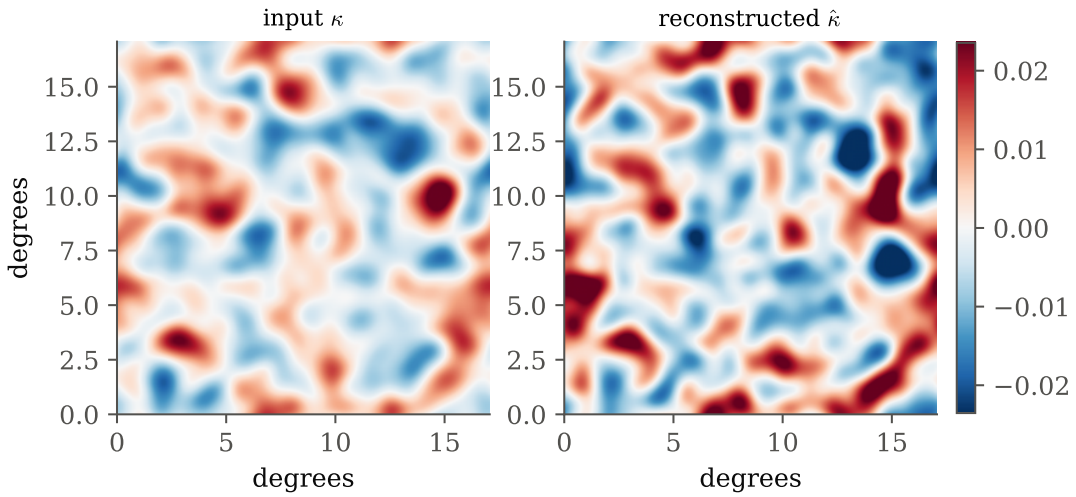


Figure T1.13: One simulated patch. Left: the input convergence $\kappa = \frac{1}{2}\nabla^2\psi$. Right: the convergence reconstructed by the quadratic estimator from the lensed, noisy temperature map. Both are filtered with the same weights $C_L^{\psi\psi}/(C_L^{\psi\psi} + N_L^{(0)})$, which suppress the multipoles where the reconstruction is worse than its noise, and smoothed with a Gaussian of 1° full width at half maximum. The structures of a degree and more agree; the finer ones in the reconstruction are mostly $N^{(0)}$.

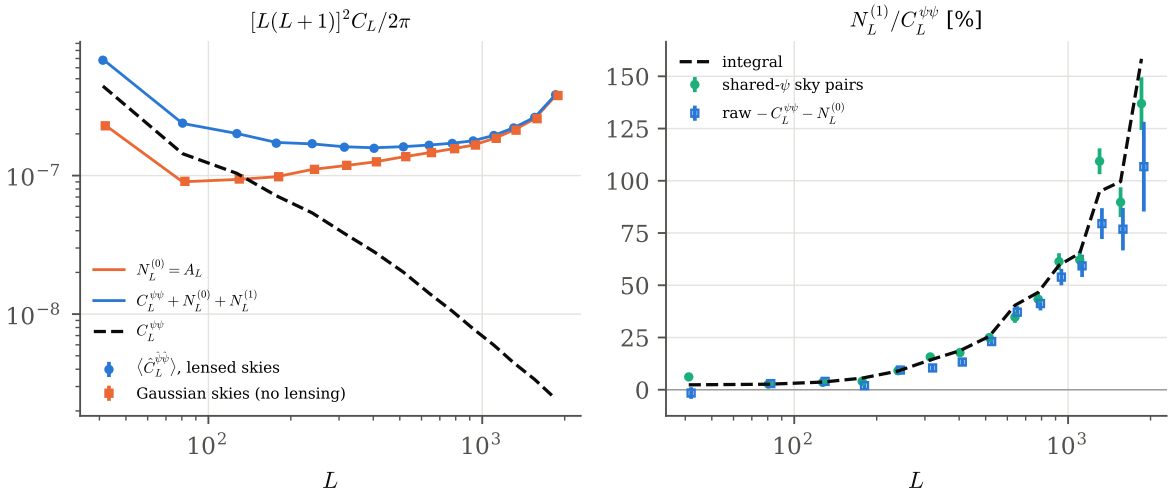


Figure T1.14: Left: the mean spectrum of the reconstructed potential over 400 lensed skies (blue dots) against the input $C_L^{\psi\psi}$ (black dashed), the noise $N_L^{(0)} = A_L$ of (T1.24) (orange line) and its simulation on Gaussian skies (orange squares), and the sum $C_L^{\psi\psi} + N_L^{(0)} + N_L^{(1)}$ (blue line). Right: $N_L^{(1)}$ as a percentage of $C_L^{\psi\psi}$, from pairs of skies sharing one potential (green dots), from the raw spectrum minus each sky's own input power and $N_L^{(0)}$ (open squares), and from the integral (T1.28) (black dashed).

$N^{(0)}$ is the normalisation. On the grid, A_L agrees with the integral of (T1.24) to 0.6% at single modes. On Gaussian skies with no lensing, the mean spectrum of the estimate matches A_L in all 15 bins, with $\chi^2 = 19.5$ for 15 degrees of freedom. It is smaller than the signal only on the largest scales: $N_L^{(0)}$ equals $C_L^{\psi\psi}$ near $L \approx 139$; at $L \approx 100$ it is 0.63 times the signal, and at $L \approx 1000$ it is 21.2 times. A single reconstructed mode is noise-dominated above $L \approx 139$, and the spectrum is measured there only by averaging many modes, as the shape-noise-dominated spectra of Section T1.8.4 were.

Where $N^{(0)}$ comes from. Recompute A_L with the noise removed from $C^{\rm tot}$: it drops by only 2% at $L \approx 100$ and 1% at $L \approx 1000$. At 5 μK-arcmin the instrument contributes almost nothing; the noise of the reconstruction is the chance pairing of the unlensed CMB with itself. This decides what can and cannot remove it. The trick of Section 8.3, crossing two maps with independent noise, removes the instrument's share of $N^{(0)}$ when the four modes are taken from four splits of the data [60]. It cannot remove the rest, because every split sees the same CMB. That part must be subtracted, which is why its formula and its simulation matter.

$N^{(1)}$ is there, and the integral gets it right. The pairs of skies sharing one potential detect $N^{(1)}$ at 44σ over the bins, and they agree with the integral (T1.28), $\chi^2 = 24.9$ for 15 degrees of freedom. As a fraction of the signal it grows with L: 2.6% of $C_L^{\psi\psi}$ at $L \approx 100$, 59.6% at $L \approx 1000$. Small next to $N^{(0)}$, it is not small next to the signal: an amplitude fitted to the spectrum with $N^{(1)}$ left in, weighting the bins by their Fisher information, would come out 16.3% too high. The raw spectrum tells the same story less cleanly. After subtracting each sky's own input power and A_L, what remains (open squares in Figure T1.14) follows the integral and stays within about three standard errors of it in every bin, but its χ^2 is 46.5 for 15 degrees of freedom, worse than the pairs give. The reason is the size of what was subtracted. Above $L \approx 1000$, $N^{(0)}$ is twenty to a hundred and fifty times the signal, so a few tenths of a per cent of $N^{(0)}$, from the higher orders of lensing or from the scatter of $C^{\rm tot}$, is as large as $N^{(1)}$ itself. A residual that is a small difference of large numbers inherits their errors. The pairs avoid that difference, because the same two CMB skies enter both of their terms.

So the spectrum of a reconstructed map is the signal plus two biases with the structure of (T1.27). $N^{(0)}$ is the Gaussian part of a four-point function, the same object as the noise bias of a power spectrum, and the best estimator makes it equal to the normalisation. $N^{(1)}$ is the part of the lensing trispectrum that joins the two copies of the estimate, proportional to $C^{\psi\psi}$ itself. Published analyses subtract both, with refinements the experiment did not need. $N^{(0)}$ is computed from the data's own spectrum, sky by sky (the realization-dependent $N^{(0)}$), which removes the sensitivity to a wrong $C^{\rm tot}$ and part of the scatter. $N^{(1)}$ comes from simulations sharing one potential, as in step 5. And the mean field of the mask and the uneven noise is subtracted first [48, 59, 61].

T1.9.5 Solving for the lensing spectrum: the answer on both sides

The two biases are not subtracted in the same way. $N^{(0)}$ is computed from the spectrum of the observed maps, and it does not contain $C^{\psi\psi}$, so subtracting it is one step. $N^{(1)}$ is different: by (T1.28) it is proportional to $C^{\psi\psi}$, the very spectrum we are trying to measure. The debiased spectrum is therefore defined by an equation with the unknown on both sides, the situation of Section T5.7. What does that equation look like, and how is it solved?

The kernel. In (T1.28) the potential enters only through its spectrum at one wavevector, $\boldsymbol{L}' \equiv \boldsymbol{\ell}_1 - \boldsymbol{\ell}_1'$. Group the terms by that wavevector:

$$N_L^{(1)} \overset{(1)}{=} \int \frac{\mathrm{d}^2 L'}{(2\pi)^2}\, \mathcal{K}(\boldsymbol{L}, \boldsymbol{L}')\, C_{L'}^{\psi\psi},$$

$$\mathcal{K}(\boldsymbol{L},\boldsymbol{L}') \overset{(2)}{=} 2A_L^2 \int \frac{\mathrm{d}^2\boldsymbol{\ell}_1}{(2\pi)^2} F(\boldsymbol{\ell}_1,\boldsymbol{\ell}_2)F(\boldsymbol{\ell}_1',\boldsymbol{\ell}_2')\, f(\boldsymbol{\ell}_1,-\boldsymbol{\ell}_1')\, f(\boldsymbol{\ell}_2,-\boldsymbol{\ell}_2'), \qquad \boldsymbol{\ell}_1' = \boldsymbol{\ell}_1 - \boldsymbol{L}',\ \boldsymbol{\ell}_2' = \boldsymbol{L} - \boldsymbol{\ell}_1',$$
$$N_b^{(1)} \overset{(3)}{=} \sum_{b'} K_{bb'}\, C_{b'}^{\psi\psi} + N_b^{(1),\mathrm{out}}. \tag{T1.29}$$

(1) In the double integral of (T1.28), change the variable of the inner integral from $\boldsymbol{\ell}_1'$ to $\boldsymbol{L}' = \boldsymbol{\ell}_1 - \boldsymbol{\ell}_1'$ at fixed $\boldsymbol{\ell}_1$; the shift has unit Jacobian, so $\mathrm{d}^2\boldsymbol{\ell}_1' = \mathrm{d}^2L'$.

(2) What multiplies $C_{L'}^{\psi\psi}$: the weights and the response f, which contain only the CMB spectra. The kernel $\mathcal{K}$ does not depend on $C^{\psi\psi}$ at all.

(3) Average over the L of one band b and split the L' integral into the bands b' in which the spectrum is measured. $K_{bb'}$ is the part from band b' divided by the band's value of $C^{\psi\psi}$, which is exact when the spectrum has the fiducial shape inside each band. The last term collects the potential outside the measured bands, computed with the fiducial spectrum and treated as known.

So $N^{(1)}$ is a fixed matrix acting on the unknown spectrum. Collect the band powers into a vector $\boldsymbol{c}$, and the measured spectrum minus $N^{(0)}$ and the outside part into $\boldsymbol{d}$, with components $d_b = \hat{C}_b^{\hat{\psi}\hat{\psi}} - N_b^{(0)} - N_b^{(1),\mathrm{out}}$. Then (T1.27) reads

$$\langle \boldsymbol{d} \rangle = (\mathbf{I} + \mathbf{K})\, \boldsymbol{c}, \qquad \text{solved by} \qquad \boldsymbol{c}_* = (\mathbf{I} + \mathbf{K})^{-1} \boldsymbol{d}. \tag{T1.30}$$

The equation is linear, because $N^{(1)}$ is linear in $C^{\psi\psi}$, and one matrix inversion solves it.

The same equation by iteration. Move the bias to the other side, $\boldsymbol{c} = \boldsymbol{d} - \mathbf{K}\boldsymbol{c}$, and replace the unknown inside the correction by the latest estimate:

$$\boldsymbol{c}^{(n+1)} = \boldsymbol{d} - \mathbf{K}\, \boldsymbol{c}^{(n)}, \qquad \boldsymbol{c}^{(0)} = \boldsymbol{c}_{\mathrm{fid}}. \tag{T1.31}$$

The first pass, $\boldsymbol{c}^{(1)} = \boldsymbol{d} - \mathbf{K}\boldsymbol{c}_{\mathrm{fid}}$, is exactly the usual recipe: subtract the $N^{(1)}$ that the fiducial spectrum predicts. Every further pass recomputes the bias with the corrected spectrum. Does this settle on $\boldsymbol{c}_*$? Subtract the equation that $\boldsymbol{c}_*$ satisfies, $\boldsymbol{c}_* = \boldsymbol{d} - \mathbf{K}\boldsymbol{c}_*$, and call $\boldsymbol{e}^{(n)} = \boldsymbol{c}^{(n)} - \boldsymbol{c}_*$ the error:

$$\boldsymbol{e}^{(n+1)} \overset{(1)}{=} -\mathbf{K}\, \boldsymbol{e}^{(n)}, \qquad \boldsymbol{e}^{(n)} \overset{(2)}{=} (-\mathbf{K})^n\, \boldsymbol{e}^{(0)} \overset{(3)}{=} \sum_k a_k (-\lambda_k)^n \boldsymbol{v}_k. \tag{T1.32}$$

(1) Subtract the two equations; $\boldsymbol{d}$ cancels.

(2) Apply the first result n times.

(3) Expand the starting error in the eigenvectors $\boldsymbol{v}_k$ of $\mathbf{K}$, $\boldsymbol{e}^{(0)} = \sum_k a_k \boldsymbol{v}_k$; each application of $-\mathbf{K}$ multiplies the k-th component by $-\lambda_k$.

Every component shrinks if every eigenvalue has $|\lambda_k| < 1$, that is if the **spectral radius** $\rho(\mathbf{K}) = \max_k |\lambda_k|$ is below one, and the error then falls by the factor ρ per pass. This is the contraction condition $|g'| < 1$ of Section T5.7.1, with the slope replaced by the matrix $\mathbf{K}$.

What our kernel does. The script computes $\mathbf{K}$ from the same polar sums that gave $N^{(1)}$ in Figure T1.14, now kept apart band by band (Figure T1.15, left). Its spectral radius is $\rho(\mathbf{K}) = 0.26$, and the iteration on the mean simulated spectrum shrinks its error by 0.26 per pass (right panel), as (T1.32) says; started from twice the fiducial spectrum it is within 10^{-3} of $\boldsymbol{c}_*$ after 8 passes. This is perhaps surprising, since $N^{(1)}$ exceeds the signal at the highest L (up to 158.4% of it). The kernel shows why. The large entries sit in the high-L rows, which draw on the potential at intermediate L', where the spectrum is large; the low-L rows, which would carry an error back, are weak. For an error to grow it must go round that loop, from high L back to low L and up again, and one leg of the loop is small. The size of the rows controls a single pass, the spectral radius controls many: started from half the fiducial spectrum, the largest relative error first grows, then falls at the rate ρ.

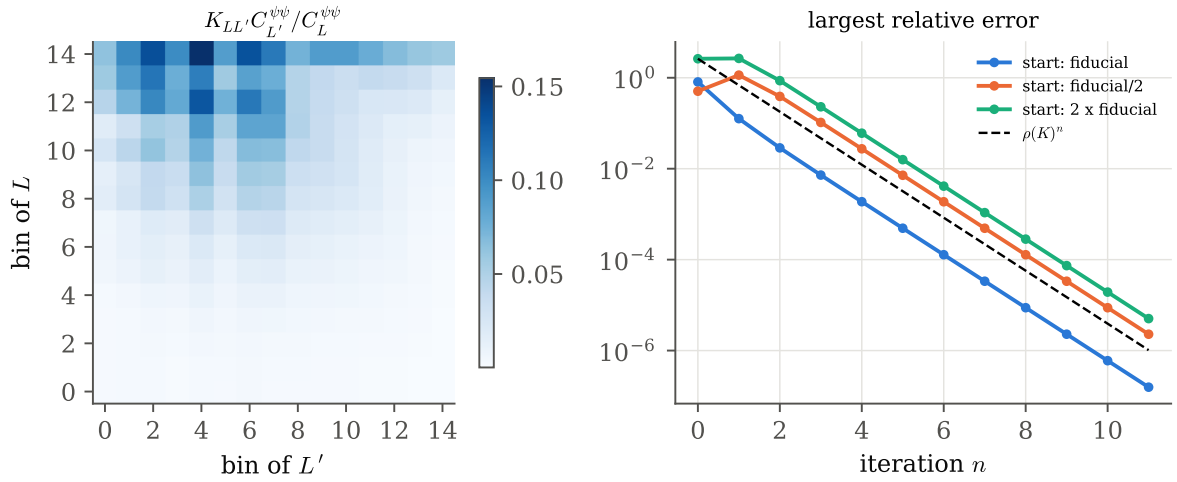


Figure T1.15: Left: the $N^{(1)}$ kernel of our simulated experiment, as the share of $N_L^{(1)}$ in band L (row) that comes from the potential in band L' (column), relative to $C_L^{\psi\psi}$; the rows add up to $N_L^{(1)}/C_L^{\psi\psi}$ apart from the part outside the bands. Right: the largest relative error of the iteration (T1.31) on the mean simulated spectrum, from three starting spectra, against the pass number; the dashed line falls by $\rho(\mathbf{K})$ per pass.

Is one pass enough? After the first pass the error is $-\mathbf{K}(\boldsymbol{c}_{\rm fid} - \boldsymbol{c}_*)$. A fiducial spectrum wrong by 10% leaves an error of up to 14.0% in the highest band, where $N^{(0)}$ is a hundred and fifty times the signal and the band carries almost no weight, and 1.6% on an amplitude fitted with the Fisher weights of Section T1.9.4; a second pass would cut that by ρ again. That is why Planck keeps to one pass: it subtracts $N^{(1)}$ computed in a fiducial model and lets the likelihood correct, to first order, for its dependence on the model. It also tested the direct inversion (T1.30) and found band powers that differ by a small fraction of their error [48]. Iterating on the data has a price the matrix form hides: $\boldsymbol{d}$ is noisy, and every pass feeds that noise back through $\mathbf{K}$, while a fiducial correction keeps the bias in the model, where it belongs. And the kernel itself depends on the assumed CMB spectra. If those are updated between passes as well, the iteration is a general fixed point whose convergence has to be checked, not assumed.

Recap

- Lensing distorts galaxy images by about 1%: $\kappa = \frac{1}{2}\nabla^2\psi$ magnifies, the spin-2 shear $\gamma_1 = \frac{1}{2}(\psi_{,11} - \psi_{,22})$, $\gamma_2 = \psi_{,12}$ stretches; a galaxy measures $\varepsilon \approx \varepsilon^{(0)} + \gamma$, with the reduced shear $\gamma/(1-\kappa)$ as its mean.
- Lensing makes only E modes; a B mode is a null test for systematics.
- $C_\ell^{\kappa\kappa}$ is the Limber integral of $[g/a]^2 P_\delta$ over distance, with the broad lensing efficiency $g(\chi)$; CAMB computes it with lensing source windows, and our own integral agrees to 0.1%.
- Cosmic shear measures $\sigma_8\Omega_m^\alpha$ with $\alpha \approx 0.55$ for a Stage-III sample (S_8 takes $\alpha = 0.5$); along that line the spectrum changes by a few per cent, across it by $\sim 20\%$ for a 9% change in S_8.
- Shape noise $N_i = \sigma_e^2/\bar{n}_i$ is white; for a KiDS-like survey the signal beats it only up to $\ell \approx 21$–194 per bin; total $S/N \approx 48$, $\sigma(\ln\sigma_8) = 0.8\%$ with everything else fixed, against $\approx 3\%$ achieved with nuisances marginalised.

- Nuisances: photo-z shifts, shear bias m, intrinsic alignments, baryonic feedback. CMB lensing (ψ reconstructed by a quadratic estimator) weights $z \approx 2$ and measures $\sigma_8\Omega_m^{0.25}$.
- The spectrum of a reconstructed lensing map is $C_L^{\psi\psi} + N_L^{(0)} + N_L^{(1)}$: $N^{(0)}$, the Gaussian part of its four-point function, equals the normalisation A_L of the best (Hu–Okamoto) weights and comes mostly from the CMB itself; $N^{(1)}$, from lensing pairings across the two copies, is proportional to $C^{\psi\psi}$ (2.6% of it at $L \approx 100$, 59.6% at $L \approx 1000$ in our setup). Both are confirmed by simulation; a mask adds a mean field. Because $N^{(1)}$ is linear in $C^{\psi\psi}$, the debiased spectrum solves $(\mathbf{I} + \mathbf{K})\boldsymbol{c} = \boldsymbol{d}$; iterating $\boldsymbol{c} \to \boldsymbol{d} - \mathbf{K}\boldsymbol{c}$ from a fiducial converges when $\rho(\mathbf{K}) < 1$ (here 0.26), and its first pass is the usual fiducial subtraction.

T1.10 Polarization: sticks on the sky, and two kinds of pattern

So far the microwave sky has been a map of one number per direction, the temperature. Its statistics are the $a_{\ell m}$ and their variance C_ℓ (Section T1.2), and one sky is one draw from them. That description already handles everything in the temperature map: its acoustic peaks, its parameters, its scatter. Weak lensing then showed us a sky whose value at each point is not a number but a short stick without an arrowhead, the shear of a galaxy image, and it gave us the rule for how such a stick is written and how it changes when the axes turn (Section T1.5).

The microwave light carries a stick too. It is partly linearly polarized: at each point of the sky the electric field oscillates a little more along one direction than along the one at right angles. A temperature map cannot tell left from right. A field of sticks can.

Picture two wheels. The first is an old cart wheel, its spokes running straight out from the hub. The second is a toy pinwheel, its blades all leaning the same way, so that it swirls. Hold each in front of a mirror. The cart wheel in the mirror is the same cart wheel. The pinwheel in the mirror swirls the other way: a left-handed pinwheel has become a right-handed one. Now take the cart wheel and twist every spoke by the same small angle at the hub. It has become a faint pinwheel, and a pinwheel of one particular hand, set by the sense of the twist. A smudge on the glass that blurs the spokes in random directions would also spoil the clean radial pattern, but it would not prefer one hand: averaged over many wheels, it makes as many left swirls as right ones.

The polarization of the microwave sky behaves in the same way. Its sticks can be split into a spoke-like part, called E, and a pinwheel-like part, called B. The ordinary physics that made the sky at last scattering does not know left from right, so it cannot produce a net handedness. If something on the way to us twists every stick by the same angle β, part of E turns into B, and the new B has a definite hand, which appears as a correlation between E and B. Gravitational lensing, which plays the part of the smudge, also turns E into B, but without a hand. Everything that follows is a search for that hand.

The twist in symbols. Take one stick of length P at an angle ψ from a fixed reference axis, and recall from Eq. (T1.5) that a headless stick is described by the complex number $Pe^{2i\psi}$, the factor 2 because the stick at $\psi + \pi$ is the same stick. Let the stick be a spoke: it points from a hub at position angle φ, so $\psi = \varphi$. Twist it by β, so $\psi = \varphi + \beta$, and ask how much of it is still a spoke and how much now leans at 45° to the spoke direction, the pinwheel direction. Measured from the spoke direction, the twisted stick is $Pe^{2i\beta} = P\cos 2\beta + iP\sin 2\beta$: a spoke part $P\cos 2\beta$ and a pinwheel part $P\sin 2\beta$. There are two dials, the twist β and the hand, which is the sign of β. The naive way to detect the twist is to look for pinwheel power. That power is $P^2\sin^2 2\beta$, which for a twist of a third of a degree is about 1.5×10^{-4} of the spoke power, and lensing makes pinwheel power of its own. The good choice is the product of the spoke part and the pinwheel part, $P^2\cos 2\beta \sin 2\beta = \frac{1}{2}P^2\sin 4\beta$. It is linear in β for small twists, it changes

sign with the hand, and lensing gives zero for it on average. Its price is that we must know which way our own reference axis points to better than β: an instrument whose axis is off by an angle α adds α to every ψ, and that is a twist too. For the Planck satellite the uncertainty of α is 0.28° [62], as large as the twist that has been reported. Lesson: a twist shows up first in a cross product that changes sign in a mirror, not in power, and the instrument's own orientation is the first thing it can be confused with.

A picture: gradient and curl. An ordinary vector field in a plane splits into a part that is the gradient of something, flowing straight out of hills and into valleys, and a part that is a curl, circulating around them. E and B are the same split for headless sticks. Around a hot or cold patch, E sticks line up along the radius or around the circle, the spokes and rims of a wheel; B sticks lean at 45° and swirl. A mirror leaves gradients alone and reverses circulation, so it leaves E alone and reverses B.

One idea, many names. A quantity that changes sign in a mirror is called a pseudoscalar, and a nonzero average of one means that the laws producing the data distinguish left from right. In particle physics this is a parity-violation test: in 1957 the electrons from polarized cobalt-60 nuclei were found to leave preferentially opposite to the nuclear spin, a correlation between a vector and an axial vector that a mirror reverses. In chemistry it is chirality, the handedness of a molecule, and in fluids it is helicity. In each case the statistic that detects the handedness is linear in a parity-odd quantity, and its expected value is zero when the laws are mirror symmetric. The mathematical language is that of how objects change under rotations and reflections, which is why spin weights and Fourier modes do all the work below.

We first write the stick in terms of the measured Stokes numbers and derive what a twist does to them, then split a field of sticks into E and B, show what a mirror does to each, and see why the ordinary sky has no EB and no TB. Section T1.11 then follows the twist and the smudge through to the power spectra.

T1.10.1 Stokes Q and U, and what a twist does to them

An instrument does not see sticks. It sees intensities. Fix two axes $\boldsymbol{e}_1, \boldsymbol{e}_2$ across the line of sight, and write the electric field of a wave as $a_1 e^{i\theta_1}$ along $\boldsymbol{e}_1$ and $a_2 e^{i\theta_2}$ along $\boldsymbol{e}_2$, with amplitudes a_1, a_2 and phases θ_1, θ_2. The cosmology notes define the four **Stokes parameters** as time averages,

$$I = \langle a_1^2 \rangle + \langle a_2^2 \rangle, \qquad Q = \langle a_1^2 \rangle - \langle a_2^2 \rangle, \qquad U = 2\langle a_1 a_2 \cos(\theta_1 - \theta_2) \rangle, \tag{T1.33}$$

plus a fourth, V, for circular polarization, which Thomson scattering does not produce and which we drop. I is the total intensity, which for the CMB we write as the temperature T. Q is the excess of intensity along $\boldsymbol{e}_1$ over $\boldsymbol{e}_2$, measured by a detector that compares the two. U is the same excess for a pair of axes turned by 45°.

Where is the stick? Take a wave linearly polarized along a direction at angle ψ from $\boldsymbol{e}_1$, with amplitude a. Its two components oscillate in step, $\theta_1 = \theta_2$, with $a_1 = a\cos\psi$ and $a_2 = a\sin\psi$. Then

$$\begin{aligned} Q &\overset{(1)}{=} a^2\cos^2\psi - a^2\sin^2\psi \overset{(2)}{=} a^2\cos 2\psi, \\ U &\overset{(3)}{=} 2a^2\cos\psi\sin\psi \overset{(4)}{=} a^2\sin 2\psi, \\ Q + iU &\overset{(5)}{=} P\, e^{2i\psi}, \qquad P \equiv \sqrt{Q^2+U^2} = a^2. \end{aligned} \tag{T1.34}$$

(1) Insert $a_1 = a\cos\psi$, $a_2 = a\sin\psi$ into the definition of Q; for a steady wave the time average does nothing.

(2) The double-angle identity $\cos^2\psi - \sin^2\psi = \cos 2\psi$.
(3) The definition of U with $\cos(\theta_1 - \theta_2) = \cos 0 = 1$.
(4) The double-angle identity $2\sin\psi\cos\psi = \sin 2\psi$.
(5) Euler's formula, $\cos 2\psi + i\sin 2\psi = e^{2i\psi}$. P is the polarized intensity, here all of I because the wave is fully polarized; an unpolarized part would add to I and nothing to Q or U.

So the measured pair (Q, U) is the stick of Section T1.5: its length is P, its angle is $\psi = \frac{1}{2}\arctan(U/Q)$ (taken in the right quadrant), and the factor two makes ψ and $\psi + 180°$ the same stick. Figure T1.16 is the dictionary: $Q > 0$ is a stick along $\boldsymbol{e}_1$, $Q < 0$ along $\boldsymbol{e}_2$, $U > 0$ and $U < 0$ the two diagonals.

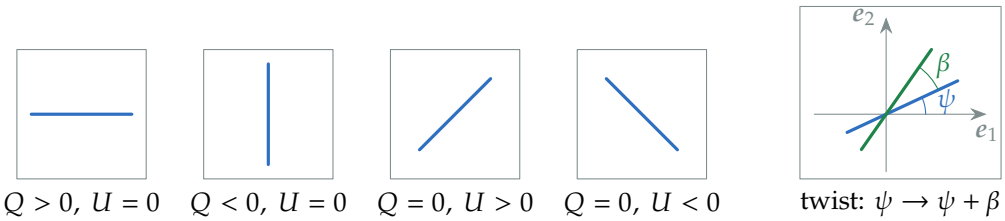


Figure T1.16: The Stokes dictionary. Each square shows one point of the sky with the polarization stick that the pair (Q, U) describes: along the first reference axis, along the second, and along the two diagonals. On the right, a stick at angle ψ (blue) is twisted by β (green); the twist adds β to ψ and so multiplies $Q + iU$ by $e^{2i\beta}$.

What a twist does. Suppose something turns the plane of polarization by an angle β, in the sense from $\boldsymbol{e}_1$ towards $\boldsymbol{e}_2$, without changing its strength. The stick at ψ becomes a stick at $\psi + \beta$ with the same P. Then

$$Q' + iU' \overset{(1)}{=} P\,e^{2i(\psi+\beta)} \overset{(2)}{=} e^{2i\beta}\,(Q + iU), \qquad Q' - iU' \overset{(3)}{=} e^{-2i\beta}\,(Q - iU),$$
$$\begin{pmatrix} Q' \\ U' \end{pmatrix} \overset{(4)}{=} \begin{pmatrix} \cos 2\beta & -\sin 2\beta \\ \sin 2\beta & \cos 2\beta \end{pmatrix} \begin{pmatrix} Q \\ U \end{pmatrix}. \tag{T1.35}$$

(1) (T1.34) for the twisted stick, whose angle is $\psi + \beta$ and whose length is unchanged.
(2) Split the exponential, $e^{2i(\psi+\beta)} = e^{2i\beta}e^{2i\psi}$, and use (T1.34) again for the untwisted stick.
(3) Q and U are real, so the complex conjugate of the first equation is the second.
(4) Write $e^{2i\beta} = \cos 2\beta + i\sin 2\beta$, multiply out $(\cos 2\beta + i\sin 2\beta)(Q + iU)$, and read off the real part Q' and the imaginary part U'.

The pair (Q, U) turns like a vector, but by twice the angle. A partly polarized wave obeys the same rule, because its unpolarized part sits in I and does not move, and every point of the sky obeys it with the same β if the cause of the twist is the same everywhere.

Example: three twists of one stick. Let a stick lie along $\boldsymbol{e}_1$, so $(Q, U) = (P, 0)$. A twist of $\beta = 45°$ gives $(Q', U') = (P\cos 90°, P\sin 90°) = (0, P)$: the stick is now on the diagonal and all of the signal has moved from Q to U. A twist of 90° gives $(-P, 0)$: the stick stands along $\boldsymbol{e}_2$, and Q has changed sign. A twist of $\beta = 0.35°$ gives $Q' = P\cos 0.70° = 0.99993\,P$ and $U' = P\sin 0.70° = 0.0122\,P$: a 1.2% leak of signal into U and a loss in Q of only 7×10^{-5}. Small twists move signal linearly in β and power quadratically.

A twist of the sticks or a turn of the axes? Equation (T1.5) said that turning the *axes* by α multiplies $Q + iU$ by $e^{-2i\alpha}$. That is the same stick described by a turned observer, so it changes the numbers and nothing physical. A twist turns the sticks themselves. The two are mirror images of each other in the formula, and that matters for an instrument: a detector whose

reference axis is turned by $-\alpha$ from where we believe it to be reads every stick as if it had been twisted by $+\alpha$. In the data a miscalibrated axis and a physical twist look identical. We return to this in Sections T1.11.6 and T1.13.3.

Watch out: Two sign conventions for U

The angle ψ can be measured from north towards east, as astronomers do (the IAU convention), or towards west, which is the convention of HEALPix and of the Planck maps. Changing convention flips the sign of U, and therefore the sign of every twist angle, of B, and of the EB and TB spectra built below. A twist reported as positive in one convention is negative in the other. Always check which convention a map and a paper use before comparing signs.

T1.10.2 E and B on a small patch of sky

The numbers Q and U at a point depend on the axes we chose there, so they are not a good description of a pattern: turn the axes and they change. We want two fields that describe the pattern itself. On a small patch of sky, treated as flat with angular coordinates $\boldsymbol{\theta} = (\theta_1, \theta_2)$ as in Section T1.5, the natural way is to take the pattern apart into waves. Write each field as a sum of plane waves, $f(\boldsymbol{\theta}) = \int \frac{\mathrm{d}^2\boldsymbol{\ell}}{(2\pi)^2} f(\boldsymbol{\ell})\, e^{i\boldsymbol{\ell}\cdot\boldsymbol{\theta}}$, where the wavevector $\boldsymbol{\ell}$ points at the angle $\varphi_{\boldsymbol{\ell}}$ from $\boldsymbol{e}_1$, and its length ℓ plays the part of the multipole.

Which axes does one wave pick? A single plane wave has one direction of its own, the direction of $\boldsymbol{\ell}$. So for that wave there is a natural pair of axes: the first along $\boldsymbol{\ell}$, the second across it. Read the Stokes numbers of the wave in those axes, which means turning the axes by $\varphi_{\boldsymbol{\ell}}$ and applying Eq. (T1.5). The two numbers we get are called E and B:

$$E(\boldsymbol{\ell}) + iB(\boldsymbol{\ell}) \overset{(1)}{=} e^{-2i\varphi_{\boldsymbol{\ell}}}\big[Q(\boldsymbol{\ell}) + iU(\boldsymbol{\ell})\big], \qquad E(\boldsymbol{\ell}) - iB(\boldsymbol{\ell}) \overset{(2)}{=} e^{2i\varphi_{\boldsymbol{\ell}}}\big[Q(\boldsymbol{\ell}) - iU(\boldsymbol{\ell})\big],$$
$$E(\boldsymbol{\ell}) \overset{(3)}{=} \cos 2\varphi_{\boldsymbol{\ell}}\, Q(\boldsymbol{\ell}) + \sin 2\varphi_{\boldsymbol{\ell}}\, U(\boldsymbol{\ell}), \qquad B(\boldsymbol{\ell}) \overset{(4)}{=} -\sin 2\varphi_{\boldsymbol{\ell}}\, Q(\boldsymbol{\ell}) + \cos 2\varphi_{\boldsymbol{\ell}}\, U(\boldsymbol{\ell}). \quad \text{(T1.36)}$$

(1) Turning the axes by $\varphi_{\boldsymbol{\ell}}$ multiplies $Q + iU$ by $e^{-2i\varphi_{\boldsymbol{\ell}}}$ (T1.5); this holds wave by wave because the turn is the same at every point of the wave.
(2) The same for $Q - iU$, which picks up the opposite phase. In Fourier space $Q(\boldsymbol{\ell})$ and $U(\boldsymbol{\ell})$ are complex, so the second line is a separate statement, not the complex conjugate of the first.
(3) Add the two equations of the first line and divide by two: $e^{\mp 2i\varphi}$ combine into $\cos 2\varphi$ and $\sin 2\varphi$.
(4) Subtract them and divide by $2i$.

In words: E is the part of each wave whose sticks lie along or across its wavevector, and B is the part whose sticks lean at 45° to it. The first row of Figure T1.17 shows one wave of each. This choice of axes uses nothing but the wave itself, so E and B do not depend on how we oriented $\boldsymbol{e}_1, \boldsymbol{e}_2$; they describe the pattern.

Are these the same E and B that weak lensing used, Section T1.6, where they were defined by $\Delta E = (\partial_1^2 - \partial_2^2)Q + 2\partial_1\partial_2 U$ and $\Delta B = -2\partial_1\partial_2 Q + (\partial_1^2 - \partial_2^2)U$ with γ_1, γ_2 in place of Q, U? On one wave each derivative ∂_a becomes a factor $i\ell_a$, with $\ell_1 = \ell\cos\varphi_{\boldsymbol{\ell}}$ and $\ell_2 = \ell\sin\varphi_{\boldsymbol{\ell}}$, so

$$-\ell^2 E(\boldsymbol{\ell}) \overset{(1)}{=} -(\ell_1^2 - \ell_2^2)\, Q(\boldsymbol{\ell}) - 2\ell_1\ell_2\, U(\boldsymbol{\ell}) \overset{(2)}{=} -\ell^2\big[\cos 2\varphi_{\boldsymbol{\ell}}\, Q(\boldsymbol{\ell}) + \sin 2\varphi_{\boldsymbol{\ell}}\, U(\boldsymbol{\ell})\big],$$
$$-\ell^2 B(\boldsymbol{\ell}) \overset{(1)}{=} 2\ell_1\ell_2\, Q(\boldsymbol{\ell}) - (\ell_1^2 - \ell_2^2)\, U(\boldsymbol{\ell}) \overset{(2)}{=} -\ell^2\big[-\sin 2\varphi_{\boldsymbol{\ell}}\, Q(\boldsymbol{\ell}) + \cos 2\varphi_{\boldsymbol{\ell}}\, U(\boldsymbol{\ell})\big]. \quad \text{(T1.37)}$$

(1) The Laplacian becomes $-\ell^2$, $\partial_1^2 - \partial_2^2$ becomes $-(\ell_1^2 - \ell_2^2)$, and $2\partial_1\partial_2$ becomes $-2\ell_1\ell_2$.

(2) $\ell_1^2 - \ell_2^2 = \ell^2(\cos^2\varphi - \sin^2\varphi) = \ell^2\cos 2\varphi$ and $2\ell_1\ell_2 = \ell^2 \sin 2\varphi$.

Dividing by $-\ell^2$ gives back (T1.36): the two definitions agree. The derivative form says that E and B at a point depend on Q and U all around it. One stick does not know whether it belongs to an E or a B pattern; only its neighbours do.

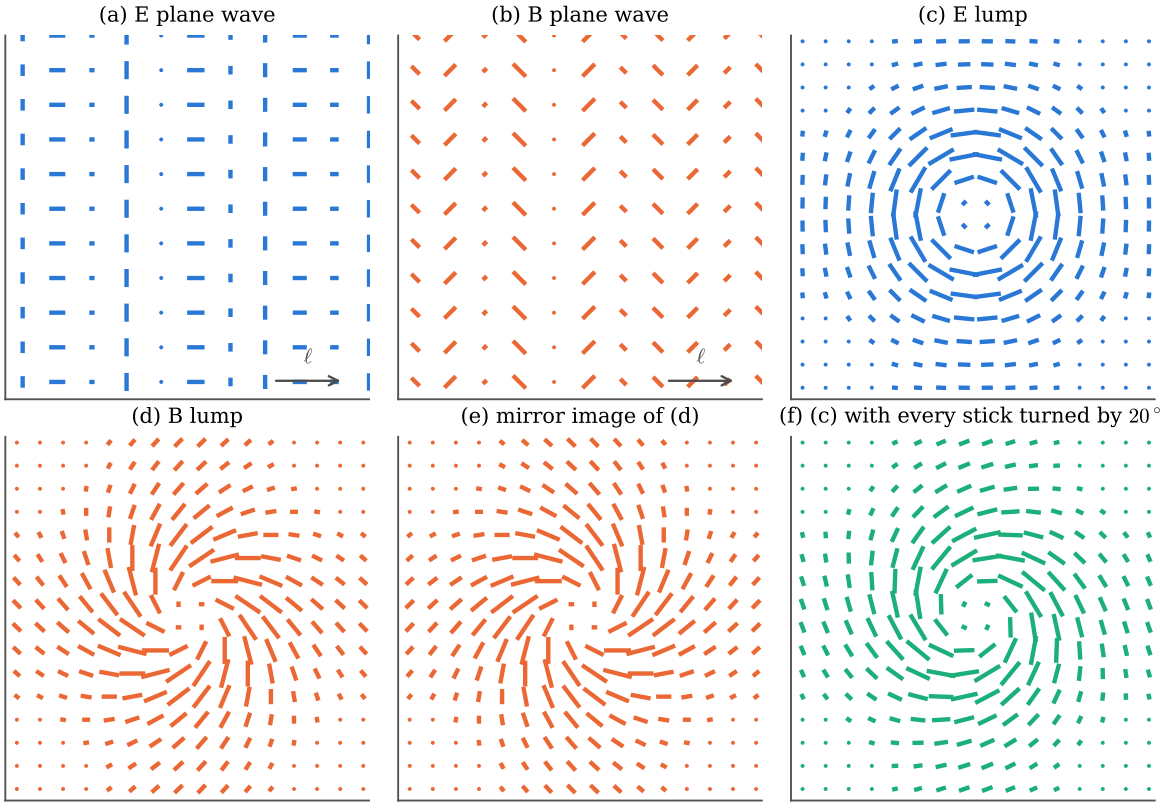


Figure T1.17: Pure E and pure B patterns as fields of sticks (stick length proportional to P). (a) One plane wave of E, wavevector along the first axis (arrow): the sticks lie along the wavevector where the wave is positive and across it where it is negative. (b) The same wave of B: the sticks lean at ±45°. (c) A localised lump of E: rings and spokes. (d) A lump of B: a swirl. (e) The mirror image of (d), reflected top to bottom: it swirls the other way, and it is exactly the B lump with its sign reversed. (f) The E lump of (c) with every stick twisted by 20°: it now swirls, and it is $\cos 40°$ of the E lump plus $\sin 40°$ of a B lump.

Example: what each lump is made of. The script `code/chT1/05_eb_patterns.py` builds the six panels of Figure T1.17 from (T1.36) run backwards, $Q + iU = e^{2i\varphi_\ell}(E + iB)$ wave by wave, and then recovers E and B from the sticks. For the twisted lump of panel (f) the recovered fields are $\cos 40°$ times the original E and $\sin 40°$ times it as B, to a relative error of 2.600×10^{-11}, round-off. That is the rule (T1.35) at work on a whole pattern, and Section T1.11 derives it in general.

Working definition T1.1 (E and B modes)

For each Fourier wave of a polarization pattern, E and B are the Stokes parameters Q and U read in axes aligned with the wavevector: $E(\boldsymbol{\ell}) + iB(\boldsymbol{\ell}) = e^{-2i\varphi_\ell}[Q(\boldsymbol{\ell}) + iU(\boldsymbol{\ell})]$. E sticks lie along or across the wavevector, B sticks at 45° to it. *Operational:* Fourier-transform the Q and U maps, combine them with (T1.36), and on the full sky use the spin-2 harmonic transform of `healpy` (`map2alm` with T, Q, U), which returns $T_{\ell m}, E_{\ell m}, B_{\ell m}$.

T1.10.3 The whole sky

On the whole sphere the plane waves are replaced by spherical harmonics, and the stick-carrying fields need their spin-weighted version. Spin-weighted harmonics ${}_{\pm2}Y_{\ell m}$ are the functions that turn by $e^{\mp 2i\psi}$ when the local axes turn by ψ, just as $Q \pm iU$ do, so they are the natural basis for $Q \pm iU$; they are built from the ordinary $Y_{\ell m}$ by two derivatives (see the cosmology notes, and [63, 64]). With the sign convention of the cosmology notes and of [63],

$$(Q \pm iU)(\boldsymbol{n}) = -\sum_{\ell\geq 2}\sum_{m=-\ell}^{\ell} \left(E_{\ell m} \pm iB_{\ell m}\right) {}_{\pm2}Y_{\ell m}(\boldsymbol{n}). \tag{T1.38}$$

The sum starts at $\ell = 2$ because a stick pattern has no monopole or dipole. On small scales this becomes the flat-sky (T1.36), wave by wave, and the overall minus sign is a convention that the flat-sky form hides in its Fourier phases. The temperature has its own coefficients $T_{\ell m}$, the $a_{\ell m}$ of Section T1.2.

The three sets $T_{\ell m}$, $E_{\ell m}$, $B_{\ell m}$ are Gaussian random numbers with mean zero, like the $a_{\ell m}$ of Chapter 7, but now they can be correlated with each other at the same (ℓ, m). For any two of them, $X, Y \in \{T, E, B\}$, the **angular cross-spectrum** is

$$\langle X_{\ell m} Y^*_{\ell' m'}\rangle = \delta_{\ell\ell'}\delta_{mm'}\, C_\ell^{XY}, \qquad \widehat{C}_\ell^{XY} = \frac{1}{2\ell+1}\sum_{m=-\ell}^{\ell} \mathrm{Re}\big(X_{\ell m} Y^*_{\ell m}\big). \tag{T1.39}$$

There are six: TT, EE, BB, TE, EB and TB. On a flat patch the same definition reads $\langle X(\boldsymbol{\ell}) Y^*(\boldsymbol{\ell}')\rangle = (2\pi)^2\delta^2(\boldsymbol{\ell}-\boldsymbol{\ell}')\, C_\ell^{XY}$. The CAMB fiducial of `code/camb_fiducial.py` returns four of them, TT, EE, BB and TE, and none of EB or TB. The next subsection says why.

T1.10.4 What a mirror does

The event. The fiducial spectra contain no EB and no TB. *What would it take to produce them?* The answer comes from the mirror test of the opening, now done with the formulas.

Reflect the patch of sky in the line $\theta_2 = 0$: the mirror sends the point (θ_1, θ_2) to $(\theta_1, -\theta_2)$. A stick at angle ψ appears in the mirror at $-\psi$. By (T1.34) that leaves $Q = P\cos 2\psi$ unchanged and flips $U = P\sin 2\psi$. So the mirrored sky has $Q^{\mathrm{M}}(\theta_1,\theta_2) = Q(\theta_1,-\theta_2)$, $U^{\mathrm{M}}(\theta_1,\theta_2) = -U(\theta_1,-\theta_2)$ and $T^{\mathrm{M}}(\theta_1,\theta_2) = T(\theta_1,-\theta_2)$. In Fourier space a field flipped in θ_2 has its waves flipped in ℓ_2, so with $\bar{\boldsymbol{\ell}} \equiv (\ell_1, -\ell_2)$, whose angle is $\varphi_{\bar\ell} = -\varphi_\ell$,

$$\begin{aligned}
E^{\mathrm{M}}(\boldsymbol{\ell}) &\overset{(1)}{=} \cos 2\varphi_\ell\, Q(\bar{\boldsymbol{\ell}}) - \sin 2\varphi_\ell\, U(\bar{\boldsymbol{\ell}}) \overset{(2)}{=} \cos 2\varphi_{\bar\ell}\, Q(\bar{\boldsymbol{\ell}}) + \sin 2\varphi_{\bar\ell}\, U(\bar{\boldsymbol{\ell}}) \overset{(3)}{=} E(\bar{\boldsymbol{\ell}}),\\
B^{\mathrm{M}}(\boldsymbol{\ell}) &\overset{(1)}{=} -\sin 2\varphi_\ell\, Q(\bar{\boldsymbol{\ell}}) - \cos 2\varphi_\ell\, U(\bar{\boldsymbol{\ell}}) \overset{(2)}{=} -\big[-\sin 2\varphi_{\bar\ell}\, Q(\bar{\boldsymbol{\ell}}) + \cos 2\varphi_{\bar\ell}\, U(\bar{\boldsymbol{\ell}})\big] \overset{(3)}{=} -B(\bar{\boldsymbol{\ell}}),\\
T^{\mathrm{M}}(\boldsymbol{\ell}) &= T(\bar{\boldsymbol{\ell}}).
\end{aligned} \tag{T1.40}$$

(1) (T1.36) applied to the mirrored maps: $Q^{\mathrm{M}}(\boldsymbol{\ell}) = Q(\bar{\boldsymbol{\ell}})$ and $U^{\mathrm{M}}(\boldsymbol{\ell}) = -U(\bar{\boldsymbol{\ell}})$.
(2) $\varphi_{\bar\ell} = -\varphi_\ell$, so $\cos 2\varphi_\ell = \cos 2\varphi_{\bar\ell}$ and $\sin 2\varphi_\ell = -\sin 2\varphi_{\bar\ell}$.
(3) (T1.36) again, now at the wavevector $\bar{\boldsymbol{\ell}}$.

A mirror carries E into E and B into $-B$, as panels (d) and (e) of Figure T1.17 show; the script finds the mirrored B lump equal to minus the original to a relative error of 3.400×10^{-11}. On the whole sky the corresponding statement is about the reflection through the centre, $\boldsymbol{n} \to -\boldsymbol{n}$, under which $T_{\ell m} \to (-1)^\ell T_{\ell m}$, $E_{\ell m} \to (-1)^\ell E_{\ell m}$ and $B_{\ell m} \to (-1)^{\ell+1} B_{\ell m}$ [62, 64]: again B has the opposite sign to E and T. E is called a scalar and B a **pseudoscalar**.

Now the spectra. Suppose the laws that produced the sky do not distinguish left from right. Then the mirror image of our sky is exactly as probable as our sky itself, so its spectra are the same as ours. But by (T1.40) its cross-spectra are

$$C_\ell^{EB,\mathrm{M}} \overset{(1)}{=} -C_\ell^{EB}, \qquad C_\ell^{TB,\mathrm{M}} \overset{(1)}{=} -C_\ell^{TB}, \qquad C_\ell^{XY,\mathrm{M}} \overset{(2)}{=} C_\ell^{XY} \quad (XY = TT, EE, BB, TE),$$
$$C_\ell^{EB} \overset{(3)}{=} -C_\ell^{EB} \implies C_\ell^{EB} = 0, \qquad C_\ell^{TB} = 0. \tag{T1.41}$$

(1) Insert $E^{\mathrm{M}}(\boldsymbol{\ell}) = E(\bar{\boldsymbol{\ell}})$, $B^{\mathrm{M}}(\boldsymbol{\ell}) = -B(\bar{\boldsymbol{\ell}})$, $T^{\mathrm{M}}(\boldsymbol{\ell}) = T(\bar{\boldsymbol{\ell}})$ into $\langle X(\boldsymbol{\ell}) Y^*(\boldsymbol{\ell}')\rangle$; one minus sign comes out, and $\bar{\boldsymbol{\ell}}$ has the same length ℓ as $\boldsymbol{\ell}$, so the spectrum is evaluated at the same multipole.
(2) In BB the minus sign appears twice, in the others not at all.
(3) A mirror-symmetric law makes the mirrored sky as probable as ours, so its spectra equal ours; with the first line, C^{EB} equals its own negative.

Four of the six spectra survive the mirror, and two, EB and TB, change sign. In a universe whose physics treats left and right alike, the two parity-odd spectra vanish on average. On one sky they are not exactly zero: $\widehat{C}_\ell^{EB}$ scatters around zero from chance alignments, by an amount set by C^{EE} and C^{BB}, just as $\widehat{C}_\ell$ scatters around C_ℓ.

Key idea: EB and TB detect handedness

Under a mirror T and E keep their sign and B reverses. A universe whose laws are mirror symmetric therefore has $C_\ell^{EB} = C_\ell^{TB} = 0$ on average. A nonzero EB or TB on the sky means that something between the emission and our detector picked a hand: a twist of the polarization by new physics, a twist by our own instrument, or a source whose own structure is handed.

T1.10.5 The same mirror, written with matrices

The mirror of Section T1.10.4 was worked out by following one stick: its angle ψ became $-\psi$, so $Q = P\cos 2\psi$ stayed and $U = P\sin 2\psi$ flipped. The same answer comes out of a short matrix calculation, and that calculation does more: it applies unchanged to the strain of a gravitational wave, and it shows which statements about even and odd are properties of a pattern and which are properties of the axes we happened to choose.

The mirror as a matrix. Put the line of sight along z and the two sky axes $\boldsymbol{e}_1, \boldsymbol{e}_2$ along x and y. The mirror $\theta_2 \to -\theta_2$ is the reflection $(x, y, z) \to (x, -y, z)$, which acts on the components of any arrow as the matrix

$$\mathsf{P}_T = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \mathsf{P}_T^{\mathsf{T}} = \mathsf{P}_T, \qquad \mathsf{P}_T^2 = \mathbb{1}. \tag{T1.42}$$

The subscript T says that it reflects a transverse direction, across the line of sight, and leaves the direction of travel alone. Reflecting twice changes nothing, which is $\mathsf{P}_T^2 = \mathbb{1}$.

How does a tensor feel the mirror? The simplest tensor is a product of two arrows. Take the electric field of a wave, an arrow with components E_a, and form the table of products $T_{ab} = \langle E_a E_b\rangle$. In the mirror every arrow becomes $\boldsymbol{E}' = \mathsf{P}_T \boldsymbol{E}$, so

$$T'_{ab} \overset{(1)}{=} \langle E'_a E'_b\rangle \overset{(2)}{=} \sum_{c,d} (\mathsf{P}_T)_{ac} (\mathsf{P}_T)_{bd} \langle E_c E_d\rangle \overset{(3)}{=} \left(\mathsf{P}_T\, T\, \mathsf{P}_T^{\mathsf{T}}\right)_{ab},$$
$$T'_{ab} \overset{(4)}{=} s_a s_b\, T_{ab}, \qquad (s_x, s_y, s_z) = (1, -1, 1). \tag{T1.43}$$

(1) The mirrored table is built from the mirrored arrow.
(2) Write each component of $\boldsymbol{E}' = \mathsf{P}_T \boldsymbol{E}$ as a sum, $E'_a = \sum_c (\mathsf{P}_T)_{ac} E_c$; the matrix entries are fixed numbers and come out of the average.
(3) The double sum is the matrix product, with the second factor transposed because its index b sits on the left of $(\mathsf{P}_T)_{bd}$.
(4) P_T is diagonal with entries s_a, so only $c = a$ and $d = b$ survive.

The rule $T' = \mathsf{P}_T T \mathsf{P}_T^{\mathsf{T}}$ holds for every two-index object that turns like a product of two arrows, which includes the strain of space. The last line says it in words: an entry keeps its sign if it carries an even number of y indices and flips if it carries one.

Example: the two polarizations of a gravitational wave. A gravitational wave travelling along z stretches and squeezes space across its direction of travel. Its strain is $h_{ij} = h_+ e^+_{ij} + h_\times e^\times_{ij}$, with the two polarization tensors

$$e^+ = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad e^\times = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{T1.44}$$

The + state stretches along x while it squeezes along y; the × state does the same along the two diagonals (see the cosmology notes). In the mirror,

$$\begin{aligned} \mathsf{P}_T\, e^+\, \mathsf{P}_T^{\mathsf{T}} &\overset{(1)}{=} \begin{pmatrix} 1\cdot 1\cdot 1 & 0 & 0 \\ 0 & (-1)(-1)(-1) & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix} = e^+, \\ \mathsf{P}_T\, e^\times\, \mathsf{P}_T^{\mathsf{T}} &\overset{(2)}{=} \begin{pmatrix} 0 & (1)(1)(-1) & 0 \\ (-1)(1)(1) & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = -e^\times, \\ \begin{pmatrix} h_+ \\ h_\times \end{pmatrix} &\overset{(3)}{=} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} h_+ \\ h_\times \end{pmatrix}. \end{aligned} \tag{T1.45}$$

(1) (T1.43) entry by entry: each diagonal entry e^+_{aa} is multiplied by $s_a^2 = 1$, so nothing changes.
(2) The only nonzero entries of $e^\times$ are xy and yx, and each carries one y index, so each is multiplied by $s_x s_y = -1$.
(3) The mirrored strain is $h_+ e^+ - h_\times e^\times$: written in the same two tensors, the mirror keeps the amount h_+ and reverses the amount $h_\times$.

Figure T1.18 shows why. A ring of free particles squeezed into an ellipse along x looks the same in a mirror that flips y. An ellipse along the diagonal from lower left to upper right turns, in that mirror, into an ellipse along the other diagonal, which is the × pattern with the opposite sign.

Example: the Stokes numbers as a tensor. The polarization of the microwave light is a two-index object too. The field across the line of sight has the two complex components $E_1 = a_1 e^{i\theta_1}$ and $E_2 = a_2 e^{i\theta_2}$ of (T1.33). The real part of the table of averages $\langle E_a E_b^* \rangle$ is

$$\begin{aligned} \mathrm{Re} \begin{pmatrix} \langle E_1 E_1^* \rangle & \langle E_1 E_2^* \rangle \\ \langle E_2 E_1^* \rangle & \langle E_2 E_2^* \rangle \end{pmatrix} &\overset{(1)}{=} \begin{pmatrix} \langle a_1^2 \rangle & \langle a_1 a_2 \cos(\theta_1 - \theta_2) \rangle \\ \langle a_1 a_2 \cos(\theta_1 - \theta_2) \rangle & \langle a_2^2 \rangle \end{pmatrix} \overset{(2)}{=} \frac{1}{2} \begin{pmatrix} I + Q & U \\ U & I - Q \end{pmatrix} \\ &\overset{(3)}{=} \frac{I}{2} \mathbb{1} + \mathcal{P}, \qquad \mathcal{P} \equiv \frac{1}{2} \begin{pmatrix} Q & U \\ U & -Q \end{pmatrix} \overset{(4)}{=} \frac{Q}{2} e^+ + \frac{U}{2} e^\times. \end{aligned} \tag{T1.46}$$

(1) $E_a E_b^* = a_a a_b e^{i(\theta_a - \theta_b)}$, whose real part is $a_a a_b \cos(\theta_a - \theta_b)$; on the diagonal the phases cancel.

(2) The definitions (T1.33): $\langle a_1^2\rangle = \frac{1}{2}(I+Q)$, $\langle a_2^2\rangle = \frac{1}{2}(I-Q)$, and $\langle a_1 a_2 \cos(\theta_1-\theta_2)\rangle = \frac{1}{2}U$.
(3) Split off the part proportional to the unit matrix; what is left has zero trace.
(4) Compare entry by entry with the 2×2 blocks of (T1.44) that lie across the line of sight.

(The imaginary part of the table, which is antisymmetric, holds the circular polarization V that Thomson scattering does not produce.) The traceless part $\mathcal{P}_{ab}$ is built from e^+ and $e^\times$ exactly as the strain is, with $Q/2$ in place of h_+ and $U/2$ in place of $h_\times$. So the calculation (T1.45) can be copied line by line: the mirror sends $(Q,U)\to(Q,-U)$ at the mirrored point, and the unit-matrix part, the intensity I, does not change. This is what the stick argument of Section T1.10.4 found.

Example: E and B. By (T1.36), E and B are Q and U turned through the angle $2\varphi_{\boldsymbol{\ell}}$ of each wavevector:

$$\begin{pmatrix} E \\ B \end{pmatrix}(\boldsymbol{\ell}) = R(-2\varphi_{\boldsymbol{\ell}}) \begin{pmatrix} Q \\ U \end{pmatrix}(\boldsymbol{\ell}), \qquad R(\vartheta) \equiv \begin{pmatrix} \cos\vartheta & -\sin\vartheta \\ \sin\vartheta & \cos\vartheta \end{pmatrix}. \tag{T1.47}$$

Write $D = \mathrm{diag}(1,-1)$ for the action of the mirror on the pair (Q,U). The mirror does two things to a wave: it flips U, and it turns the wavevector $\boldsymbol{\ell}$ into $\bar{\boldsymbol{\ell}} = (\ell_1, -\ell_2)$, whose angle is $\varphi_{\bar{\boldsymbol{\ell}}} = -\varphi_{\boldsymbol{\ell}}$. Then

$$\begin{aligned} \begin{pmatrix} E^{\mathrm{M}} \\ B^{\mathrm{M}} \end{pmatrix}(\boldsymbol{\ell}) &\overset{(1)}{=} R(-2\varphi_{\boldsymbol{\ell}})\, D \begin{pmatrix} Q \\ U \end{pmatrix}(\bar{\boldsymbol{\ell}}) \overset{(2)}{=} D\, R(2\varphi_{\boldsymbol{\ell}}) \begin{pmatrix} Q \\ U \end{pmatrix}(\bar{\boldsymbol{\ell}}) \\ &\overset{(3)}{=} D\, R(-2\varphi_{\bar{\boldsymbol{\ell}}}) \begin{pmatrix} Q \\ U \end{pmatrix}(\bar{\boldsymbol{\ell}}) \overset{(4)}{=} D \begin{pmatrix} E \\ B \end{pmatrix}(\bar{\boldsymbol{\ell}}). \end{aligned} \tag{T1.48}$$

(1) (T1.47) applied to the mirrored maps, whose Fourier coefficients at $\boldsymbol{\ell}$ are $Q(\bar{\boldsymbol{\ell}})$ and $-U(\bar{\boldsymbol{\ell}})$.
(2) A reflection followed by a turn equals the opposite turn followed by the reflection, $R(\vartheta)D = DR(-\vartheta)$. Check it by multiplying: both sides equal $\begin{pmatrix} \cos\vartheta & \sin\vartheta \\ \sin\vartheta & -\cos\vartheta \end{pmatrix}$.
(3) $2\varphi_{\boldsymbol{\ell}} = -2\varphi_{\bar{\boldsymbol{\ell}}}$.
(4) (T1.47) again, now at the wavevector $\bar{\boldsymbol{\ell}}$.

E and B transform with the same matrix D as (Q,U) and $(h_+,h_\times)$: E is even and B is odd. This is (T1.40) in two lines. The reason it works is visible in step 2: the mirror flips U and also flips the angle of the wave, and the two flips meet so that E and B come out with D again.

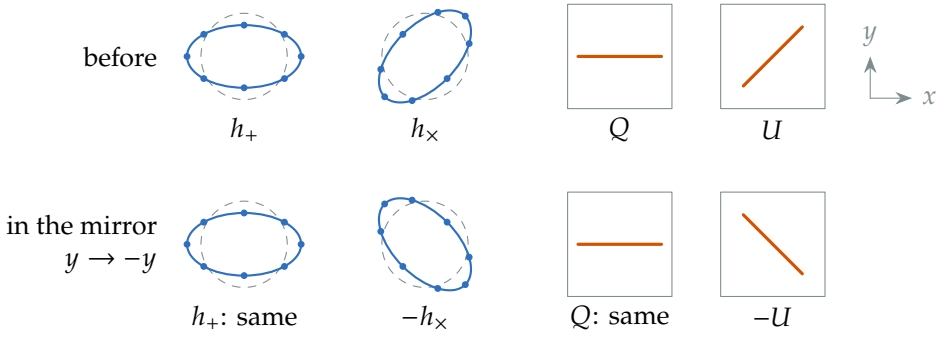


Figure T1.18: One mirror, four objects. Top: a ring of free particles (dashed) deformed by the + and the × polarization of a gravitational wave, and a polarization stick with only Q and with only U. Bottom: the same after the reflection $y\to -y$. The ellipse along x and the stick along x are unchanged. The ellipse and the stick along a diagonal move to the other diagonal, which is the same pattern with its amount reversed.

All in one table. Under the mirror P_T:

	even (sign kept)	odd (sign reversed)
gravitational-wave strain	h_+	$h_\times$
Stokes numbers at a point	$I = T, Q$	U
patterns on the sky	T, E	B

The same matrix is not the same quantity. The table lines up Q, E and h_+ because the mirror acts on each pair with the same matrix $\mathrm{diag}(+1, -1)$. That is a statement about the even and odd structure, and nothing more: Q is not E, and h_+ is not E. The difference shows as soon as we use another mirror. Reflect instead in the diagonal line $x = y$, which swaps the two sky axes: its transverse block is $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, and applied to the blocks of (T1.44) it gives $e^+ \to -e^+$ and $e^\times \to e^\times$ (swapping x and y swaps the two diagonal entries of e^+ and leaves $e^\times$ as it was). For that mirror Q and h_+ are odd, and U and $h_\times$ are even. So whether Q is even or odd depends on how the mirror line sits relative to the axes we chose: Q, U and $h_+, h_\times$ are components of one spin-2 tensor in a chosen basis, and the basis decides which component a given mirror flips. E and B were built differently, by reading each wave in axes set by the wave itself, so the answer does not depend on the mirror: every mirror line is the reflection $y \to -y$ seen from turned axes, and E, B do not care about the turn of the axes. E is even and B is odd for every mirror, which is why they, and not Q and U, are the quantities with a definite parity. Nor does the table say that gravitational waves make only E: h_+ and $h_\times$ are the strain in the frame of the wave, and the polarization patterns a gravitational wave leaves on the sky contain both E and B (Section T1.10.6).

Why it matters. A correlation of two quantities picks up the product of their signs. EE, BB, TT and TE multiply even by even or odd by odd and are unchanged by the mirror. EB and TB multiply an even quantity by an odd one and change sign, which is the content of (T1.41). Only physics that tells left from right can make them nonzero on average.

T1.10.6 Where the E modes come from, and how large they are

The CMB is polarized because of Thomson scattering at the last-scattering surface: an electron illuminated more strongly from one direction than from the direction at right angles (a quadrupole in the radiation it sees) scatters light that is linearly polarized [65]; the physics is in the cosmology notes. The quadrupole comes from the flows of the primordial plasma, and a second, smaller dose of polarization was added at reionization, at much larger angular scales.

Which patterns can a density wave produce? A single scalar wave, a plane wave of density in the early universe, has one special direction, its wavevector. Reflect the whole set-up in any plane that contains the wavevector and the line of sight, and nothing changes. By (T1.40) the B part of the pattern it produces must then equal its own negative, so it is zero. Scalar perturbations make only E [64, 66]. A gravitational wave is different: its two polarization states stretch space along crossed axes, and the $\times$ state is not symmetric under that reflection (it changes sign, Eq. (T1.45)), so tensor perturbations make both E and B. This is why B modes became the target of searches for primordial gravitational waves, and why the next sections ask what else can make them.

The fiducial spectra fix the sizes. $D_\ell^{EE} = \ell(\ell+1)C_\ell^{EE}/2\pi$ peaks at $42.4\,\mu\mathrm{K}^2$ near $\ell \approx 990$, so the polarization fluctuations are roughly a tenth of the temperature ones in amplitude. The TE spectrum, with both signs, connects the two. The fiducial BB is not zero, but it contains no primordial gravitational waves: it is the lensing B of Section T1.11.4, which peaks at $D_\ell^{BB} = 0.101\,\mu\mathrm{K}^2$ near $\ell \approx 886$, four hundred times below EE.

T1.11 When power moves from E to B

We now have the rule for one stick, (T1.35): a twist by β multiplies $Q + iU$ by $e^{2i\beta}$. And we have the split of a field of sticks into E and B, with the six spectra C_ℓ^{XY} of (T1.39). Together these rules give the rotated pattern in panel (f) of Figure T1.17. To connect this construction to data, we need the effect of the rotation on the measured spectra. Their response tells us how to distinguish a coherent rotation from other processes that convert E into B.

Think of two radio stations on neighbouring frequencies, a loud one and a nearly silent one. A slightly mistuned receiver lets a little of the loud station into the quiet channel. You hear the leak as extra sound in the quiet channel, but there is a better test than loudness: play the loud station's programme next to the leak and they match, word for word and with the same sign. Now suppose instead the antenna wobbles in the wind, so that the loud programme is added to the quiet channel sometimes with a plus sign and sometimes with a minus sign. You still hear extra sound, as loud as before. But the match with the loud programme averages to zero, because the plus and minus moments cancel. Loudness cannot tell the two leaks apart; the correlation with the source can.

On the sky the loud station is E and the quiet one is B. A twist of every stick by the same angle is the mistuned receiver: it moves a fixed fraction of E into B with a fixed sign, so the new B is correlated with E (and with T, which is correlated with E). Gravitational lensing is the wobbling antenna: it moves E into B through the lensing potential, which is positive in some places and negative in others, so it adds BB power but leaves no EB and no TB on average.

The two leaks in symbols. For one wave, write the mistuned receiver as $B' = sE$ with a fixed number s, and the wobbling antenna as $B' = wE$ with a random w of mean zero and variance σ_w^2, independent of E. The extra power is $\langle |B'|^2 \rangle = s^2 C^{EE}$ in the first case and $\sigma_w^2 C^{EE}$ in the second: equal when $\sigma_w = |s|$, so power cannot tell them apart. The correlation is $\langle B' E^* \rangle = s\, C^{EE}$ in the first case and $\langle w \rangle C^{EE} = 0$ in the second. The dial is the mean of the leak coefficient: a fixed leak has one, a random one does not. Lesson: to identify a twist, measure the correlation EB, not the power BB.

A picture. Plot every mode of the sky as a point in a plane, its E along the horizontal axis and its B along the vertical one. Before any leak the cloud of points is a thin horizontal cigar. A twist turns every point by the same angle, so the cigar tilts, and a tilted cigar is exactly a correlation between the two axes. A random leak pushes each point up or down by a random amount, so the cigar gets fatter but stays horizontal, and a fat horizontal cigar has no correlation.

One idea, many names. A leak with a fixed sign is what an electronics engineer calls coherent cross-talk, and a leak with a random sign incoherent cross-talk; a lock-in amplifier finds a weak signal by correlating with a reference, not by measuring power. In data analysis the same split separates a systematic error, which has a mean and therefore correlates with its cause, from a random one, which only adds variance. The rest of the section turns this into the exact formulas, first for the twist and then for lensing.

T1.11.1 From sticks to harmonic coefficients

What does a uniform twist do to $E_{\ell m}$ and $B_{\ell m}$? Twist every stick of the sky by the same angle β. At every direction $\boldsymbol{n}$, (T1.35) gives $(Q' \pm iU')(\boldsymbol{n}) = e^{\pm 2i\beta}(Q \pm iU)(\boldsymbol{n})$. Insert the harmonic expansion (T1.38) on both sides:

$$-\sum_{\ell m} \left(E'_{\ell m} \pm iB'_{\ell m}\right) {}_{\pm 2}Y_{\ell m}(\boldsymbol{n}) \overset{(1)}{=} e^{\pm 2i\beta} \left[-\sum_{\ell m} \left(E_{\ell m} \pm iB_{\ell m}\right) {}_{\pm 2}Y_{\ell m}(\boldsymbol{n}) \right]$$

$$\overset{(2)}{=} -\sum_{\ell m} e^{\pm 2i\beta}\big(E_{\ell m} \pm iB_{\ell m}\big)\,{}_{\pm 2}Y_{\ell m}(\boldsymbol{n}),$$

$$E'_{\ell m} \pm iB'_{\ell m} \overset{(3)}{=} e^{\pm 2i\beta}\big(E_{\ell m} \pm iB_{\ell m}\big). \tag{T1.49}$$

(1) The twist rule at each point, with both sides written in the harmonic basis.
(2) β is the same in every direction, so $e^{\pm 2i\beta}$ is a constant and moves inside the sum.
(3) The spin-weighted harmonics are a basis, so two expansions of the same function have the same coefficients.

Now separate E' and B', adding and subtracting the two signs:

$$E'_{\ell m} \overset{(1)}{=} \tfrac{1}{2}\big[e^{2i\beta}(E_{\ell m} + iB_{\ell m}) + e^{-2i\beta}(E_{\ell m} - iB_{\ell m})\big] \overset{(2)}{=} \cos 2\beta\, E_{\ell m} - \sin 2\beta\, B_{\ell m},$$
$$B'_{\ell m} \overset{(3)}{=} \tfrac{1}{2i}\big[e^{2i\beta}(E_{\ell m} + iB_{\ell m}) - e^{-2i\beta}(E_{\ell m} - iB_{\ell m})\big] \overset{(4)}{=} \sin 2\beta\, E_{\ell m} + \cos 2\beta\, B_{\ell m}. \tag{T1.50}$$

(1) Add the + and − versions of (T1.49); the $\pm iB'$ cancel and $2E'$ is left.
(2) Collect: the E terms give $\frac{1}{2}(e^{2i\beta} + e^{-2i\beta}) = \cos 2\beta$; the B terms give $\frac{i}{2}(e^{2i\beta} - e^{-2i\beta}) = \frac{i}{2}\cdot 2i\sin 2\beta = -\sin 2\beta$.
(3) Subtract the two versions; the E' cancel and $2iB'$ is left.
(4) The E terms give $\frac{1}{2i}(e^{2i\beta} - e^{-2i\beta}) = \sin 2\beta$; the B terms give $\frac{i}{2i}(e^{2i\beta} + e^{-2i\beta}) = \cos 2\beta$.

The pair $(E_{\ell m}, B_{\ell m})$ turns like (Q, U) at a point, by the angle 2β, mode by mode and with the same angle for every ℓ and m. The temperature does not change at all, $T'_{\ell m} = T_{\ell m}$, because a twist changes the direction of the polarization and not the intensity. The flat-sky version is the same with $E(\boldsymbol{\ell})$, $B(\boldsymbol{\ell})$ in place of $E_{\ell m}$, $B_{\ell m}$: panel (f) of Figure T1.17, with $2\beta = 40°$, is exactly $\cos 40°$ of E and $\sin 40°$ of B.

T1.11.2 All the rotated spectra

What happens to each of the six spectra? We take them one at a time, starting from a sky that is mirror symmetric before the twist: by (T1.41), $C_\ell^{EB} = C_\ell^{TB} = 0$. Write $c \equiv \cos 2\beta$ and $s \equiv \sin 2\beta$, both real numbers, and use the definition $\langle X_{\ell m} Y^*_{\ell m}\rangle = C_\ell^{XY}$ of (T1.39). The expectation of a sum is the sum of the expectations, and real constants come out of it.

EE: how much E is left?

$$C_\ell^{EE\prime} = \langle E'_{\ell m} E'^*_{\ell m}\rangle \overset{(1)}{=} \big\langle (cE_{\ell m} - sB_{\ell m})(cE^*_{\ell m} - sB^*_{\ell m})\big\rangle$$
$$\overset{(2)}{=} c^2 C_\ell^{EE} - cs\big(C_\ell^{EB} + C_\ell^{BE}\big) + s^2 C_\ell^{BB} \overset{(3)}{=} \cos^2 2\beta\, C_\ell^{EE} + \sin^2 2\beta\, C_\ell^{BB}. \tag{T1.51}$$

(1) Insert E' from (T1.50) and its complex conjugate.
(2) Multiply out and take the expectation term by term.
(3) The sky before the twist has $C_\ell^{EB} = C_\ell^{BE} = 0$.

BB: how much B is there now?

$$C_\ell^{BB\prime} \overset{(1)}{=} \big\langle (sE_{\ell m} + cB_{\ell m})(sE^*_{\ell m} + cB^*_{\ell m})\big\rangle \overset{(2)}{=} s^2 C_\ell^{EE} + sc\big(C_\ell^{EB} + C_\ell^{BE}\big) + c^2 C_\ell^{BB}$$
$$\overset{(3)}{=} \sin^2 2\beta\, C_\ell^{EE} + \cos^2 2\beta\, C_\ell^{BB}. \tag{T1.52}$$

(1) Insert B' from (T1.50).
(2) Multiply out and take expectations.
(3) No EB before the twist.

EB: does the new B remember E?

$$C_\ell^{EB\prime} \overset{(1)}{=} \left\langle (cE_{\ell m} - sB_{\ell m})(sE^*_{\ell m} + cB^*_{\ell m}) \right\rangle \overset{(2)}{=} cs\, C_\ell^{EE} + c^2 C_\ell^{EB} - s^2 C_\ell^{BE} - sc\, C_\ell^{BB}$$
$$\overset{(3)}{=} \cos 2\beta \sin 2\beta \left(C_\ell^{EE} - C_\ell^{BB}\right) \overset{(4)}{=} \tfrac{1}{2} \sin 4\beta \left(C_\ell^{EE} - C_\ell^{BB}\right). \tag{T1.53}$$

(1) E' times the complex conjugate of B'.
(2) Multiply out and take expectations.
(3) No EB before the twist; factor out cs.
(4) The double-angle identity $2 \sin x \cos x = \sin 2x$ with $x = 2\beta$.

TE and TB: what does the temperature see? The temperature is untouched, so only the second factor changes:

$$C_\ell^{TE\prime} \overset{(1)}{=} \left\langle T_{\ell m}(cE^*_{\ell m} - sB^*_{\ell m}) \right\rangle \overset{(2)}{=} c\, C_\ell^{TE} - s\, C_\ell^{TB} \overset{(3)}{=} \cos 2\beta\, C_\ell^{TE},$$
$$C_\ell^{TB\prime} \overset{(1)}{=} \left\langle T_{\ell m}(sE^*_{\ell m} + cB^*_{\ell m}) \right\rangle \overset{(2)}{=} s\, C_\ell^{TE} + c\, C_\ell^{TB} \overset{(3)}{=} \sin 2\beta\, C_\ell^{TE}, \qquad C_\ell^{TT\prime} = C_\ell^{TT}. \tag{T1.54}$$

(1) $T'_{\ell m} = T_{\ell m}$, and E', B' from (T1.50), conjugated.
(2) Take the expectations.
(3) No TB before the twist.

Had we kept an EB already present before the twist, the same steps would give $C_\ell^{EB\prime} = \frac{1}{2} \sin 4\beta\, (C_\ell^{EE} - C_\ell^{BB}) + \cos 4\beta\, C_\ell^{EB}$, with the C^{EB} terms of the first two lines adding $\mp \sin 4\beta\, C_\ell^{EB}$ to EE' and BB'; we need this once, in Section T1.13.3.

Key idea: The rotated spectra

A uniform twist β of a parity-symmetric sky gives

$$C_\ell^{EE\prime} = \cos^2 2\beta\, C_\ell^{EE} + \sin^2 2\beta\, C_\ell^{BB}, \qquad C_\ell^{BB\prime} = \sin^2 2\beta\, C_\ell^{EE} + \cos^2 2\beta\, C_\ell^{BB},$$
$$C_\ell^{EB\prime} = \tfrac{1}{2} \sin 4\beta \left(C_\ell^{EE} - C_\ell^{BB}\right), \qquad C_\ell^{TE\prime} = \cos 2\beta\, C_\ell^{TE}, \qquad C_\ell^{TB\prime} = \sin 2\beta\, C_\ell^{TE},$$

and leaves C_ℓ^{TT} alone [67, 68]. Power leaves EE and appears in BB; EB and TB appear because the twist has a hand.

T1.11.3 Reading the formulas

Power moves; nothing is lost. Add (T1.51) and (T1.52): $C_\ell^{EE\prime} + C_\ell^{BB\prime} = C_\ell^{EE} + C_\ell^{BB}$, because $\cos^2 + \sin^2 = 1$. The total polarized power at each ℓ is unchanged, as it must be, since a twist does not change the length P of any stick. The fraction $\sin^2 2\beta$ of the E power has moved into B, and the same fraction of the (small) B power has moved into E. Subtract instead, and $C_\ell^{EE\prime} - C_\ell^{BB\prime} = \cos 4\beta\, (C_\ell^{EE} - C_\ell^{BB})$, by $\cos^2 - \sin^2 = \cos(2 \cdot 2\beta)$.

EB and TB appear because the twist picks a hand. Before the twist they were zero by the mirror argument. After it they are proportional to $\sin 4\beta$ and $\sin 2\beta$, which change sign with β: twisting the other way gives the other sign, and the mirror image of a sky twisted by β is a sky twisted by $-\beta$. EB follows the shape of $C_\ell^{EE} - C_\ell^{BB}$, which on the real sky is essentially C_ℓ^{EE}, and TB follows the shape of C_ℓ^{TE}, with its changes of sign. These shapes are known from the ΛCDM fit, so a twist is a one-parameter template.

Small twists: linear against quadratic. For $|\beta| \ll 1$ (in radians), $\sin 2\beta \approx 2\beta$ and $\sin 4\beta \approx 4\beta$, so

$$C_\ell^{EB\prime} \approx 2\beta \left(C_\ell^{EE} - C_\ell^{BB}\right), \qquad C_\ell^{TB\prime} \approx 2\beta\, C_\ell^{TE}, \qquad C_\ell^{BB\prime} - C_\ell^{BB} \approx 4\beta^2 \left(C_\ell^{EE} - C_\ell^{BB}\right). \tag{T1.55}$$

The parity-odd spectra grow like β, the extra BB like β^2. For a small twist the signal is therefore in EB and TB, as in the radio story.

Example 1: a quarter turn of the sticks. $\beta = 45°$: $\cos 2\beta = 0$, $\sin 2\beta = 1$, $\sin 4\beta = 0$. Then $C^{EE\prime} = C^{BB}$, $C^{BB\prime} = C^{EE}$, $C^{TE\prime} = 0$, $C^{TB\prime} = C^{TE}$, and $C^{EB\prime} = 0$. E and B have swapped completely, so the sky looks like a pure-B sky with a TB correlation, and EB is zero because each mode is again purely of one kind. Panel (a) of Figure T1.17, with every stick turned by 45°, becomes panel (b).

Example 2: a half turn of the sticks. $\beta = 90°$: $\cos 2\beta = -1$, $\sin 2\beta = \sin 4\beta = 0$. Every stick along a wavevector now lies across it, so $E \to -E$ and $B \to -B$. All auto-spectra and EB are unchanged, but $C^{TE\prime} = -C^{TE}$: only the correlation with the temperature notices.

Example 3: the reported angle. $\beta = 0.35°$ is the value reported from Planck by [62]. Then $\sin 2\beta = 1.22\%$ and $\sin^2 2\beta = 1.490 \times 10^{-4}$. Figure T1.19 draws the rotated spectra computed by `code/chT1/06_rotated_spectra.py` from the fiducial spectra. The new EB reaches $D_\ell^{EB} = 0.517\,\mu\text{K}^2$ at $\ell \approx 990$, where EE peaks, and the new TB reaches $|D_\ell^{TB}| = 1.637\,\mu\text{K}^2$ near $\ell \approx 748$, larger because TE is larger than EE there. The extra BB is a small fraction of the lensing BB: 0.036 of it at $\ell = 100$ and 0.065 at $\ell = 1000$. Only at the very largest scales, $\ell \lesssim 4$, where the reionization peak of EE meets the smallest lensing power, does it exceed lensing, by up to a factor 2.539. A search for β in BB would be fighting lensing at every multipole where the data are good; a search in EB has no lensing to fight.

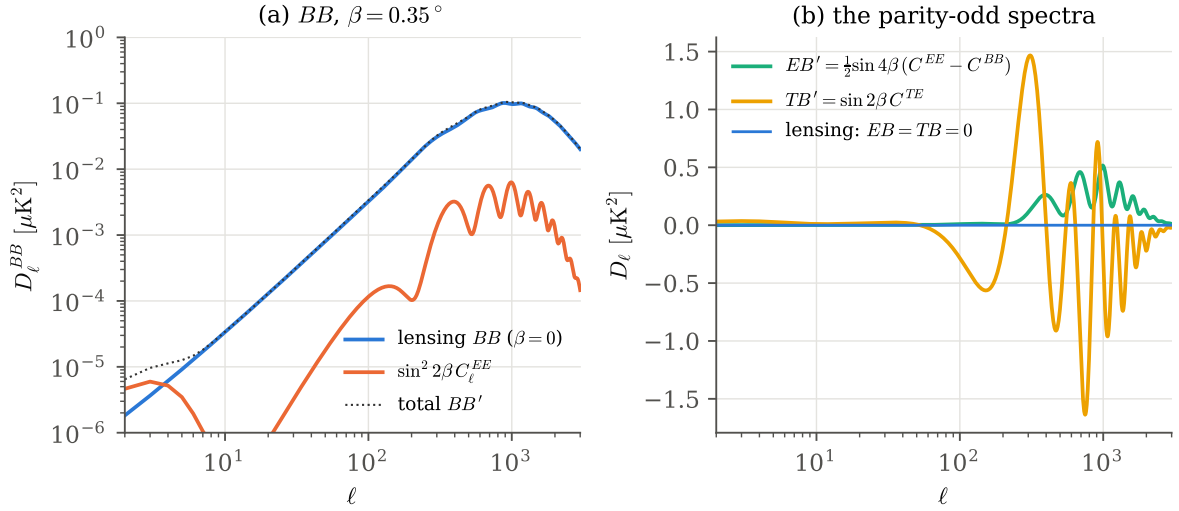


Figure T1.19: The spectra of the fiducial sky twisted by $\beta = 0.35°$. (a) BB: the lensing B modes that the untwisted sky already has (blue), the part moved over from EE by the twist, $\sin^2 2\beta\, C_\ell^{EE}$ (orange), and their sum (dotted), which is indistinguishable from lensing except on the largest scales. (b) The two parity-odd spectra created by the twist: EB follows the shape of EE, and TB follows TE with its changes of sign. Lensing alone leaves both at zero (blue line).

T1.11.4 Lensing: E into B without a hand

Lensing was met twice already: as the remapping $\hat{T}(\boldsymbol{n}) = T(\boldsymbol{n} + \boldsymbol{\nabla}\phi)$ of the temperature by the gradient of the lensing potential ϕ (Section T1.2, where the cosmology notes' ψ is CAMB's ϕ), and in Eq. (T1.7), where the shear made by a single potential was shown to have no B. That second result does not apply here. There, lensing *created* the sticks, and sticks made as second derivatives of one potential have no curl. Here the sticks already exist, created at last scattering as a pure E pattern, and lensing *moves* them: the stick seen in direction $\boldsymbol{n}$ is the one

that, without lensing, would have been seen in direction $\boldsymbol{n} + \boldsymbol{\nabla}\phi$. To the accuracy needed here the stick is carried along without being turned [47]. A pure E pattern whose pieces are shifted by different amounts in different places is no longer a pure E pattern. *How much B does that make, and does it have a hand?*

First order in the deflection. The deflection is a few arcminutes, small compared with the scales of the E pattern, so we expand to first order on a flat patch, for each of $X = T, Q, U$:

$$\tilde{X}(\boldsymbol{\theta}) = X(\boldsymbol{\theta} + \boldsymbol{\nabla}\phi) \approx X(\boldsymbol{\theta}) + \boldsymbol{\nabla}\phi \cdot \boldsymbol{\nabla} X(\boldsymbol{\theta}). \tag{T1.56}$$

The correction is a product of two fields. In Fourier space, with the convention of Section T1.10.2, a product becomes a convolution, $(gh)(\boldsymbol{\ell}) = \int \frac{\mathrm{d}^2\boldsymbol{\ell}'}{(2\pi)^2}\, g(\boldsymbol{\ell} - \boldsymbol{\ell}')\, h(\boldsymbol{\ell}')$, and a gradient becomes a factor $i\boldsymbol{\ell}$. Before lensing $B = 0$, so $Q \pm iU = e^{\pm 2i\varphi_{\ell'}} E(\boldsymbol{\ell}')$ by (T1.36). Write $K(\boldsymbol{\ell}, \boldsymbol{\ell}') \equiv (\boldsymbol{\ell} - \boldsymbol{\ell}') \cdot \boldsymbol{\ell}'$. Then

$$\begin{aligned}
\delta(Q \pm iU)(\boldsymbol{\ell}) &\overset{(1)}{=} \int \frac{\mathrm{d}^2\boldsymbol{\ell}'}{(2\pi)^2} \left[i(\boldsymbol{\ell} - \boldsymbol{\ell}')\phi(\boldsymbol{\ell} - \boldsymbol{\ell}') \right] \cdot \left[i\boldsymbol{\ell}' (Q \pm iU)(\boldsymbol{\ell}') \right] \\
&\overset{(2)}{=} - \int \frac{\mathrm{d}^2\boldsymbol{\ell}'}{(2\pi)^2}\, K(\boldsymbol{\ell}, \boldsymbol{\ell}')\, \phi(\boldsymbol{\ell} - \boldsymbol{\ell}')\, e^{\pm 2i\varphi_{\ell'}} E(\boldsymbol{\ell}'), \\
\delta(E \pm iB)(\boldsymbol{\ell}) &\overset{(3)}{=} e^{\mp 2i\varphi_{\ell}}\, \delta(Q \pm iU)(\boldsymbol{\ell}) \overset{(4)}{=} - \int \frac{\mathrm{d}^2\boldsymbol{\ell}'}{(2\pi)^2}\, K\, \phi(\boldsymbol{\ell} - \boldsymbol{\ell}')\, e^{\pm 2i(\varphi_{\ell'} - \varphi_{\ell})} E(\boldsymbol{\ell}'), \\
\tilde{B}(\boldsymbol{\ell}) &\overset{(5)}{=} - \int \frac{\mathrm{d}^2\boldsymbol{\ell}'}{(2\pi)^2}\, K(\boldsymbol{\ell}, \boldsymbol{\ell}')\, \phi(\boldsymbol{\ell} - \boldsymbol{\ell}')\, E(\boldsymbol{\ell}')\, \sin 2(\varphi_{\ell'} - \varphi_{\ell}).
\end{aligned} \tag{T1.57}$$

(1) The correction term of (T1.56) is the product of $\boldsymbol{\nabla}\phi$ and $\boldsymbol{\nabla}(Q \pm iU)$; transform the product as a convolution, each gradient bringing down i times its own wavevector.
(2) $i \cdot i = -1$; insert $Q \pm iU = e^{\pm 2i\varphi_{\ell'}} E$ for the unlensed, pure-E sky.
(3) The definition (T1.36) of E and B at the wavevector $\boldsymbol{\ell}$, applied to the change.
(4) Combine the two phases.
(5) B was zero, so $\tilde{B} = \delta B = [\delta(E + iB) - \delta(E - iB)]/2i$, and $(e^{ix} - e^{-ix})/2i = \sin x$ with $x = 2(\varphi_{\ell'} - \varphi_{\ell})$.

Each lensed B wave collects E waves at other wavevectors $\boldsymbol{\ell}'$, weighted by the lensing potential at the difference $\boldsymbol{\ell} - \boldsymbol{\ell}'$ and by the sine of twice the angle between the two wavevectors. That is the wobbling antenna of the opening: a leak from E into B whose coefficient is the random field ϕ.

Does the new B remember E? Correlate (T1.57) with E or with T:

$$\langle \tilde{B}(\boldsymbol{\ell})\, E^*(\boldsymbol{\ell}'') \rangle \overset{(1)}{=} - \int \frac{\mathrm{d}^2\boldsymbol{\ell}'}{(2\pi)^2}\, K\, \sin 2(\varphi_{\ell'} - \varphi_{\ell}) \left\langle \phi(\boldsymbol{\ell} - \boldsymbol{\ell}')\, E(\boldsymbol{\ell}')\, E^*(\boldsymbol{\ell}'') \right\rangle \overset{(2)}{=} 0, \tag{T1.58}$$

and in the same way $\langle \tilde{B}\, T^* \rangle = 0$.

(1) Insert (T1.57); the expectation passes inside the integral.
(2) ϕ, E and T are jointly Gaussian with mean zero. The average of a product of three such variables vanishes: changing the sign of all of them leaves their joint distribution unchanged but flips the sign of the product, so its average equals its own negative. This holds even though ϕ is slightly correlated with T.

At the next order the lensed E also has a correction δE, and $\langle \delta E\, \delta B^* \rangle$ contains the integral of $\cos 2(\varphi_{\ell'} - \varphi_{\ell}) \sin 2(\varphi_{\ell'} - \varphi_{\ell})$ against a weight that does not change when $\boldsymbol{\ell}'$ is reflected across $\boldsymbol{\ell}$; the sine changes sign under that reflection, so this term integrates to zero as well. The general reason holds at every order: ϕ is a scalar, a mirror carries a sky lensed by ϕ into the mirrored

sky lensed by the mirrored ϕ, and both are equally probable, so the argument of (T1.41) goes through unchanged. Lensing makes B, but no EB and no TB.

How much B power? Multiply (T1.57) by its complex conjugate at $\boldsymbol{\ell}''$ and average. Treat ϕ and E as independent Gaussian fields (their correlation is negligible), so the four-field average splits into $\langle\phi\phi^*\rangle\langle EE^*\rangle$, each a delta function times a spectrum, as in (T1.39):

$$\begin{aligned}\langle\tilde{B}(\boldsymbol{\ell})\tilde{B}^*(\boldsymbol{\ell}'')\rangle &\overset{(1)}{=} \int \frac{\mathrm{d}^2\boldsymbol{\ell}'}{(2\pi)^2}\frac{\mathrm{d}^2\boldsymbol{\ell}'''}{(2\pi)^2}\,KK'''\,\sin\sin'''\,(2\pi)^2\delta^2(\boldsymbol{\ell}'-\boldsymbol{\ell}''')C^{EE}_{\ell'}\,(2\pi)^2\delta^2\big(\boldsymbol{\ell}-\boldsymbol{\ell}'-\boldsymbol{\ell}''+\boldsymbol{\ell}'''\big)C^{\phi\phi}_{|\boldsymbol{\ell}-\boldsymbol{\ell}'|}\\ &\overset{(2)}{=} (2\pi)^2\delta^2(\boldsymbol{\ell}-\boldsymbol{\ell}'')\int\frac{\mathrm{d}^2\boldsymbol{\ell}'}{(2\pi)^2}\left[(\boldsymbol{\ell}-\boldsymbol{\ell}')\cdot\boldsymbol{\ell}'\right]^2\sin^2 2(\varphi_{\boldsymbol{\ell}'}-\varphi_{\boldsymbol{\ell}})\,C^{\phi\phi}_{|\boldsymbol{\ell}-\boldsymbol{\ell}'|}\,C^{EE}_{\ell'},\\ C^{BB}_{\ell} &\overset{(3)}{=} \int\frac{\mathrm{d}^2\boldsymbol{\ell}'}{(2\pi)^2}\left[(\boldsymbol{\ell}-\boldsymbol{\ell}')\cdot\boldsymbol{\ell}'\right]^2\sin^2 2(\varphi_{\boldsymbol{\ell}'}-\varphi_{\boldsymbol{\ell}})\,C^{\phi\phi}_{|\boldsymbol{\ell}-\boldsymbol{\ell}'|}\,C^{EE}_{\ell'}.\end{aligned}\tag{T1.59}$$

(1) Insert (T1.57) twice (the second with integration variable $\boldsymbol{\ell}'''$ and its factors K''', $\sin'''$), and replace $\langle EE^*\rangle$ and $\langle\phi\phi^*\rangle$ by their delta functions and spectra; the other pairings vanish because ϕ and E are independent with mean zero.

(2) Do the $\boldsymbol{\ell}'''$ integral with the first delta function, which sets $\boldsymbol{\ell}''' = \boldsymbol{\ell}'$; the second then becomes $\delta^2(\boldsymbol{\ell}-\boldsymbol{\ell}'')$.

(3) Read off the spectrum from the definition $\langle\tilde{B}(\boldsymbol{\ell})\tilde{B}^*(\boldsymbol{\ell}'')\rangle = (2\pi)^2\delta^2(\boldsymbol{\ell}-\boldsymbol{\ell}'')C^{BB}_{\ell}$.

This is the standard first-order lensing B spectrum [47, 69]. Its strength is set by the product $C^{\phi\phi}C^{EE}$ and nothing else: no parameter of new physics, and no hand.

Why is lensing B white on large scales? For ℓ much smaller than the multipoles $\ell' \sim 1000$ where E has its power, $\boldsymbol{\ell}-\boldsymbol{\ell}' \approx -\boldsymbol{\ell}'$. Then $K \approx -\ell'^2$, $C^{\phi\phi}_{|\boldsymbol{\ell}-\boldsymbol{\ell}'|} \approx C^{\phi\phi}_{\ell'}$, and only the sine still depends on the direction of $\boldsymbol{\ell}'$. Write $\mathrm{d}^2\boldsymbol{\ell}' = \ell'\mathrm{d}\ell'\,\mathrm{d}\varphi_{\boldsymbol{\ell}'}$:

$$C^{BB}_{\ell\to 0} \overset{(1)}{=} \int_0^\infty \frac{\ell'\mathrm{d}\ell'}{(2\pi)^2}\,\ell'^4 C^{\phi\phi}_{\ell'}C^{EE}_{\ell'}\int_0^{2\pi}\mathrm{d}\varphi\,\sin^2 2\varphi \overset{(2)}{=} \frac{1}{4\pi}\int_0^\infty \mathrm{d}\ell'\,\ell'^5\,C^{\phi\phi}_{\ell'}\,C^{EE}_{\ell'}.\tag{T1.60}$$

(1) Insert the large-ℓ' approximations into (T1.59); the angle between $\boldsymbol{\ell}'$ and $\boldsymbol{\ell}$ is now the only angle left.

(2) $\int_0^{2\pi}\sin^2 2\varphi\,\mathrm{d}\varphi = \pi$, and $\pi/(2\pi)^2 = 1/4\pi$.

The answer does not depend on ℓ: a flat spectrum, like white noise. Large-scale lensing B modes are a sum of many small-scale E modes, each moved a little, and such a sum has no preferred large scale. Figure T1.20 checks both formulas against CAMB's full calculation. The first-order integral (T1.59) gives between 0.985 and 0.997 times CAMB's lensed BB everywhere between $\ell = 20$ and 1500, and the white limit (T1.60) gives $C^{BB} = 1.890\times10^{-6}\,\mu\mathrm{K}^2$, against CAMB's mean $1.940\times10^{-6}\,\mu\mathrm{K}^2$ over $10 \le \ell \le 50$. Written as the noise level of a map, that white floor is $\sqrt{C^{BB}} = 4.7\,\mu$K-arcmin: an experiment whose polarization noise is below about 5 μK-arcmin sees lensing, not its own noise, as the floor of its B-mode map.

Lensing B is not Gaussian. Equation (T1.57) makes $\tilde{B}$ a product of two Gaussian fields, and a product of Gaussians is not Gaussian: its four-point function has a part that the power spectrum does not describe, which correlates the errors of different ℓ bins. The same product structure is what allows **delensing** [47]: reconstruct ϕ (by the quadratic estimator of Section T1.6, or from tracers of the matter), take E from the measured map, compute $\tilde{B}$ from (T1.57) and subtract it. A Gaussian noise could not be removed this way, because nothing in the data predicts it.

T1.11.5 Both on simulated skies

The two derivations make claims about averages over many skies. The script `code/chT1/08_flat_rotation_lensing.py` checks them on 500 simulated flat patches of $17^\circ \times 17^\circ$ (256 pixels

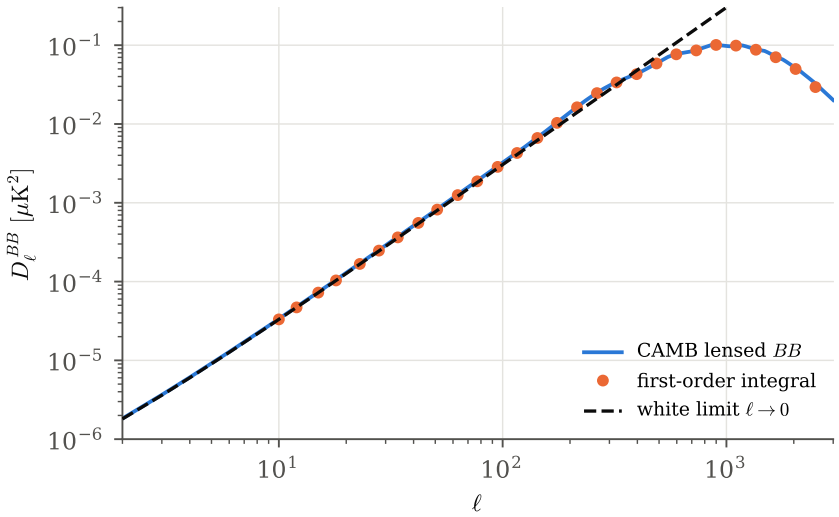


Figure T1.20: The lensing B-mode spectrum. Blue: CAMB's lensed BB for the fiducial model with no gravitational waves. Orange dots: the first-order integral (T1.59) evaluated from CAMB's unlensed EE and lensing potential. Dashed: its large-scale limit (T1.60), a constant C_ℓ^{BB}, which appears as a line of slope two in D_ℓ. The first-order formula holds up to $\ell \approx 1500$; beyond, higher orders in the deflection matter.

of 4′ on a side). Each patch is a Gaussian sky with the unlensed T and E of CAMB, correlated through TE, no B, and an independent Gaussian lensing potential with CAMB's $C^{\phi\phi}$. Each sky is processed twice: once twisted by $\beta = 2°$, chosen so that the B power it makes is about that of lensing (their median ratio over $100 < \ell < 1500$ is 1.00), and once lensed to first order with (T1.56). For each sky we measure the binned BB, EB and TB in 25 bins between $\ell = 40$ and 2040 and average over the skies.

Figure T1.21 shows the answer. In BB the two are alike, each following its own prediction. In EB and TB they are not: the twist produces the shapes of C^{EE} and C^{TE} from Section T1.11.2, and lensing produces nothing. Quantitatively, the averaged lensing BB divided by the first-order integral is 1.000 ± 0.001 over $100 < \ell < 1500$, and the twisted EB divided by $\frac{1}{2}\sin 4\beta\, C^{EE}$ is 1.0000 ± 0.0004. The lensing EB and TB bin averages scatter around zero by the amount their own Monte Carlo errors allow: the sums of their squared ratios to those errors are 18.1 and 25.3 for 25 bins, where pure chance gives 25 ± 7. The largest lensing EB average is 0.15% of the peak of the twisted EB.

The simulation also shows a property of the twist that is exact sky by sky, not only on average. Since the simulated skies have no B before the twist, $\widehat{C}_\ell^{EB\prime} = \frac{1}{2}\sin 4\beta\, \widehat{C}_\ell^{EE}$ and $\widehat{C}_\ell^{EE\prime} - \widehat{C}_\ell^{BB\prime} = \cos 4\beta\, \widehat{C}_\ell^{EE}$ hold for the measured spectra of each sky, because (T1.50) acts mode by mode. Dividing one by the other, $\widehat{C}_\ell^{EB\prime} = \frac{1}{2}\tan 4\beta\, (\widehat{C}_\ell^{EE\prime} - \widehat{C}_\ell^{BB\prime})$: the twisted EB is predicted by the twisted EE and BB of the same sky, with no reference to any model. The script finds this to hold to 5.100×10^{-16} in every sky. On the real sky, where some B is present before the twist, chance EB adds scatter to this relation; Section T1.13.3 builds the measurement on it.

T1.11.6 The instrument twists too

Every detector that measures Q and U has its own reference axis on the sky. If the true axis is turned by $-\alpha$ from where the analysis believes it is, then by the argument at the end of Section T1.10.1 every measured stick appears twisted by $+\alpha$. The twist of the instrument acts on the maps at the end of the line, after any physical twist β the sky light received on its way,

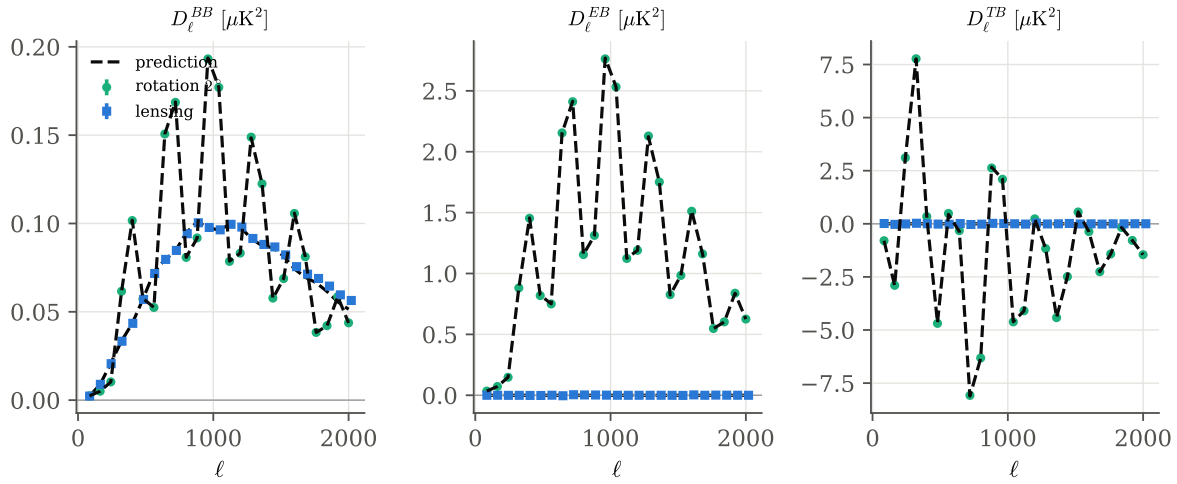


Figure T1.21: A twist and lensing on 500 simulated flat skies. Green circles: skies twisted by 2°; blue squares: the same skies lensed to first order; error bars are the errors of the averages over the skies; dashed: the predictions ($\sin^2 2\beta\, C^{EE}$, $\frac{1}{2}\sin 4\beta\, C^{EE}$, $\sin 2\beta\, C^{TE}$ for the twist; the first-order integral for lensing BB, zero for lensing EB and TB). Both make B power of about the same size. Only the twist makes EB and TB.

so two twists in a row add: $e^{2i\alpha}e^{2i\beta} = e^{2i(\alpha+\beta)}$. Every formula derived above therefore holds with β replaced by $\alpha+\beta$: $C_\ell^{EB,\mathrm{obs}} = \frac{1}{2}\sin(4\alpha+4\beta)(C_\ell^{EE} - C_\ell^{BB})$ for the CMB, and likewise for TB, EE and BB. The CMB spectra alone measure the sum and nothing else, so a miscalibrated polarization angle and a physical twist are exactly degenerate in them. Section T1.13.3 shows how the light of our own Galaxy, which is twisted by α but not by β, breaks this degeneracy.

Watch out: EB alone does not prove new physics

A nonzero EB with the shape of C^{EE} measures $\alpha+\beta$. Without an independent handle on α (a calibration source of known polarization angle, or the Galactic foreground method of Section T1.13.3), an EB signal is first of all a statement about the orientation of the detectors.

T1.12 Where B modes come from

We have met three makers of B so far. Gravitational waves make it at last scattering, because their pattern is not mirror symmetric (Section T1.10.6). Lensing makes it from E, with no hand (Section T1.11.4). A twist makes it from E, with a hand (Section T1.11.2). Each was studied alone. A real B-mode map contains all of them at once, plus the light of our own Galaxy and the imperfections of the instrument, and nothing in the map is labelled. Given a measured excess of B power, or a measured EB, how do we decide what made it?

Imagine a photograph with stains on it, and three ways to examine it. First, look at it through a red, a green and a blue filter. A stain that is equally dark through all three is grey; one that is darker through the red filter is reddish. Second, look at the size of the blotches: some are a centimetre across, some are a fine grain. Third, hold it to a mirror: stains with no shape of their own look the same, but a stain shaped like a swirl or a letter is reversed. No single test sorts every stain, but together the colour, the size and the mirror nearly always do.

The B-mode sky is examined in exactly these three ways. The filters are the observing frequencies. In the temperature units used for the CMB, everything that comes from the CMB itself (gravitational waves, lensing, a twist on the way) has the same amplitude in every channel,

a grey stain. The Galaxy is coloured: dust brightens towards high frequency, synchrotron towards low. The blotch size is the multipole ℓ. The mirror is the parity of the spectrum: EB and TB are nonzero only for a source with a hand. One sentence holds the idea of the section: *each source of B leaves fingerprints in frequency, in ℓ and in parity, and the fingerprints are read together.*

The colour test in symbols. Take one multipole, $\ell = 80$, and two channels, 150 and 353 GHz. A grey source has the same D_ℓ^{BB} in both. Dust, scaled with the frequency law of Section T1.12.3, has 410 times more power at 353 than at 150 GHz. Measure D_{80}^{BB} in both channels, $x_{150} = g + d$ and $x_{353} = g + 410\,d$, with g the grey part and d the dust at 150 GHz. The naive reading of the 150 GHz channel alone takes all of x_{150} as cosmological. In the cleanest one per cent of the sky, where the dust is weakest, it would be wrong by a large factor: there the dust at 150 GHz is $D_{80}^{BB} = 0.011\,\mu\text{K}^2$, 5.2 times the lensing B and 16.5 times the B of gravitational waves with $r = 0.01$. The good choice uses both channels, $d = (x_{353} - x_{150})/(410 - 1)$ and $g = x_{150} - d$, the two-reading weighting of Section 3.7.4 in its simplest form. Its price is the noise of the 353 GHz channel, which enters g with a weight of only $1/(410 - 1)$ but is large, and the assumption that the dust frequency law is known. Lesson: frequency separates the Galaxy from the CMB, but it cannot separate one grey source from another; for that we need the other two tests.

One idea, many names. Telling sources apart by how they respond to several probes is what an astronomer does with colours, what a particle physicist does with a calorimeter and a tracker together, and what a statistician calls a mixing model: each reading is a sum of sources, each source with its own pattern across the readings. When the patterns are linearly independent, linear algebra separates the sources; the internal linear combination of Section 3.7.4 and the parametric fits of B-mode experiments are both this idea.

We now take the sources one at a time, each with its physical origin and its fingerprints, and then collect them in one table.

T1.12.1 Gravitational waves from the early universe

Origin. Inflation predicts a background of gravitational waves in addition to the density perturbations (see the cosmology notes). Its strength is quoted as the **tensor-to-scalar ratio** r [the primordial power of the gravitational waves divided by that of the curvature perturbations, at the pivot $k_* = 0.05\,\text{Mpc}^{-1}$ of Section T1.2]. As a gravitational wave passes, it stretches the plasma in a quadrupole pattern, and Thomson scattering turns that quadrupole into polarization with both E and B [64, 66]. The best limit is $r < 0.036$ at 95% confidence [70].

Fingerprints. In ℓ, two bumps (blue curve of Figure T1.22, computed with CAMB for $r = 0.01$). The *reionization bump* at $\ell \lesssim 10$ comes from polarization generated when the universe was reionized, late and therefore seen on very large angles. The *recombination bump*, here at $\ell \approx 88$ with $D_\ell^{BB} = 6.590 \times 10^{-4}\,\mu\text{K}^2$, comes from last scattering: it sits at the angle that the horizon at last scattering subtends, because a gravitational wave decays once it is inside the horizon, so waves smaller than the horizon then had already faded. The amplitude is proportional to r. In frequency, grey: the signal is a pattern in the CMB itself, so it has the black-body spectrum of the CMB and the same amplitude in every channel in CMB temperature units. In statistics, Gaussian, like the primordial density field. In parity, symmetric: standard inflation has no hand, so $EB = TB = 0$. (Models with chiral gravitational waves, where one circular polarization of the wave is stronger, would break this [67].)

T1.12.2 Lensing of E

Origin. The remapping of the E pattern by the matter between us and last scattering, derived in Section T1.11.4.

Fingerprints. In ℓ, a broad peak at $\ell \approx 886$, inherited from the peak of EE, and a white spectrum, $C_\ell^{BB} \approx 1.890{\times}10^{-6}\,\mu\mathrm{K}^2$, on large scales (T1.60). At $\ell = 80$ lensing is $D_{80}^{BB} = 0.00205\,\mu\mathrm{K}^2$, about three times the recombination bump of $r = 0.01$; for smaller r the gravitational-wave bump hides under lensing, which is why delensing matters for future searches. In frequency, grey, since gravity bends all wavelengths alike. In statistics, non-Gaussian (a product of ϕ and E), which is what delensing exploits. In parity, symmetric: $EB = TB = 0$ on average (T1.58).

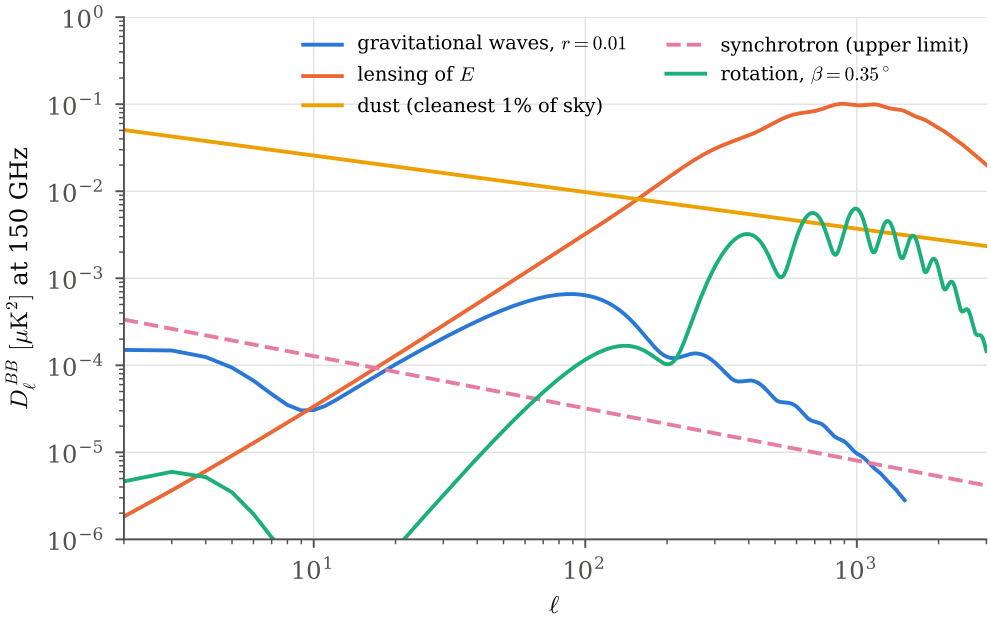


Figure T1.22: The B-mode spectra of the different sources at 150 GHz. Gravitational waves with $r = 0.01$ (blue): the reionization bump at the lowest ℓ and the recombination bump near $\ell \approx 88$. Lensing (orange): white on large scales, peaking near $\ell \approx 886$. Dust (yellow) in the cleanest one per cent of the sky, from the BICEP/Keck model [70], and the 95% upper limit on synchrotron (pink dashed): power laws that fall with ℓ. A twist of $\beta = 0.35°$ (green): the shape of EE, a factor 1.490×10^{-4} down. Over most of the sky the dust is tens of times higher than drawn here.

T1.12.3 The Galaxy: dust and synchrotron

Origin. Two kinds of emission from the Milky Way are polarized [19, 71]. Interstellar *dust* grains are elongated and spin with their long axes across the Galactic magnetic field, so their thermal emission is polarized across the field. *Synchrotron* emission comes from relativistic electrons spiralling around the magnetic field lines, and it is polarized too. Both trace the turbulent, filamentary magnetized gas of the Galaxy, which has structure on every scale.

Fingerprints in ℓ. Power laws, $D_\ell \propto \ell^\alpha$, falling towards small scales. For dust, Planck finds D_ℓ^{EE} and D_ℓ^{BB} both close to $\ell^{-0.42}$ over $40 \le \ell \le 600$, with $BB/EE \approx 0.5$ and a positive TE [71]. For synchrotron the slope is between -0.6 and -1 [70]. A power law that falls with ℓ is unlike both the bumps of gravitational waves and the rising lensing spectrum.

Fingerprints in frequency. This is the strongest test. Dust emits like a black body at $T_d \approx 19.6\,\mathrm{K}$ multiplied by ν^{β_d} with $\beta_d \approx 1.5$ (a modified black body); synchrotron falls like $\nu^{-3.1}$ in brightness temperature [70]. Converted to CMB temperature units, dust rises steeply with frequency and synchrotron falls steeply, and the CMB is flat. Figure T1.23(a), from `code/chT1/07_bmode_sources.py`, plots the B-mode amplitude at $\ell = 80$ against frequency in the cleanest patch of sky: the dust power grows by a factor 410 from 150 to 353 GHz, the synchrotron power (at its upper limit) by a factor 7504 from 150 down to 30 GHz, and the sum

of the two is smallest near 90 GHz. This is why CMB experiments put their most sensitive channels near 100 to 150 GHz and add channels on both sides to measure the Galaxy.

Fingerprints in statistics. Neither emission is Gaussian or statistically isotropic. Its amplitude follows the column of gas along each line of sight, so it is far brighter near the Galactic plane than at high latitude, and its maps are filamentary. Over 71% of the sky Planck measures a dust amplitude $D_{80}^{EE} = 315\,\mu\mathrm{K}^2$ at 353 GHz [71], so $D_{80}^{BB} \approx 0.53 \times 315 \approx 170\,\mu\mathrm{K}^2$, against $4.4\,\mu\mathrm{K}^2$ in the cleanest patch [70]: about forty times more. A Galactic mask is therefore part of every analysis, and the answer depends on which mask.

Fingerprints in parity. The Galaxy can have a hand. The intensity filaments of the dust tend to be aligned with the magnetic field, which makes the dust polarization mostly E-like; where filament and field are misaligned, the misalignment angle has a sign, and the polarization acquires correlations TB and EB [72]. Planck finds a positive dust TB at 353 GHz [71], and [72] predicts a positive dust EB of up to $D_\ell^{EB} \approx 2.5\,\mu\mathrm{K}^2$ at 353 GHz for $100 \lesssim \ell \lesssim 500$ over their mask, with a sign and size that can change with the sky area. This small Galactic hand is the main worry for birefringence measurements (Section T1.13.4).

T1.12.4 Twists: birefringence and Faraday rotation

Cosmic birefringence. A twist of the plane of polarization during the journey from last scattering, by an angle β that is the same for every frequency. Its physical origin is discussed in Section T1.13. Fingerprints: BB, EB and TB with the shapes of Section T1.11.2 (EB follows $C^{EE} - C^{BB}$, TB follows C^{TE}); grey in frequency; Gaussian (it is a fixed linear operation on Gaussian fields); and the only cosmological source that creates EB and TB.

Faraday rotation. A plane of polarization travelling along a magnetic field through an ionized gas is also twisted, by an angle proportional to the square of the wavelength, that is to $1/\nu^2$ [73]; Section T1.13.2 derives this law from the motion of the electrons. A primordial magnetic field of 10^{-9} G today would twist the CMB polarization by about 1° at 30 GHz [73], hence by $(30/150)^2 \times 1^\circ = 0.04^\circ$ at 150 GHz. The twist formulas of Section T1.11 apply channel by channel (with the angle varying across the sky when the field does), so Faraday rotation makes EB and TB like birefringence, but its strong frequency dependence (Figure T1.23(b)) tells it apart: a constant β and an angle falling as ν^{-2} cannot be confused once several channels are compared.

T1.12.5 The instrument

Three imperfections of the telescope matter most [74]. A *miscalibrated polarization angle* α is a twist applied to the maps by the instrument (Section T1.11.6): in the CMB it is exactly degenerate with β, and it is grey only by accident, since each channel, or each detector, has its own α. *Temperature-to-polarization leakage* [a small fraction of the temperature signal read as Q or U, from mismatched gains or bandpasses of the paired detectors] copies the much larger temperature anisotropy into polarization, producing spurious TE and EE, and TB and BB as well, with shapes inherited from C^{TT}. *Beam asymmetry* [the two detectors of a pair seeing slightly different, non-circular patches of sky] does the same with the gradients and second derivatives of the temperature. These effects depend on how the sky was scanned, so they differ between channels, detectors and data splits, and they are hunted with null tests on differences of data splits and corrected with models of the beams [75].

All these fingerprints fit in one table, which is the tool to reach for when a B-mode signal needs a name.

Taxonomy: Where a B mode comes from: the fingerprints

source	shape in ℓ (BB)	frequency (CMB units)	statistics	EB, TB
gravitational waves (r)	reionization bump at $\ell < 10$, recombination bump near $\ell \approx 80$, amplitude $\propto r$	flat	Gaussian	none
lensing of E	white at low ℓ, broad peak near $\ell \approx 1000$	flat	non-Gaussian, removable by delensing	none on average
dust	power law $D_\ell \propto \ell^{-0.42}$, rising to low ℓ	rises steeply with ν	non-Gaussian, anisotropic	small EB, TB from filament–field misalignment
synchrotron	power law, rising to low ℓ	falls steeply with ν	non-Gaussian, anisotropic	small, if any
cosmic birefringence (β)	$BB = \sin^2 2\beta\, C^{EE}$; EB follows $C^{EE} - C^{BB}$, TB follows C^{TE}	flat	Gaussian	yes: the only cosmological source
Faraday rotation	as birefringence, angle varying over the sky	angle $\propto \nu^{-2}$	follows the magnetized plasma	yes
angle miscalibration (α)	exactly as birefringence	own value per channel or detector	as the CMB	yes
$T \to P$ leakage, beam asymmetry	shapes of C^{TT} and its derivatives	per channel and detector	depends on scan	can be nonzero

T1.12.6 Reading the fingerprints together

The table is read with three questions, each of which splits the list along a different line. *Does the signal change with frequency?* If it does, it comes from the Galaxy, from Faraday rotation, or from the instrument; if it does not, it comes from the CMB, which leaves gravitational waves, lensing and birefringence (and an angle miscalibration that happens to be equal in every channel). *What is its shape in ℓ?* Among the grey sources, the recombination bump near $\ell \approx 80$ marks gravitational waves, the white spectrum rising to a peak near $\ell \approx 1000$ marks lensing, and the shape of C^{EE} marks a twist. *Is it parity-odd?* EB and TB separate the twists (birefringence, Faraday, miscalibration) and the slightly handed dust from everything that preserves parity, which includes gravitational waves and lensing. Here are four cases, worked in the order of the opening: the event, what could have made it, and how each candidate fares.

Case 1: a BB excess at $\ell \approx 80$. *The event:* D_{80}^{BB} at 150 GHz is a few times the lensing prediction. *What could make it?* Gravitational waves (bump there), dust (rising to low ℓ), leakage. *Test each.* Measure the same patch at 353 GHz. Gravitational waves predict the same D_{80}^{BB}; dust predicts

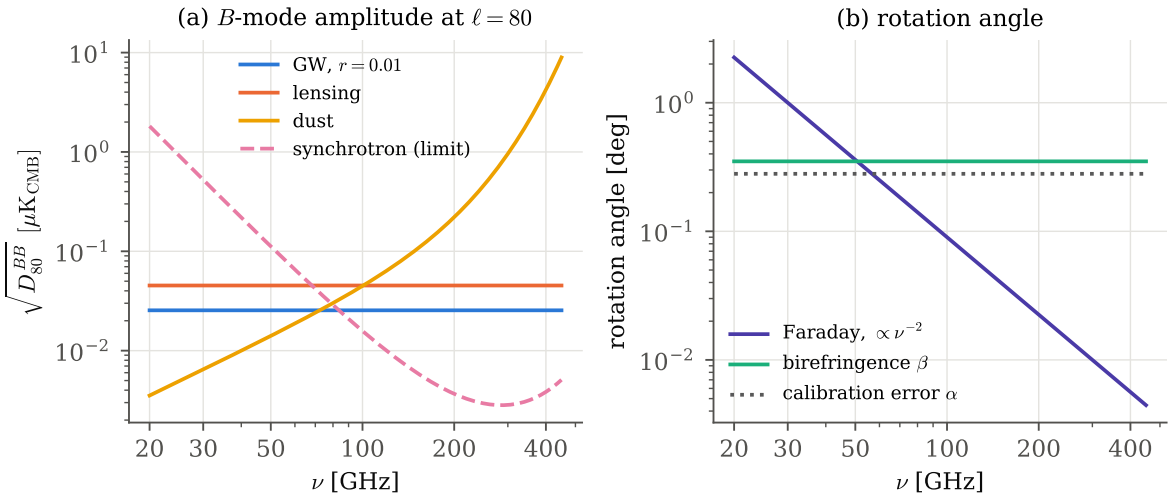


Figure T1.23: The colour test. (a) The rms B-mode amplitude at $\ell = 80$ against observing frequency, in CMB temperature units, for the cleanest patch of sky: gravitational waves ($r = 0.01$) and lensing are flat, dust rises and synchrotron (upper limit) falls; the Galaxy is weakest near 90 GHz. (b) The twist angle against frequency: cosmic birefringence ($\beta = 0.35°$) is flat, Faraday rotation by a 1 nG primordial field falls as ν^{-2}; the dotted line marks the 0.28° calibration uncertainty of Planck, which a single channel cannot distinguish from β.

410 times more. Leakage predicts a different answer for each detector set. Only after this does r get a number, from what is left of the 150 GHz power once the dust inferred at 353 GHz is subtracted, as in the colour test of the opening.

Case 2: BB power that rises from a flat floor to a peak at $\ell \approx 1000$, the same in every channel. Grey, so cosmological. The shape is that of lensing, and the identification can be confirmed in a way no other source allows: reconstruct ϕ, compute the B it predicts from the measured E with (T1.57), and correlate. If the correlation matches, the power is lensing.

Case 3: an EB with the shape of C^{EE}, equal in all channels. Parity-odd, grey, and with the shape of a twist: the candidates are birefringence and a common angle miscalibration. The CMB cannot tell them apart (Section T1.11.6). Something that is twisted by α but not by β is needed, and Section T1.13.3 finds it in the Galaxy.

Case 4: an EB that is four times larger at 75 GHz than at 150 GHz. A ratio of $(150/75)^2 = 4$ is the ν^{-2} law of Faraday rotation; birefringence would give a ratio of one, dust, which is faint at low frequency, a ratio below one, and synchrotron, which brightens much faster towards low frequency, a ratio of tens. The candidates are Faraday rotation or, if the ratio differs from channel to channel without a pattern, independent miscalibrations of the two channels.

The lesson of the four cases: frequency separates the Galaxy (and Faraday rotation) from the CMB; the shape in ℓ separates gravitational waves from lensing; and EB, TB separate a rotation from everything that preserves parity. A real analysis does all three at once, as one likelihood over all the auto- and cross-spectra of all channels [70].

T1.13 Cosmic birefringence, the instrument and the Galaxy

The previous sections leave us with a precise statement and a problem. The statement: a uniform twist of the CMB polarization by an angle produces $EB = \frac{1}{2}\sin(4\times\text{angle})\,(C^{EE} - C^{BB})$ and $TB = \sin(2\times\text{angle})\,C^{TE}$, and nothing else on the sky produces that pattern in every channel. The problem: the angle the CMB measures is $\alpha + \beta$, the twist of the instrument plus the twist of the sky (Section T1.11.6), and from the CMB alone there is no way to split the sum.

Think of a painting in a room that looks crooked to you. Either the painting hangs crooked, or your head is tilted, or both. Looking harder at the painting cannot decide: all you see is the sum of the two tilts. So you look at the door frame next to it. The door frame is close to you, and nothing made it crooked except, possibly, your own head. If the door frame looks straight and the painting crooked, the painting is crooked; if both look crooked by the same amount, it is your head. The method works as long as the door frame really is straight. If the builder hung it slightly crooked, you will misjudge the painting by exactly that amount.

The microwave sky has a door frame. The polarized emission of our Galaxy reaches the telescope after a journey of a few hundred to a few thousand light years, far too short for any cosmological twist to act on it, but it passes through the same detectors as the CMB. It is twisted by α and not by β [76]. The idea of the section, in one sentence: *measure the twist of the Galaxy to get α, measure the twist of the CMB to get $\alpha + \beta$, and take the difference.*

The door frame in symbols. Write the twisted EB of each component to first order in the angles, using (T1.55): the CMB gives $2(\alpha + \beta)\Delta_\ell^{\rm CMB}$ and the dust gives $2\alpha\,\Delta_\ell^{\rm dust}$, where $\Delta_\ell \equiv C_\ell^{EE} - C_\ell^{BB}$ of each. The two dials are α and β; the two patterns across multipoles are $\Delta_\ell^{\rm CMB}$, with its acoustic peaks, and $\Delta_\ell^{\rm dust}$, a smooth power law. Without dust the data are one number times one pattern and fix only $\alpha + \beta$: the naive analysis has one equation for two unknowns. With dust the data are $2\alpha(\Delta_\ell^{\rm dust} + \Delta_\ell^{\rm CMB}) + 2\beta\,\Delta_\ell^{\rm CMB}$, two patterns with two different shapes, and both numbers follow. The price is that β is now the difference of two measured quantities, so its error is at least as large as the larger of their two errors. And the reference is not free: dust bright enough to fix α is also extra scatter in the CMB's own EB. In the forecast of Section T1.13.3 for one channel, the CMB in a dust-free map fixes $\alpha + \beta$ to $0.036°$; at 353 GHz the dust fixes α to $0.1129°$, but the sum is now known only to $0.271°$, and β alone to $0.3717°$, ten times worse than the dust-free sum. Lesson: a degeneracy is broken by a second component that sees only one of the two parameters, and the price is paid in the precision of that second component.

One idea, many names. Calibrating an instrument on a source whose answer is known, while measuring an unknown one at the same time, is the oldest trick of experimental physics: a reference star in the same field, a known line in a spectrum, a tag-and-probe sample at a collider. What is special here is that the reference is not known in advance. Only one of its properties is assumed, that the Galaxy itself has no EB, and its own $EE - BB$ is measured from the same data.

We first say what physics could twist the CMB light, and how ordinary matter twists light, then derive the Minami–Komatsu relation that turns the door-frame idea into a measurement, and end with what has been measured and why it is not yet a discovery.

T1.13.1 What could twist the light

A twist is a phase difference between two helicities. Write the two circular polarization states of a wave travelling along $\boldsymbol{e}_3$ as $\boldsymbol{e}_\pm = (\boldsymbol{e}_1 \pm i\boldsymbol{e}_2)/\sqrt{2}$. A wave linearly polarized at angle ψ, with

polarization vector $\boldsymbol{\epsilon} = \cos\psi\,\boldsymbol{e}_1 + \sin\psi\,\boldsymbol{e}_2$, is a sum of the two:

$$\frac{e^{-i\psi}\boldsymbol{e}_+ + e^{i\psi}\boldsymbol{e}_-}{\sqrt{2}} \overset{(1)}{=} \tfrac{1}{2}\left[e^{-i\psi}(\boldsymbol{e}_1 + i\boldsymbol{e}_2) + e^{i\psi}(\boldsymbol{e}_1 - i\boldsymbol{e}_2)\right] \overset{(2)}{=} \cos\psi\,\boldsymbol{e}_1 + \sin\psi\,\boldsymbol{e}_2 = \boldsymbol{\epsilon}. \tag{T1.61}$$

(1) Insert the definitions of $\boldsymbol{e}_\pm$.
(2) Collect: $\frac{1}{2}(e^{-i\psi} + e^{i\psi}) = \cos\psi$ multiplies $\boldsymbol{e}_1$, and $\frac{i}{2}(e^{-i\psi} - e^{i\psi}) = \frac{i}{2}(-2i\sin\psi) = \sin\psi$ multiplies $\boldsymbol{e}_2$.

Now suppose that on its way the + helicity is delayed in phase by β relative to the average, and the − helicity advanced by β, so the left side becomes $(e^{-i(\psi+\beta)}\boldsymbol{e}_+ + e^{i(\psi+\beta)}\boldsymbol{e}_-)/\sqrt{2}$. By the same identity this is the linear polarization at $\psi + \beta$. A relative phase of 2β between the two helicities is a twist of the plane by β. A medium in which the two helicities travel at different speeds is called birefringent, and a twist of the CMB plane accumulated over its whole journey is called **cosmic birefringence**.

What could make the helicities differ? A pseudoscalar field ϕ that couples to photons through the term $\frac{1}{4}g_{\phi\gamma}\phi F_{\mu\nu}\tilde{F}^{\mu\nu}$ in the Lagrangian, where F is the electromagnetic field tensor, $\tilde{F}$ its dual and $g_{\phi\gamma}$ a coupling constant, does exactly this when ϕ changes in time [77, 78]. The product $F\tilde{F}$ is proportional to $\boldsymbol{E}\cdot\boldsymbol{B}$ of the electromagnetic field, which changes sign in a mirror; multiplied by a pseudoscalar, which also changes sign, it keeps the Lagrangian mirror symmetric, but a background ϕ that evolves in one direction picks a hand. The result is a twist

$$\beta = \tfrac{1}{2}g_{\phi\gamma}\left(\phi_{\rm obs} - \phi_{\rm LSS}\right), \tag{T1.62}$$

set by how much the field changed between last scattering and today, and independent of the frequency of the light [62, 67, 78]. Such "axion-like" fields are candidates for dark matter and dark energy. If the field had always been constant, or did not couple, $\beta = 0$. Because the CMB has travelled for almost the whole age of the universe, it is the most sensitive test: a nonzero β would be a sign of parity violation in the laws of physics beyond the Standard Model. For the data analyst the microphysics matters through three consequences, all met already: β is the same in every channel, it is (for a field that is uniform on the sky) the same in every direction, and its signature is the set of spectra of Section T1.11.2. Equation (T1.62) also says that the early proposals and searches for a twist in the CMB [67, 68] are, in this language, measurements of $g_{\phi\gamma}\Delta\phi$.

Faraday rotation is also a helicity phase difference, caused by free electrons in a magnetic field, but its phase grows with the square of the wavelength (Section T1.12.4). That is the physical reason it is told apart by frequency, while a pseudoscalar twist is not. Both statements can be derived from the motion of a charge pushed by the wave, which is the subject of the next subsection.

T1.13.2 How matter turns light: the electron on a spring

The helicity identity (T1.61) turned the question "what twists the plane of polarization?" into "what gives the two helicities different phases?". For the pseudoscalar field the answer was stated with a citation. For ordinary matter it can be derived from a model every physicist has met, a charge on a spring, and the derivation shows what a twist needs and what it does not need.

Lay a crystal of calcite on a printed page and every letter appears twice. Light polarized along one direction in the crystal travels at one speed, light polarized at right angles to it at another, and the two images are made of the two polarizations. Now put a glass of clear syrup between two polarizing filters, as in the polarimeters that sugar refineries use. The light that

leaves the syrup is still linearly polarized, but along a turned direction, and the turn grows with the depth of syrup and with the amount of sugar in it. In white light the turn differs from colour to colour, so the filters show colours. Both materials have two speeds of light. They do different things to the light because the two kinds of light that keep their shape inside them are different: two straight lines in the crystal, two circles, one turning each way, in the syrup, whose sugar molecules are handed. A strong magnet laid along the beam makes plain glass behave like the syrup; Faraday found this in 1845.

In the language of physics: the electrons of a medium are held by springs. In a crystal the springs are stiffer along some directions than others. A magnetic field adds a sideways force on a moving electron that couples its motion along x with its motion along y, with a sense of rotation. The whole idea in one sentence: *the medium acts on the wave through a matrix, the polarizations that keep their shape are the eigenvectors of that matrix, and a wave that is not an eigenvector changes shape because its two eigen-components drift out of phase.* Whether that change is a rotation depends only on which eigenvectors the matrix has. The naive guess, that a crystal whose axes are turned would turn the light, fails: the crystal's eigenvectors are still straight lines, and a mixture of two straight lines drifting out of phase becomes an ellipse. To turn the plane and keep it a plane, the eigenvectors must be the two helicities.

One idea, many names. A state that is a mixture of two eigenstates with different eigenvalues changes as the eigenstates drift in phase. Two coupled pendulums pass their swinging back and forth this way; a neutrino born as one flavour turns into another; a spin in a magnetic field precesses. Faraday rotation is the precession of the light's polarization about the direction of the field, and the mathematics is that of a two-state system in quantum mechanics.

The electron on a spring. Let a charge q of mass m be bound to an atom, displaced from its rest position by $\boldsymbol{r}$. It feels a restoring force $-m\,\Omega_0^2\boldsymbol{r}$, where Ω_0^2 is a symmetric 3×3 matrix of spring constants divided by the mass, equal to $\omega_0^2\mathbb{1}$ for a spring that is equally stiff in every direction; a friction force $-m\gamma\dot{\boldsymbol{r}}$; the electric field $\boldsymbol{E}(t)$ of the wave; and, if there is one, a steady magnetic field $\boldsymbol{B}_0$. Newton's law is

$$m\ddot{\boldsymbol{r}} = -m\,\Omega_0^2\boldsymbol{r} - m\gamma\dot{\boldsymbol{r}} + q\boldsymbol{E} + q\,\dot{\boldsymbol{r}}\times\boldsymbol{B}_0. \tag{T1.63}$$

(The magnetic field of the wave itself also pushes, but with a force smaller by the speed of the electron over c, which we neglect.) *How does the charge answer a wave of one frequency?* Write the wave as $\boldsymbol{E}e^{-i\omega t}$ and look for the steady answer $\boldsymbol{r}e^{-i\omega t}$, with the physical quantities the real parts. Then $\dot{\boldsymbol{r}}\to-i\omega\boldsymbol{r}$ and $\ddot{\boldsymbol{r}}\to-\omega^2\boldsymbol{r}$. Without $\boldsymbol{B}_0$,

$$\begin{gathered} -m\omega^2\boldsymbol{r} \overset{(1)}{=} -m\,\Omega_0^2\boldsymbol{r} + im\gamma\omega\,\boldsymbol{r} + q\boldsymbol{E}, \\ \left[\Omega_0^2 - \omega^2\mathbb{1} - i\omega\gamma\mathbb{1}\right]\boldsymbol{r} \overset{(2)}{=} \frac{q}{m}\boldsymbol{E}, \qquad \boldsymbol{r} \overset{(3)}{=} \frac{q}{m}\left[\Omega_0^2 - \omega^2\mathbb{1} - i\omega\gamma\mathbb{1}\right]^{-1}\boldsymbol{E}. \end{gathered} \tag{T1.64}$$

(1) Insert the oscillating forms into (T1.63) with $\boldsymbol{B}_0=0$; the common factor $e^{-i\omega t}$ cancels, and $-m\gamma(-i\omega)\boldsymbol{r} = im\gamma\omega\boldsymbol{r}$.
(2) Move every term with $\boldsymbol{r}$ to the left and divide by m.
(3) Multiply by the inverse matrix, which exists as long as ω is not exactly a resonance of an undamped spring.

With N such charges per unit volume, each displaced by $\boldsymbol{r}$, the medium carries a dipole moment $Nq\boldsymbol{r}$ per unit volume. Writing it as $\varepsilon_0\chi\boldsymbol{E}$ defines the **susceptibility** matrix, and with it the relative permittivity:

$$\chi(\omega) = \omega_p^2\left[\Omega_0^2 - \omega^2\mathbb{1} - i\omega\gamma\mathbb{1}\right]^{-1}, \qquad \varepsilon_r(\omega) = \mathbb{1} + \chi(\omega), \qquad \omega_p^2 \equiv \frac{Nq^2}{\varepsilon_0 m}. \tag{T1.65}$$

The frequency ω_p is the **plasma frequency** of the charges: it measures how many there are and how easily they move.

Which waves can travel? In the medium there are no free charges or currents beyond the bound ones already inside ε_r, so Maxwell's equations read $\nabla\times\boldsymbol{E} = -\partial_t\boldsymbol{B}$ and $\nabla\times\boldsymbol{B} = \mu_0\varepsilon_0\varepsilon_r\,\partial_t\boldsymbol{E}$. Try a plane wave along z, $\boldsymbol{E}\,e^{i(kz-\omega t)}$, for which $\nabla\times$ becomes $ik\,\hat{\boldsymbol{z}}\times$ and ∂_t becomes $-i\omega$:

$$\begin{aligned} \boldsymbol{B} &\overset{(1)}{=} \frac{k}{\omega}\,\hat{\boldsymbol{z}}\times\boldsymbol{E}, \\ ik\,\frac{k}{\omega}\,\hat{\boldsymbol{z}}\times(\hat{\boldsymbol{z}}\times\boldsymbol{E}) &\overset{(2)}{=} -\frac{i\omega}{c^2}\,\varepsilon_r\boldsymbol{E}, \\ \frac{c^2k^2}{\omega^2}\,\boldsymbol{E}_\perp &\overset{(3)}{=} \varepsilon_r\boldsymbol{E}, \qquad \left[\varepsilon_{r\perp} - n^2\mathbb{1}\right]\boldsymbol{E}_\perp \overset{(4)}{=} 0, \qquad n \equiv \frac{ck}{\omega}. \end{aligned} \tag{T1.66}$$

(1) The first Maxwell equation: $ik\,\hat{\boldsymbol{z}}\times\boldsymbol{E} = i\omega\boldsymbol{B}$.
(2) The second, with $\boldsymbol{B}$ from the first line inserted and $\mu_0\varepsilon_0 = 1/c^2$.
(3) The vector identity $\hat{\boldsymbol{z}}\times(\hat{\boldsymbol{z}}\times\boldsymbol{E}) = \hat{\boldsymbol{z}}(\hat{\boldsymbol{z}}\cdot\boldsymbol{E}) - \boldsymbol{E} = -\boldsymbol{E}_\perp$, the part of $\boldsymbol{E}$ across z; then multiply by ic^2/ω.
(4) In both media below ε_r does not mix z with x and y. The z row then reads $\varepsilon_{r,zz}E_z = 0$, so $E_z = 0$, and the x, y rows form a 2×2 problem with the transverse block $\varepsilon_{r\perp}$.

This is an eigenvalue problem. *The polarizations that can travel unchanged are the eigenvectors of $\varepsilon_{r\perp}$, and their refractive indices squared are its eigenvalues.* A medium is **birefringent** when the two eigenvalues differ. Away from a resonance the friction γ can be dropped, the matrix is then Hermitian in both cases below, and the indices are real; a nonzero γ makes them complex, which is absorption, and absorption that differs between the two eigenvectors is called dichroism. Figure T1.24 shows the idea as a sum of the two eigen-polarizations, and Figure T1.25 follows the shape step by step through the medium.

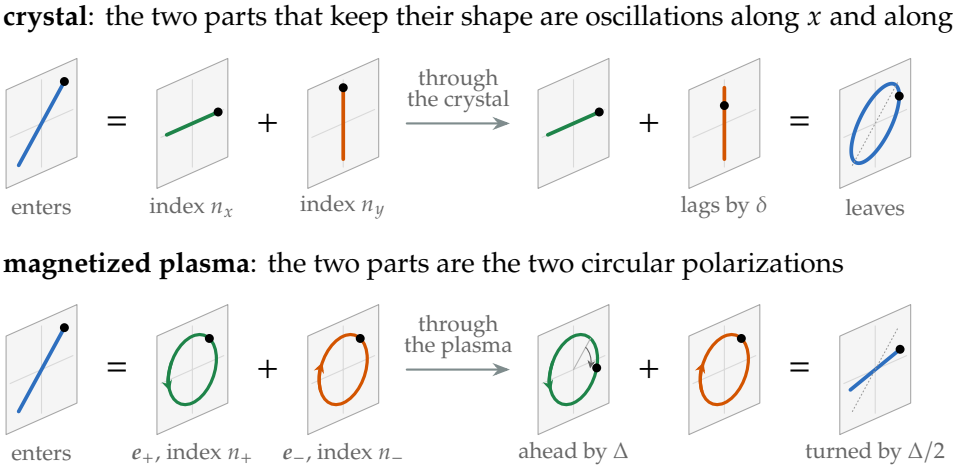


Figure T1.24: What a birefringent medium does: split the wave into the two polarizations that keep their shape, let them drift apart in phase, and add them up again. The dot in each box is the tip of that part of the field at one instant. Top: in a crystal the parts are oscillations along x and y; they enter in step, so their sum is the 45° line, and after the crystal the y part lags by δ (its dot has not yet reached the end when the x dot has), so the sum traces an ellipse instead of a line. Bottom: in a magnetized plasma the parts are the two circular polarizations turning in opposite senses; their sum is a straight line pointing midway between the two dots. One part gets ahead by Δ, the midway direction moves by $\Delta/2$, and the line leaves turned by $\Delta/2$ (dotted: the entering direction).

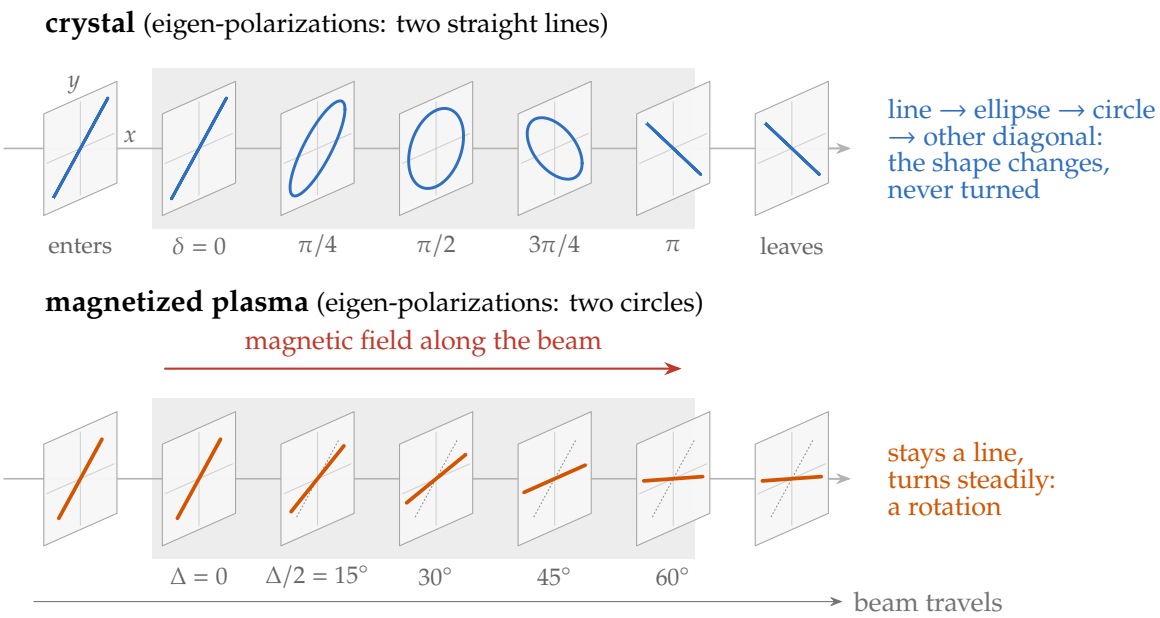


Figure T1.25: The same two media followed step by step along the beam, for a wave that enters linearly polarized at 45°. Each tilted frame is the plane at right angles to the beam at one depth inside the medium (shaded), and shows the curve traced there by the tip of the electric field. In a crystal the two waves that keep their shape are straight lines along x and y; as they drift apart in phase by δ, the line opens into an ellipse, then a circle, then closes again along the other diagonal, but it is never turned. In a magnetized plasma the two waves that keep their shape are the two circular polarizations; as they drift apart by Δ, the wave stays a straight line and turns steadily, by $\Delta/2$ (dotted: the entering direction).

Case A: a crystal. Let the springs have different stiffness along x and y, $\Omega_0^2 = \mathrm{diag}(\omega_{0x}^2, \omega_{0y}^2, \omega_{0z}^2)$, with no magnetic field and no friction. Every matrix in (T1.65) is then diagonal, and so is

$$\varepsilon_{r\perp} = \begin{pmatrix} 1 + \dfrac{\omega_p^2}{\omega_{0x}^2 - \omega^2} & 0 \\ 0 & 1 + \dfrac{\omega_p^2}{\omega_{0y}^2 - \omega^2} \end{pmatrix}. \tag{T1.67}$$

Its eigenvectors are $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}}$, with indices n_x^2 and n_y^2 equal to the two diagonal entries; they differ when $\omega_{0x} \neq \omega_{0y}$. This is **linear birefringence**: a wave polarized along x or along y crosses the crystal unchanged. *What happens to a wave at* 45°*?* At the entrance it is $\boldsymbol{E} = \frac{E_0}{\sqrt{2}}(\hat{\boldsymbol{x}} + \hat{\boldsymbol{y}})$. Each part travels with its own wavenumber $k = n\omega/c$, so after a length L

$$\boldsymbol{E}(L) \overset{(1)}{=} \frac{E_0}{\sqrt{2}}\left(\hat{\boldsymbol{x}}\, e^{in_x\omega L/c} + \hat{\boldsymbol{y}}\, e^{in_y\omega L/c}\right) \overset{(2)}{=} e^{in_x\omega L/c}\,\frac{E_0}{\sqrt{2}}\left(\hat{\boldsymbol{x}} + \hat{\boldsymbol{y}}\, e^{i\delta}\right), \qquad \delta \equiv (n_y - n_x)\frac{\omega L}{c},$$
$$Q \overset{(3)}{=} \tfrac{1}{2}E_0^2 - \tfrac{1}{2}E_0^2 = 0, \qquad U \overset{(4)}{=} 2\cdot\tfrac{1}{2}E_0^2\cos(0-\delta) = E_0^2\cos\delta. \tag{T1.68}$$

(1) Each eigenvector picks up its own phase $kL = n\omega L/c$ and keeps its direction.
(2) Take out the phase of the x part, which is common to the whole wave and invisible in the Stokes numbers.
(3) The definition (T1.33) with $a_1 = a_2 = E_0/\sqrt{2}$.
(4) The same with the phase difference $\theta_1 - \theta_2 = -\delta$.

The tip of the real field traces $E_x \propto \cos\omega t$, $E_y \propto \cos(\omega t - \delta)$: a line for $\delta = 0$, an ellipse with its axes along the two diagonals, a circle at $\delta = \pi/2$, and a line along the other diagonal at $\delta = \pi$

(Figure T1.27(a)). Q stays zero and the linear polarization $|U| = E_0^2\,|\cos\delta|$ shrinks and comes back, while the circular part V, set aside since Section T1.10.1, grows as $\sin\delta$. The stick is never turned; the light becomes partly circular. Even the plate with $\delta = \pi$ that sends the 45° line to −45° is not a rotation: it sends a stick at ψ from the x axis to $-\psi$, so a stick at 10° goes to −10° and one at 30° to −30°. That is a mirror in the x axis, not the same turn for every stick.

Case B: electrons in a magnetic field. Now let the springs be equally stiff in every direction, $\Omega_0^2 = \omega_0^2\mathbb{1}$, and put a steady field $\boldsymbol{B}_0 = B_0\hat{\boldsymbol{z}}$ along the direction of the wave. The only new force in (T1.63) is the magnetic one, $q\,\dot{\boldsymbol{r}} \times \boldsymbol{B}_0 = qB_0\,(\dot{y}, -\dot{x}, 0)$: a charge moving along x is pushed along $-y$, one moving along y is pushed along $+x$. The field therefore *couples* the two directions: the motion along y feeds the equation for x, and the motion along x feeds the equation for y.

The two equations, written out. Insert the oscillating forms $\boldsymbol{r}e^{-i\omega t}$ and $\boldsymbol{E}e^{-i\omega t}$ as before, so that $\dot{x} \to -i\omega x$ and $\dot{y} \to -i\omega y$. The x and y rows of (T1.63) become

$$
\begin{aligned}
-m\omega^2 x &\overset{(1)}{=} -m\omega_0^2 x + im\gamma\omega\, x + qE_x + qB_0(-i\omega y),\\
-m\omega^2 y &\overset{(2)}{=} -m\omega_0^2 y + im\gamma\omega\, y + qE_y - qB_0(-i\omega x),\\
a\,x + i\omega\omega_c\, y &\overset{(3)}{=} \frac{q}{m}E_x, \qquad -i\omega\omega_c\, x + a\,y \overset{(3)}{=} \frac{q}{m}E_y, \qquad a \equiv \omega_0^2 - \omega^2 - i\omega\gamma, \qquad \omega_c \equiv \frac{qB_0}{m},\\
M\begin{pmatrix} x \\ y \end{pmatrix} &\overset{(4)}{=} \frac{q}{m}\begin{pmatrix} E_x \\ E_y \end{pmatrix}, \qquad M = \begin{pmatrix} a & i\omega\omega_c \\ -i\omega\omega_c & a \end{pmatrix}.
\end{aligned}
\tag{T1.69}
$$

(1) The x component: spring, friction and electric force as in Case A, plus the x component of the magnetic force, $qB_0\dot{y}$ with $\dot{y} = -i\omega y$.
(2) The y component: the same, plus the y component of the magnetic force, $-qB_0\dot{x}$ with $\dot{x} = -i\omega x$.
(3) Move every term with x or y to the left and divide by m. In the first equation $-qB_0(-i\omega y)/m$ on the left is $+i\omega\omega_c\, y$; in the second, $+qB_0(-i\omega x)/m$ on the left is $-i\omega\omega_c\, x$.
(4) Write the two equations as one matrix equation: the first equation is the first row of M, the second is the second row.

Without the field ($\omega_c = 0$) M is a times the unit matrix and x and y answer separately. The field adds two off-diagonal entries of opposite sign, an antisymmetric part, which a crystal cannot produce because its spring matrix is symmetric. The quantity ω_c is the **cyclotron frequency**, the rate at which the charge would circle the field line; its sign is set by the charge and the direction of the field.

Which polarizations keep their shape? They are the eigenvectors of M, because $\chi = \omega_p^2 M^{-1}$ by (T1.65) and a matrix and its inverse share their eigenvectors. Find them the standard way, from the eigenvalues:

$$
\begin{aligned}
\det(M - \lambda\mathbb{1}) &\overset{(1)}{=} (a-\lambda)^2 - (i\omega\omega_c)(-i\omega\omega_c) \overset{(2)}{=} (a-\lambda)^2 - \omega^2\omega_c^2 = 0 \quad \Rightarrow \quad \lambda_\pm \overset{(3)}{=} a \mp \omega\omega_c,\\
(M - \lambda_\pm\mathbb{1})\begin{pmatrix} 1 \\ s \end{pmatrix} &\overset{(4)}{=} \begin{pmatrix} \pm\omega\omega_c + i\omega\omega_c\, s \\ -i\omega\omega_c \pm \omega\omega_c\, s \end{pmatrix} = 0 \quad \Rightarrow \quad s \overset{(5)}{=} \pm i, \qquad \boldsymbol{e}_\pm = \frac{\hat{\boldsymbol{x}} \pm i\hat{\boldsymbol{y}}}{\sqrt{2}},\\
n_\pm^2 &\overset{(6)}{=} 1 + \frac{\omega_p^2}{\omega_0^2 - \omega^2 \mp \omega\omega_c} \qquad (\gamma = 0).
\end{aligned}
\tag{T1.70}
$$

(1) The determinant of a 2×2 matrix: product of the diagonal minus product of the off-diagonal entries.
(2) $(i)(-i) = 1$.
(3) $(a-\lambda)^2 = \omega^2\omega_c^2$, so $a - \lambda = \pm\omega\omega_c$.

(4) Apply $M - \lambda_\pm \mathbb{1}$, whose diagonal entries are $a - \lambda_\pm = \pm\omega\omega_c$, to a vector with first entry 1 and second entry s.
(5) The first row gives $\pm 1 + is = 0$, so $s = \pm i$; the second row gives the same. Normalised, the eigenvectors are exactly the helicity vectors of (T1.61).
(6) $\varepsilon_{r\perp} = \mathbb{1} + \omega_p^2 M^{-1}$ has eigenvalues $1 + \omega_p^2/\lambda_\pm$ on these same vectors, and these are $n_\pm^2$ by (T1.66).

The two waves that keep their shape are the two helicities, with different indices $n_+ \neq n_-$ as soon as $\omega_c \neq 0$. This is **circular birefringence**. *What happens to a linear polarization?* By (T1.61), a stick at angle ψ is the equal mixture $(e^{-i\psi}\boldsymbol{e}_+ + e^{i\psi}\boldsymbol{e}_-)/\sqrt{2}$. After a length L each helicity has picked up its own phase $\phi_\pm = n_\pm \omega L/c$, and with $\Delta \equiv \phi_+ - \phi_-$ and $\bar\phi \equiv \frac{1}{2}(\phi_+ + \phi_-)$,

$$\frac{e^{-i\psi}e^{i\phi_+}\boldsymbol{e}_+ + e^{i\psi}e^{i\phi_-}\boldsymbol{e}_-}{\sqrt{2}} \overset{(1)}{=} e^{i\bar\phi}\,\frac{e^{-i(\psi-\Delta/2)}\boldsymbol{e}_+ + e^{i(\psi-\Delta/2)}\boldsymbol{e}_-}{\sqrt{2}} \overset{(2)}{=} e^{i\bar\phi}\left[\cos(\psi - \tfrac{\Delta}{2})\,\hat{\boldsymbol{x}} + \sin(\psi - \tfrac{\Delta}{2})\,\hat{\boldsymbol{y}}\right]. \tag{T1.71}$$

(1) Write $\phi_\pm = \bar\phi \pm \Delta/2$ and take out the common phase $e^{i\bar\phi}$.
(2) (T1.61) with ψ replaced by $\psi - \Delta/2$.

The wave leaves still linearly polarized, with the same strength, and its plane turned by $-\Delta/2$: half the phase difference of the two helicities, as found in Section T1.13.1, and the same turn for every starting angle ψ (Figure T1.27(b)). The sense of the turn is set by the sign of ω_c, that is by the sign of the charge and the direction of the field. This is a rotation, a twist of every stick by the same angle, and it is the rule (T1.35) with $\beta = -\Delta/2$.

How does the angle depend on frequency? Take free electrons in an ionized gas: no spring, $\omega_0 = 0$, charge $q = -e$, and no friction. Then $n_\pm^2 = 1 - \omega_p^2/[\omega(\omega \pm \omega_c)]$, and the gas is transparent only above the plasma frequency. For radio and microwave light crossing interstellar or intergalactic gas the plasma and cyclotron frequencies are in the kilohertz range or below, far beneath the gigahertz of the light, so we expand in the small ratios ω_p/ω and ω_c/ω:

$$\begin{aligned} n_\pm &\overset{(1)}{=} \left[1 - \frac{\omega_p^2}{\omega^2}\left(1 \pm \frac{\omega_c}{\omega}\right)^{-1}\right]^{1/2} \overset{(2)}{=} 1 - \frac{\omega_p^2}{2\omega^2}\left(1 \mp \frac{\omega_c}{\omega}\right), \\ \Delta &\overset{(3)}{=} (n_+ - n_-)\frac{\omega L}{c} \overset{(4)}{=} \frac{\omega_p^2\omega_c}{\omega^3}\,\frac{\omega L}{c} = \frac{\omega_p^2\,\omega_c\,L}{c\,\omega^2} \overset{(5)}{=} \frac{\omega_p^2\,\omega_c\,L}{4\pi^2 c^3}\,\lambda^2. \end{aligned} \tag{T1.72}$$

(1) Take ω^2 out of $\omega(\omega \pm \omega_c)$ and the square root of $n_\pm^2$.
(2) To first order, $(1 \pm x)^{-1} \approx 1 \mp x$ and $(1 - y)^{1/2} \approx 1 - y/2$; products of the two small ratios beyond $\omega_p^2\omega_c/\omega^3$ are dropped.
(3) The definition of Δ with $\phi_\pm = n_\pm \omega L/c$.
(4) In the difference of the two indices the terms $1 - \omega_p^2/2\omega^2$ cancel; what is left is $+\omega_p^2\omega_c/2\omega^3$ in n_+ and $-\omega_p^2\omega_c/2\omega^3$ in n_-, which differ by $\omega_p^2\omega_c/\omega^3$.
(5) $\omega = 2\pi c/\lambda$, so $1/\omega^2 = \lambda^2/(4\pi^2c^2)$.

The index difference falls as ω^{-3}, the path in wavelengths grows as ω, and the phase difference, hence the angle $|\Delta|/2$, grows as λ^2. That is the law of **Faraday rotation**. Inserting $\omega_p^2 = n_e e^2/(\varepsilon_0 m_e)$ for n_e electrons per unit volume and $|\omega_c| = eB_\parallel/m_e$, and adding up the contributions of a gas whose density and field change along the path, the angle is

$$\frac{|\Delta|}{2} = \mathrm{RM}\,\lambda^2, \qquad \mathrm{RM} = \frac{e^3}{8\pi^2\varepsilon_0 m_e^2 c^3}\int n_e B_\parallel \,\mathrm{d}l = 0.812\,\mathrm{rad\,m^{-2}} \int \frac{n_e}{\mathrm{cm^{-3}}}\,\frac{B_\parallel}{\mu\mathrm{G}}\,\frac{\mathrm{d}l}{\mathrm{pc}}, \tag{T1.73}$$

the standard rotation measure, which is what [73] evaluate for a primordial field. The λ^2 law is the high-frequency limit. Figure T1.27(c) compares it with the exact indices of (T1.70) for a

plasma with $\omega_c = 0.2\,\omega_p$: at ten times the plasma frequency the exact angle exceeds the law by 0.54%, at three times by 6.6%, and near the plasma frequency the law fails. For the CMB, more than a million times above the plasma frequency of the gas it crosses, the law is exact for all practical purposes.

Case C: cosmic birefringence. The medium is now empty space filled with the slowly changing pseudoscalar field ϕ of (T1.62). With the coupling $\frac{1}{4} g_{\phi\gamma} \phi F_{\mu\nu} \tilde{F}^{\mu\nu}$, Maxwell's equations again have the two helicities as the waves that keep their shape, and again give them different phases. The difference is how that phase depends on frequency. For a field changing in time at the rate $\dot{\phi}$, the two helicities with the same wavenumber have frequencies that differ by $g_{\phi\gamma}\dot{\phi}$, the same amount for every wavelength [77, 78]. Equivalently, at one frequency their wavenumbers differ by a fixed amount. Over the journey from last scattering the phase difference adds up to $g_{\phi\gamma}(\phi_{\rm obs} - \phi_{\rm LSS})$, and by (T1.71) the plane turns by half of it, which is the twist β of (T1.62) (its sign depends on the conventions for $g_{\phi\gamma}$ and for the angle). In Faraday rotation the splitting of the two indices, $n_+ - n_- \propto \omega^{-3}$, falls faster than the path in wavelengths grows, and the angle goes as λ^2; here the splitting of the wavenumbers is a constant, $n_+ - n_- \propto \omega^{-1}$, which exactly cancels the growth of the path in wavelengths, and the angle is the same at every frequency. This is the difference between the Faraday row and the birefringence row of the table in Section T1.12.5, a λ^2 law against a constant, and it is the reason the colour test of Figure T1.23(b) can separate them. Figure T1.26 shows the two cases side by side for two colours.



(a) Faraday rotation
red light (twice the wavelength)
turns four times as far: $\propto \lambda^2$

(b) cosmic birefringence
every colour turns
by the same angle β

Figure T1.26: Two colours of light, both entering linearly polarized along the dotted line, after crossing (a) a magnetized plasma and (b) the pseudoscalar field of cosmic birefringence. Blue is the shorter wavelength, red the wavelength twice as long. In (a) the red line turns four times as far as the blue, because the Faraday angle grows as λ^2. In (b) the two lines lie on top of each other (red dashed over blue): the angle β is the same at every wavelength.

Birefringence is not dispersion. Birefringence means two different eigenvalues at one frequency: two polarizations travel at two speeds. Dispersion means one eigenvalue that changes with frequency: different colours travel at different speeds. Because χ in (T1.65) depends on ω, the Lorentz model has both, and the Faraday law needs both: the two indices differ, and their difference changes with frequency. Dispersion alone, such as $n^2 = 1 - \omega_p^2/\omega^2$ in a plasma without a magnetic field, delays a pulse but leaves every polarization as it was, because both eigenvalues move together.

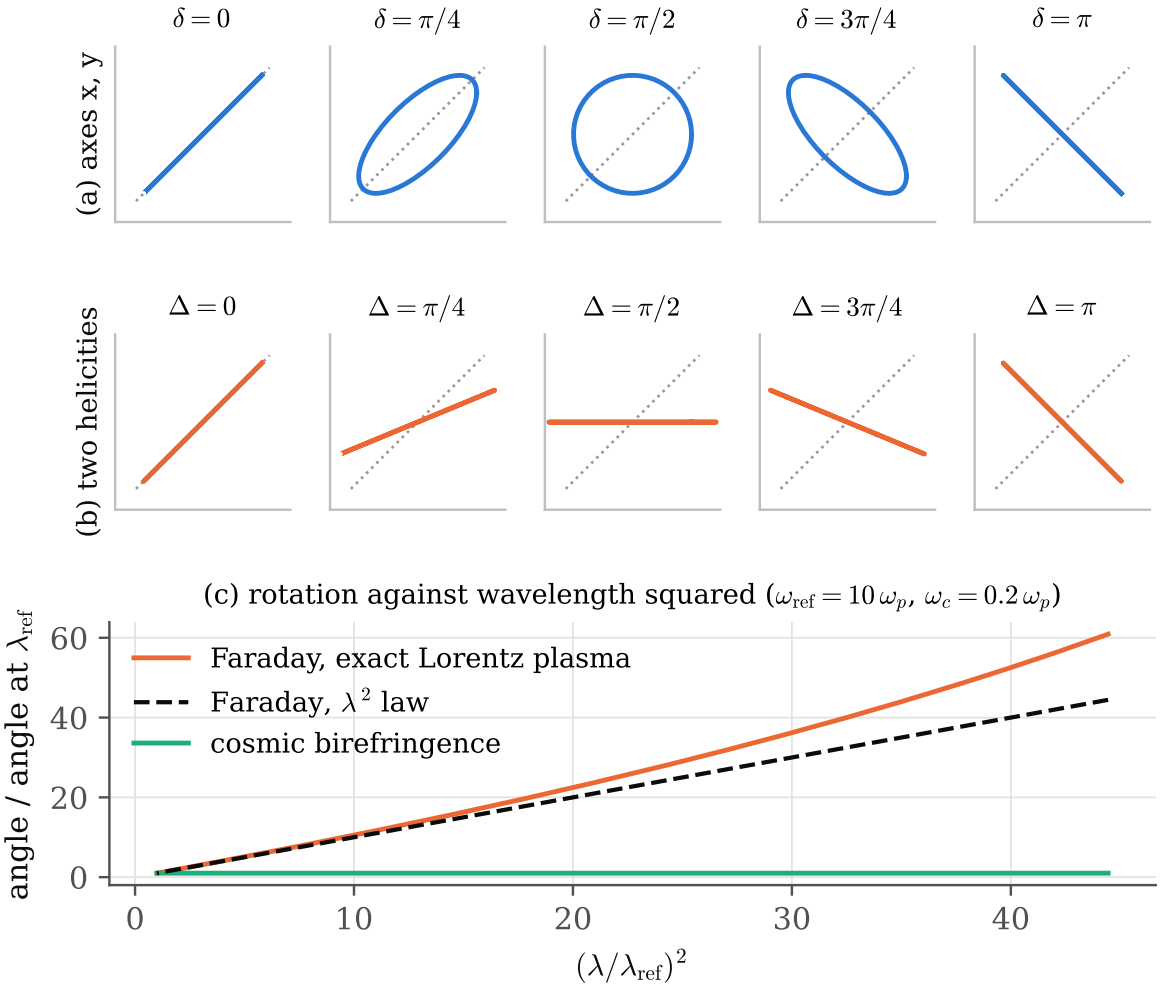


Figure T1.27: Linear against circular birefringence. A wave enters linearly polarized at 45° (dotted). (a) In a crystal, after a phase difference δ between the x and y parts, the tip of the electric field traces an ellipse, then a circle, then the line along the other diagonal; the axis of the ellipse never turns. (b) With the two helicities as the travelling waves, a phase difference Δ leaves the light linearly polarized and turns the line by $\Delta/2$. (c) The angle of turn against the square of the wavelength, each normalised to its value at a reference wavelength: Faraday rotation in a cold magnetized plasma from the exact Lorentz-model indices (orange), its high-frequency λ^2 law (dashed), and cosmic birefringence, the same at every wavelength (green). The exact curve leaves the λ^2 law only as the frequency approaches the plasma frequency.

medium	what makes the two indices differ	waves that keep their shape	a 45° line becomes
crystal	springs of different stiffness	lines along x, y	an ellipse, not turned
magnetized plasma	the force $q\,\dot{\boldsymbol{r}} \times \boldsymbol{B}_0$	the two helicities	a line turned by $\propto \lambda^2$
pseudoscalar field	$\phi F\tilde{F}$ with $\dot{\phi} \neq 0$	the two helicities	a line turned by β, any λ

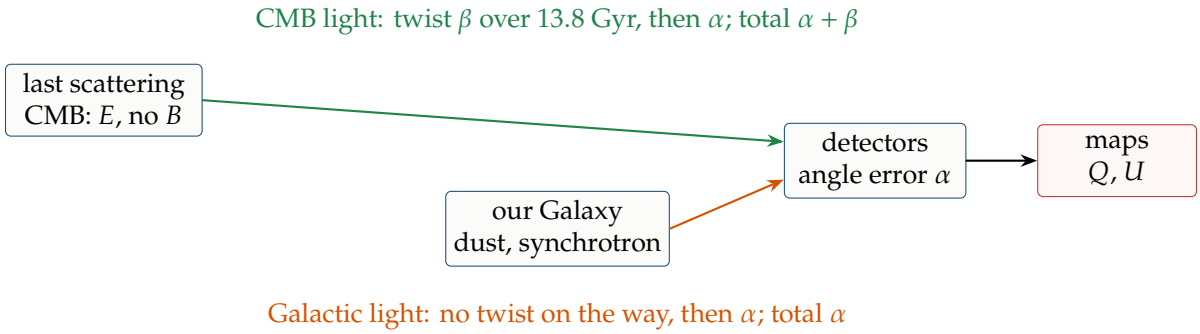


Figure T1.28: The door frame of the microwave sky. CMB light (green) is twisted by β on its long journey and then by the detector angle error α; light from the Galaxy (orange) travels too short a distance to be twisted by β and receives only α. The maps contain both, so the twists of the two components can be measured separately.

T1.13.3 The Galaxy as a reference: the Minami–Komatsu relation

The observed spectra. A map at one frequency contains the CMB, the Galactic foreground ("fg"), and noise. By Figure T1.28, the foreground is twisted by α and the CMB by $\alpha+\beta$, so by (T1.50) applied to each,

$$E^{\rm o}_{\ell m} + iB^{\rm o}_{\ell m} = e^{2i\alpha}\left(E^{\rm fg}_{\ell m} + iB^{\rm fg}_{\ell m}\right) + e^{2i(\alpha+\beta)}\left(E^{\rm CMB}_{\ell m} + iB^{\rm CMB}_{\ell m}\right) + \text{noise}. \tag{T1.74}$$

Assume, as [62, 76] do in their baseline, that neither the foreground nor the CMB has an EB of its own, and that the CMB, the foreground and the noise are uncorrelated, so the cross terms between them vanish on average. Each component then contributes to the averaged spectra exactly as in Section T1.11.2, with its own angle. Writing $\Delta^X_\ell \equiv C^{EE,X}_\ell - C^{BB,X}_\ell$ for the component X, and leaving out the noise, which has $EE = BB$ and no EB,

$$\begin{aligned} C^{EB,{\rm o}}_\ell &\overset{(1)}{=} \tfrac{1}{2}\sin 4\alpha\,\Delta^{\rm fg}_\ell + \tfrac{1}{2}\sin(4\alpha+4\beta)\,\Delta^{\rm CMB}_\ell, \\ C^{EE,{\rm o}}_\ell - C^{BB,{\rm o}}_\ell &\overset{(2)}{=} \cos 4\alpha\,\Delta^{\rm fg}_\ell + \cos(4\alpha+4\beta)\,\Delta^{\rm CMB}_\ell. \end{aligned} \tag{T1.75}$$

(1) (T1.53) for each component, with the angle α for the foreground and $\alpha+\beta$ for the CMB.
(2) $C^{EE\prime} - C^{BB\prime} = \cos 4\beta\,(C^{EE} - C^{BB})$ from Section T1.11.3, again with each component's own angle; the noise has $C^{EE} = C^{BB}$ and drops out.

Removing the unknown foreground. We do not know $\Delta^{\rm fg}_\ell$; modelling the Galaxy is exactly what we want to avoid. But it appears in both lines with known coefficients, so it can be eliminated:

$$\begin{aligned} C^{EB,{\rm o}}_\ell - \tfrac{1}{2}\tan 4\alpha\,\left(C^{EE,{\rm o}}_\ell - C^{BB,{\rm o}}_\ell\right) &\overset{(1)}{=} \tfrac{1}{2}\left[\sin(4\alpha+4\beta) - \tan 4\alpha\cos(4\alpha+4\beta)\right]\Delta^{\rm CMB}_\ell \\ &\overset{(2)}{=} \frac{\sin(4\alpha+4\beta)\cos 4\alpha - \sin 4\alpha\cos(4\alpha+4\beta)}{2\cos 4\alpha}\,\Delta^{\rm CMB}_\ell \\ &\overset{(3)}{=} \frac{\sin 4\beta}{2\cos 4\alpha}\,\Delta^{\rm CMB}_\ell. \end{aligned} \tag{T1.76}$$

(1) Insert both lines of (T1.75); the foreground terms are $\frac{1}{2}\sin 4\alpha\,\Delta^{\rm fg} - \frac{1}{2}\tan 4\alpha\cos 4\alpha\,\Delta^{\rm fg} = 0$, since $\tan 4\alpha\cos 4\alpha = \sin 4\alpha$.
(2) Put both terms over the common denominator $\cos 4\alpha$.
(3) The subtraction formula $\sin(A+B)\cos A - \sin A\cos(A+B) = \sin B$ with $A = 4\alpha$, $B = 4\beta$.

Rearranged, this is the relation of [76] (their eq. 9 without the noise and intrinsic-EB terms):

$$C^{EB,{\rm o}}_\ell = \frac{\tan 4\alpha}{2}\left(C^{EE,{\rm o}}_\ell - C^{BB,{\rm o}}_\ell\right) + \frac{\sin 4\beta}{2\cos 4\alpha}\left(C^{EE,{\rm CMB}}_\ell - C^{BB,{\rm CMB}}_\ell\right). \tag{T1.77}$$

Read it in words. The first term uses the *observed* $EE - BB$ of the map, foreground and all, and its coefficient depends only on α: everything in the map has been twisted by α. The second term uses the *theoretical* CMB spectra from the ΛCDM fit, and its coefficient carries β: only the CMB has the extra twist. Where the foreground is negligible, $C^{EE,\mathrm{o}} - C^{BB,\mathrm{o}} \approx \cos(4\alpha + 4\beta)\Delta^{\mathrm{CMB}}$ has the same shape as the second template and the two terms cannot be separated; that is the degeneracy of Section T1.11.6 again. Where the foreground contributes, the observed $EE - BB$ has a different shape in ℓ (a power law added to the acoustic peaks) and a different size in each frequency channel, and the two coefficients are separately measurable. With no foreground at all and $\beta = 0$, (T1.77) reduces to the sky-by-sky relation $\widehat{C}^{EB} = \frac{1}{2}\tan 4\alpha(\widehat{C}^{EE} - \widehat{C}^{BB})$ verified on simulations in Section T1.11.5.

How well can it work? Figure T1.29 gives a size, from a Fisher forecast in `code/chT1/09_alpha_beta_fisher.py`. The set-up is one frequency channel covering 70% of the sky with white noise of 50 μK-arcmin and a 5′ beam, $51 \le \ell \le 1500$, the CMB fiducial, and Planck's dust over 71% of the sky [71]; the data are the EB spectrum (T1.75), with the cosmic-variance error of each $\widehat{C}_\ell^{EB}$, $[C_\ell^{EE}C_\ell^{BB}/(2\ell+1)f_{\mathrm{sky}}]^{1/2}$ with total spectra, as in Chapter 10. With the CMB alone the Fisher matrix is singular: only the sum is measured, $\sigma(\alpha+\beta) = 0.036°$, the band in the figure. At 143 GHz the dust is faint: it hardly disturbs the CMB, which still fixes $\sigma(\alpha+\beta) = 0.038°$, but it anchors α only weakly, and the band closes into a long, thin ellipse with $\sigma(\beta) = 0.8578°$. At 353 GHz, where the dust is 549 times brighter in power than at 143 GHz, the dust anchors α well, $\sigma(\alpha) = 0.1129°$, but over 70% of the sky its EE and BB exceed those of the CMB at every multipole used, so it also swamps the CMB's EB with chance scatter: the sum is known only to 0.271°, and $\sigma(\beta) = 0.3717°$. One channel faces a trade-off, too little dust to calibrate or so much that it blurs the CMB. In both cases α and β are strongly anti-correlated (correlation coefficients −0.999 and −0.924), because $\alpha+\beta$ is better measured than β itself, and $\beta = (\alpha+\beta) - \alpha$. The real analyses combine several channels, each with its own α_ν and a single β, which is how [62] reach $\sigma(\beta) = 0.14°$; their posterior shows the same anti-correlation. The forecast assumes the dust spectra are known; the measurement does not need them, because (T1.77) uses the observed $EE - BB$.

When the door frame is crooked. The method assumes the Galaxy has no EB of its own. Suppose instead the dust has an intrinsic EB, and write it, as [62] do, as if the dust were twisted at the source by an angle γ: $C_\ell^{EB,\mathrm{dust}} = \frac{1}{2}\sin 4\gamma\, \Delta_\ell^{\mathrm{dust}}$. Then the dust in the map is twisted by $\alpha+\gamma$ in total, and

$$\hat{\alpha} \overset{(1)}{=} \alpha + \gamma, \qquad \hat{\alpha} + \hat{\beta} \overset{(2)}{=} \alpha + \beta \quad \Longrightarrow \quad \hat{\beta} \overset{(3)}{=} \beta - \gamma. \tag{T1.78}$$

(1) The method reads the total twist of the foreground as the detector angle.
(2) The CMB still fixes the sum of the detector angle and its own twist, which the method splits as $\hat{\alpha} + \hat{\beta}$.
(3) Subtract the first equation from the second.

The estimate of β is off by exactly the dust's own handedness. Since the dust TB and TE are both positive, the misalignment picture of Section T1.12.3 predicts $\gamma > 0$, so an analysis that assumes $\gamma = 0$ would underestimate β [62]; but the size and even the sign of the dust EB can change with the sky area [72].

T1.13.4 What has been measured, and why it is not yet a detection

The measurements. With the detector angles fixed at their ground calibration, the Planck team found an angle of $0.31° \pm 0.05°$ (statistical) $\pm 0.28°$ (systematic), where the second error is the calibration uncertainty of α [62]. Applying the relation (T1.77) to the Planck 2018 maps at 100,

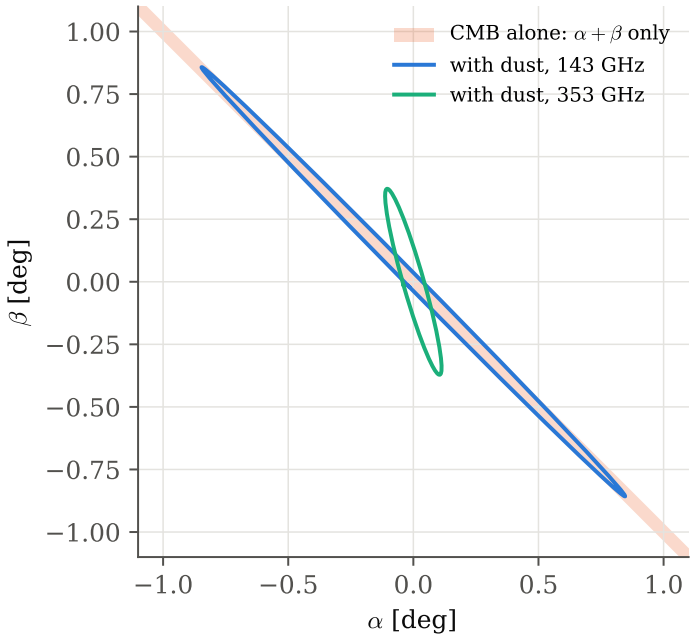


Figure T1.29: Fisher forecast for the detector angle α and the birefringence angle β from the EB spectrum of one channel. With the CMB alone only $\alpha + \beta$ is measured (shaded band, 1σ). Adding the dust of the same map, twisted by α only, closes the band into an ellipse (1σ). At 143 GHz, where the dust is faint, the ellipse is long and stays inside the band: the CMB still fixes the sum, and the dust pins α only weakly. At 353 GHz the bright dust pins α tightly, so the ellipse is short in α, but the same dust adds scatter to the CMB's EB, so the ellipse is wider than the band and β is still uncertain by about a third of a degree.

143, 217 and 353 GHz, fitting one β and four α_ν, [62] found

$$\beta = 0.35^\circ \pm 0.14^\circ \quad (68\%), \tag{T1.79}$$

which excludes zero at 99.2% confidence, a 2.4σ result, with the calibration uncertainty no longer in the error. Repeating the analysis on the reprocessed Planck data (PR4) gave $\beta = 0.30^\circ \pm 0.11^\circ$ for nearly the full sky, with values that decrease as the Galactic mask is enlarged [79]. Adding WMAP and the Planck low-frequency channels, 23 to 353 GHz, gave

$$\beta = 0.342^{\circ\,+0.094^\circ}_{\;-0.091^\circ} \quad (68\%), \tag{T1.80}$$

excluding zero at 99.987% confidence, 3.6σ, with no evidence that β depends on frequency [80]. The independence of frequency argues against Faraday rotation, whose angle would change by a factor $(353/23)^2 \approx 240$ across that range.

Why is it not yet a detection? Treat it as an inference problem, in the order used throughout the book. *The event:* an EB (and TB) with the shape of a twist, at the level of three to four standard deviations, after the detector angles were fitted using the Galaxy. *What could have produced it?* Four candidates.

1. *Chance.* With a 3.6σ result, a fluctuation of the noise and of the cosmic variance is unlikely, but particle physics and cosmology reserve the word "discovery" for 5σ, because many analyses are performed and systematic errors are often underestimated (Chapter 4).
2. *A crooked door frame.* An intrinsic dust EB shifts the estimate to $\beta - \gamma$ (T1.78). Its size is not known well enough: the dust EB predicted from filament misalignment is of the same order as the effect [72], and the dependence of β on the mask found by [79] is what a foreground

effect would produce. Those authors chose not to assign cosmological significance to the value until the foreground polarization is better understood.

3. *The instrument.* Leakage from temperature to polarization and beam effects can create EB and TB (Section T1.12.5). The analyses correct for them with beam models and test the corrections on simulations [62], but the corrections are themselves models.

4. *New physics.* A twist (T1.62) by a pseudoscalar field, $g_{\phi\gamma}\Delta\phi$ of about 10^{-2} radians [62].

Only the last is interesting, and only the last is grey, uniform over the sky and independent of the mask. A detection therefore needs evidence that removes the others: experiments with an independent calibration of α (artificial polarized sources, or well-measured astrophysical ones), a better model of the dust EB from its correlation with TB and TE [72], and agreement between instruments with different detectors and scan strategies.

How this is used in real research: How the measurement is made

The published analyses work with the cross-spectra of maps that share no noise: the two half-mission maps of each frequency channel, in all combinations of channels, so that noise bias drops out [62]. The spectra are measured on the masked sky with a pseudo-C_ℓ code (NaMaster) that corrects for the mask, which otherwise leaks E into B and mixes multipoles (the MASTER method of Chapter 8). Equation (T1.77), written for every pair of channels, is turned into a Gaussian likelihood whose covariance is estimated from the observed spectra, and α_ν, β are sampled with an MCMC (Chapter 9). The pipeline is validated on end-to-end simulations with no twist, which must return β consistent with zero. Our forecast replaces all of this with one channel, a known dust spectrum and a Fisher matrix; it gives the sizes, not the result. Chapter 13 measures the twist on the Planck SMICA maps, where only the sum $\alpha + \beta$ is accessible because a component-separated map has had its Galaxy removed.

One chain of reasoning runs from the sticks to the published angle. Polarization is a field of headless sticks; its split into E and B is a split by behaviour in a mirror; ordinary physics leaves the parity-odd spectra at zero; a uniform twist moves power from E to B and, unlike lensing, leaves a hand behind in EB and TB; and the twist of the instrument is told from the twist of the sky only by a component, the Galaxy, that sees one of them and not the other.

Recap

- The polarization stick is $Q + iU = Pe^{2i\psi}$; a twist by β multiplies it by $e^{2i\beta}$, and turning the axes by α multiplies it by $e^{-2i\alpha}$.
- E and B are Q and U read in axes aligned with each wavevector; E is mirror symmetric, B changes sign, so a mirror-symmetric universe has $EB = TB = 0$.
- Scalar perturbations make only E; gravitational waves make E and B.
- A uniform twist rotates $(E_{\ell m}, B_{\ell m})$ by 2β and gives $C^{EE\prime} = \cos^2 2\beta\, C^{EE} + \sin^2 2\beta\, C^{BB}$, $C^{BB\prime} = \sin^2 2\beta\, C^{EE} + \cos^2 2\beta\, C^{BB}$, $C^{EB\prime} = \frac{1}{2}\sin 4\beta (C^{EE} - C^{BB})$, $C^{TE\prime} = \cos 2\beta\, C^{TE}$, $C^{TB\prime} = \sin 2\beta\, C^{TE}$; total polarized power is conserved.
- Lensing also turns E into B, $C^{BB} = \int \frac{\mathrm{d}^2\ell'}{(2\pi)^2}[\boldsymbol{\ell}' \cdot (\boldsymbol{\ell} - \boldsymbol{\ell}')]^2 \sin^2 2\Delta\varphi\, C^{\phi\phi} C^{EE}$, white at low ℓ ($\approx 4.7\,\mu$K-arcmin), but leaves $EB = TB = 0$: no hand.
- Fingerprints of B: frequency separates the Galaxy and Faraday rotation from the CMB; the ℓ shape separates gravitational waves from lensing; EB, TB separate twists from parity-preserving sources.

- A detector angle error α adds to β; the Galaxy is twisted by α only, which gives the Minami–Komatsu relation (T1.77); dust EB shifts the result to $\beta - \gamma$.
- Measured: $\beta = 0.35^\circ \pm 0.14^\circ$ (Planck 2018), $0.30^\circ \pm 0.11^\circ$ (PR4), $0.342^{\circ\,+0.094^\circ}_{-0.091^\circ}$ (WMAP + Planck); not yet a detection because of the dust EB and instrumental systematics.

Chapter summary

The cosmological model is the first step of the book's pipeline. For every choice of the six ΛCDM numbers it fixes the expected value of each summary that the later chapters estimate from data, and never the sky itself. The route is short enough to hold in the head. The Friedmann equation gives the expansion rate $E(z)$, and one integral of $1/E$ gives the distances D_A and D_L. The primordial spectrum $\mathcal{P}_\zeta = A_s(k/k_*)^{n_s-1}$ is carried to the sky by the radiation transfer function, $C_\ell = \frac{2}{\pi}\int k^2 \mathrm{d}k\, P_\zeta\, g_{T\ell}^2$, and to the matter field by growth and the matter transfer function, $P_\delta(k, a_0) = P_\delta(k, a_\mathrm{i})\, T^2[D_+(a_0)/D_+(a_\mathrm{i})]^2$. The data are one draw from each: the $a_{\ell m}$ with variance C_ℓ, the Fourier modes $\delta_{\boldsymbol{k}}$ with variance $P_\delta(k)$. Weak lensing adds a third field built from the same matter: projected along the line of sight with the lensing efficiency $g(\chi)$, it becomes the convergence κ, whose power spectrum $C_\ell^{\kappa\kappa}$ is a Limber integral of P_δ, and it is read off galaxy shapes through the shear γ, buried under shape noise. The CMB light is also polarized, a field of headless sticks split into a mirror-symmetric E and a mirror-odd B: ordinary physics leaves $EB = TB = 0$, gravitational waves and lensing make B without a hand, and a uniform twist of the polarization, by new physics (β) or by a miscalibrated detector (α), moves power from E to B and creates EB and TB.

What carries into the inference chapters is how these predictions move when the parameters move. The finite-difference responses $R_i(\ell)$ of Eq. (T1.1) are the derivatives of the Fisher matrix, two nearly parallel responses are a degeneracy, and the line between linear and halofit $P(k)$ is where the Gaussian description of the matter field stops. Conventions are the fixed factors that decide whether a script reproduces these numbers. The table lists what to carry forward, what each item means in practice, and where it is used.

concept	operational meaning	used later in
model as a machine	six numbers into CAMB, expected statistics out; call it through `code/camb_fiducial.py` for the fiducial spectra	Chapters 7 to 10
one realisation	the sky is one draw; each $a_{\ell m}$ has variance C_ℓ, and $\widehat{C}_\ell$ scatters about C_ℓ as $\chi^2_{2\ell+1}/(2\ell+1)$	Chapters 7 and 8
distances and $H(z)$	$E(z)$ from the density parameters, then D_A, D_L by one integral; the model mean of supernova and BAO data	Chapters 3 and 5
conventions	raw C_ℓ vs D_ℓ; $\mu\mathrm{K}^2$; $A_s = \frac{25}{4}A_S^2$; $h^{-1}\mathrm{Mpc}$; $\phi \equiv \psi$	every CMB and LSS script
fractional response $R_i(\ell)$	central difference of C_ℓ over $\pm1\sigma$, divided by $2C_\ell$; times C_ℓ/δ_i it is $\partial C_\ell/\partial\theta_i$	Chapter 10
degeneracy	two nearly parallel response curves (A_s and τ above $\ell \approx 30$); a near-singular Fisher matrix and a ridge in the chain	Chapters 9 and 10
linear vs halofit $P(k)$	agree on large scales, differ by 10% at $k \approx 0.15\,h\,\mathrm{Mpc}^{-1}$ at $z = 0$; beyond that the field is non-Gaussian	Chapters 7 and 12
$\xi(r)$ and the BAO bump	Fourier partner of $P(k)$; a peak near $r_s(z_{\mathrm{drag}})$, a standard ruler	Chapters 7 and 12
σ_8	top-hat variance of the linear field at $8\,h^{-1}\mathrm{Mpc}$; a check that $P(k)$ is normalised correctly	Chapters 7 and 10
f_{NL}	amplitude of the primordial bispectrum; invisible to C_ℓ and $P(k)$	Chapter 12
κ, γ, E/B	projected density and its spin-2 tidal stretch (Section T1.5); lensing makes only E, so B is a null test	Chapter 12
$C_\ell^{\kappa\kappa}$ and shape noise	Limber integral of $[g/a]^2 P_\delta$ (CAMB lensing windows); measured power $C_\ell^{ij} + \delta_{ij}\sigma_\epsilon^2/\bar{n}_i$	Chapters 10 and 12
S_8 and nuisances	cosmic shear fixes $\sigma_8\Omega_m^\alpha$, $\alpha \approx 0.55$; photo-z, shear bias, intrinsic alignment, baryons are marginalised	Chapters 9, 10 and 12
E, B and parity	Q, U read along each wavevector; B changes sign in a mirror, so $EB = TB = 0$ unless something picks a hand (Section T1.10.4)	Chapters T3 and 13
rotated spectra	a twist β gives $EB = \frac{1}{2}\sin 4\beta\,(C^{EE} - C^{BB})$, $TB = \sin 2\beta\, C^{TE}$, BB gains $\sin^2 2\beta\, C^{EE}$; a detector angle error α enters as $\alpha + \beta$	Chapter 13
lensing B modes	$C^{BB} = \int [\boldsymbol{\ell}' \cdot (\boldsymbol{\ell} - \boldsymbol{\ell}')]^2 \sin^2 2\Delta\varphi\, C^{\phi\phi} C^{EE}$; white at low ℓ, peak near $\ell \approx 1000$; no EB, TB; non-Gaussian	Chapters 10 and 13
B-mode fingerprints	frequency (Galaxy vs CMB), ℓ shape (gravitational waves vs lensing), parity (EB, TB: twists vs the rest); table in Section T1.12.5	Chapter 13
Minami–Komatsu relation	the Galaxy is twisted by α only: $C^{EB,\mathrm{o}} = \frac{1}{2}\tan 4\alpha\,(C^{EE,\mathrm{o}} - C^{BB,\mathrm{o}}) + \frac{\sin 4\beta}{2\cos 4\alpha}(C^{EE} - C^{BB})^{\mathrm{CMB}}$; dust EB biases β to $\beta - \gamma$	Chapter 13

Reading the literature. The physics of every object above is developed in the cosmology notes. The Boltzmann code used throughout is CAMB [8]. The parameter values and errors are those of Planck 2018 [25]. The non-linear correction is the halofit version of [41], and the first detection of the BAO bump in the galaxy correlation function is [42]. For weak lensing, the standard reviews are [43] and [44], CMB lensing is reviewed in [47] and measured in [48], and the Stage-III cosmic-shear results are [45, 49, 50]. For CMB polarization, the E/B decomposition is due to [63, 64], with a gentle introduction in [65]; gravitational-wave B modes are predicted in [66], lensing B modes in [69], and experimental systematics are catalogued in [74]. The physics of cosmic birefringence is in [67, 77, 78]; the method that uses the Galaxy to calibrate the detector angles is [76], applied in [62, 79, 80], with the dust EB caveat of [72]. The polarized foregrounds are characterised in [71] and the current limit on r is [70].

Chapter T2

Astrophysics for the Data Analyst

Where this chapter is going

Two black holes orbit each other. Every half orbit the pair looks the same again, a dumbbell turning in space, and a turning dumbbell shakes spacetime. The shaking spreads out as a wave, the wave carries energy, and the energy comes out of the orbit. So the two holes sink closer, circle faster, and shake spacetime harder. The wave that reaches us rises in pitch and in loudness until the holes merge: a *chirp*. In the band of the space detector LISA a light black hole can circle a heavy one millions of times before the end, and each circle adds a predictable amount to the phase of the wave. If the pair lives inside a cloud of dark matter, the light hole feels a drag and sinks a little faster than it would in empty space. After a million orbits a tiny drag becomes a phase lag of many radians, and a phase lag is something a detector can measure.

Turning that lag into a number with an error bar is a statistics problem, and Chapter 10 solves it: the likelihood of a signal buried in noise, the matched filter, the Fisher matrix and the forecast. Every one of those calculations needs a short list of physical objects. It needs the signal, written as a Fourier transform $\tilde{h}(f)$ whose phase $\Psi(f)$ depends on the masses, the arrival time t_c and the phase ϕ_c. It needs the noise, as one function $S_n(f)$ of frequency. It needs the numbers that connect them, the factors $4/5$ and $3/10$ that turn a real source on the sky into an average one. And it needs the form in which any departure from the vacuum chirp enters the phase: a power of the frequency, βf^b. None of these is a matter of convention. Each follows from two physical facts, Einstein's equation for a weak gravitational field and the energy that a gravitational wave carries, together with Newton's law of gravity for the orbit; everything else can be derived from them.

Figure T2.1 shows how the ideas feed into each other. A small ripple h on flat spacetime obeys a wave equation and has exactly two polarisations, h_+ and $h_\times$, each of which stretches a ring of free masses in one direction while squeezing it in the other (Section T2.1). A source makes such a ripple only if its mass distribution changes shape: the total mass and the motion of the centre of mass are fixed by conservation laws, so the first thing that can radiate is the quadrupole, and for a binary it radiates at twice the orbital frequency (Section T2.2). The energy the wave carries away drives the chirp: one formula for $\dot{f}$, a time to merger t_c and a phase at merger ϕ_c (Section T2.4). Its Fourier transform, found by the stationary-phase approximation, is the template $\tilde{h}(f)$ of every forecast (Section T2.5). Corrections from general relativity and from the environment of the binary enter the Fourier phase as extra powers of f, which is what makes one forecast cover many kinds of physics (Section T2.6). A detector compares the light travel times along two arms, and so measures one combination of h_+ and $h_\times$ with weights that depend on where the source sits on the sky (Section T2.3). Finally, LISA's own noise, from photon counting, from stray forces on its test masses and from millions of unresolved Galactic

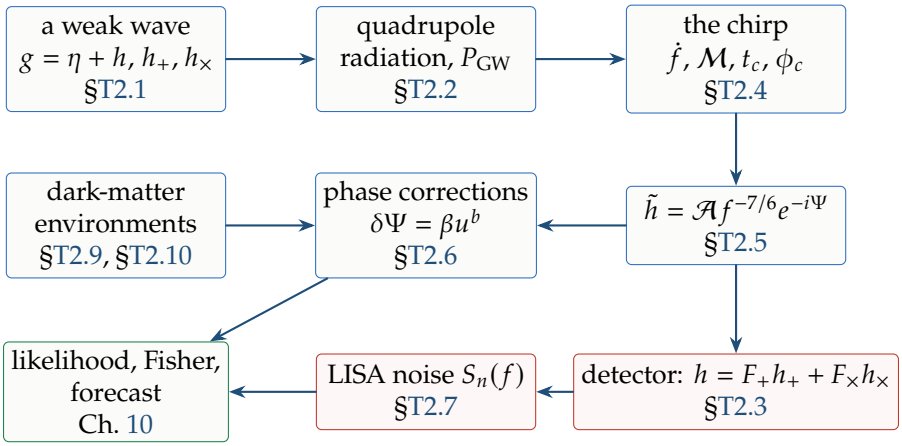


Figure T2.1: How the pieces of the chapter connect. Blue boxes are the physics of the source: a weak wave, how a moving mass makes one, how the orbit shrinks, and what the wave looks like in the frequency domain, including the small power-law terms that an environment adds to its phase. Red boxes are the detector: how it turns a wave into one number per instant, and what noise it adds. The green box is the statistics of Chapter 10, which takes the signal model and the noise spectrum and runs the chain backwards to learn about the source.

binaries, is summarised by the spectrum $S_n(f)$ (Section T2.7). Table T2.1 lists every symbol, where it is derived, and where Chapter 10 uses it.

The same reading applies to the other astrophysical data in the book. The dark-matter environments that change the chirp are built from their own physical inputs in Sections T2.9 and T2.10. Distances from standard candles, the distance ladder, selection effects and galaxy surveys follow in Section T2.11 onwards. Each time the question is the same: what physical object is there, what is measured, what is the noise, and which statistical tool turns the measurement into physics. In the language of the book's pipeline (Chapter 0), the gravitational-wave case reads as follows. The physical model is the binary and its environment. The data set is the strain time series. What matters in it is the phase of the wave. The sampling distribution is that of the detector noise, fixed by $S_n(f)$. Inference then asks the question of Chapter 5: given the recorded data, which masses, distance and environmental strength could have produced it, and how plausible is the noise realisation each of them requires?

T2.1 What a gravitational wave is

In Newtonian gravity a change in the position of a source changes its pull everywhere at once. In relativity the disturbance takes time to propagate. When two black holes orbit one another, for example, the changing distribution of mass produces a ripple that travels outwards in the gravitational field. By the time it reaches a distant detector the ripple is weak enough to treat as a small perturbation.

We write that perturbation as a correction $h_{\mu\nu}$ to the flat spacetime metric. In empty space it obeys a wave equation with speed c and has two independent polarisations, h_+ and $h_\times$. To connect those statements to a measurement, consider a ring of freely falling test masses. As a wave passes through the ring along its axis, separations grow in one transverse direction while shrinking in the perpendicular direction. A detector measures this relative change in distance.

T2.1.1 The inputs

As in the derivation of the Poisson law in Chapter 2, we first say what goes in. Four things do. Two are physics, one is a statement about coordinates, and one is a physical result we import.

1. *Physical: the field is weak.* Far from the source spacetime is nearly flat. We write the metric as
$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}, \qquad \eta_{\mu\nu} = \mathrm{diag}(-1,1,1,1), \qquad |h_{\mu\nu}| \ll 1, \tag{T2.1}$$
with coordinates $x^\mu = (ct, x, y, z)$, and keep everything to first order in h. Indices are raised and lowered with η.

2. *Physical: Einstein's equation, to first order in h.* Define the **trace-reversed perturbation** $\bar{h}_{\mu\nu} = h_{\mu\nu} - \frac{1}{2}\eta_{\mu\nu}h$, with $h = \eta^{\mu\nu}h_{\mu\nu}$ [the same ten numbers, recombined so that the field equation becomes simple]. If the coordinates are chosen so that $\partial^\mu \bar{h}_{\mu\nu} = 0$ (the **Lorenz gauge**), Einstein's equation linearised in h reads
$$\Box\, \bar{h}_{\mu\nu} = -\frac{16\pi G}{c^4} T_{\mu\nu}, \qquad \Box \equiv -\frac{1}{c^2}\frac{\partial^2}{\partial t^2} + \nabla^2, \tag{T2.2}$$
where $T_{\mu\nu}$ is the energy–momentum tensor of the matter [81, Sec. 1.1]. It is the twin of the equation $\Box A^\mu = -\mu_0 J^\mu$ for the electromagnetic potentials in the Lorenz gauge, with the energy–momentum of matter in place of the electric current.

3. *Mathematical: coordinates are labels.* Relabelling the points of spacetime by $x^\mu \to x^\mu + \xi^\mu(x)$, with a small ξ, changes the components of the metric but not the spacetime. To first order it changes $h_{\mu\nu} \to h_{\mu\nu} - \partial_\mu \xi_\nu - \partial_\nu \xi_\mu$. Anything measurable, such as the distance a ruler records between two free masses, is unchanged.

4. *Physical: the energy flux.* A gravitational wave carries energy; the formula for the flux is imported from general relativity in Section T2.1.4 and stated there.

Several things often listed as facts about gravitational waves are *not* inputs: that they travel at c, that they are transverse, that there are two polarisations, and the pattern in which they move test masses. All four are derived below.

T2.1.2 Waves in empty space, and how many kinds there are

The wave equation. Far from the source there is no matter, $T_{\mu\nu} = 0$, and (T2.2) becomes $\Box \bar{h}_{\mu\nu} = 0$. Written out, $\partial_t^2 \bar{h}_{\mu\nu} = c^2 \nabla^2 \bar{h}_{\mu\nu}$: each of the ten components obeys the ordinary wave equation with speed c. So a disturbance in the gravitational field travels at the speed of light, and nothing more had to be assumed to get this.

A plane wave. Any solution is a superposition of plane waves, so it is enough to study one,
$$\bar{h}_{\mu\nu} = \mathrm{Re}\left[\varepsilon_{\mu\nu}\, e^{ik_\alpha x^\alpha}\right], \qquad k^\alpha = \left(\frac{\omega}{c}, \boldsymbol{k}\right), \tag{T2.3}$$
with a constant symmetric matrix $\varepsilon_{\mu\nu}$ of amplitudes. Inserting it into $\Box \bar{h}_{\mu\nu} = 0$ brings down $-k_\alpha k^\alpha = \omega^2/c^2 - |\boldsymbol{k}|^2$, which must vanish: $\omega = c|\boldsymbol{k}|$, the dispersion relation of a wave moving at c.

How many of the ten amplitudes are real? A symmetric 4×4 matrix has ten entries. Not all of them describe something physical, because some can be changed by relabelling coordinates (input 3). We count them for a wave travelling along z, so that $k^\alpha = (K, 0, 0, K)$ and $k_\alpha = (-K, 0, 0, K)$ with $K = \omega/c$.

First, the Lorenz condition. The gauge condition $\partial^\mu \bar{h}_{\mu\nu} = 0$ becomes $k^\mu \varepsilon_{\mu\nu} = 0$, that is $K\varepsilon_{0\nu} + K\varepsilon_{3\nu} = 0$, or

$$\varepsilon_{3\nu} = -\varepsilon_{0\nu} \qquad (\nu = 0, 1, 2, 3). \tag{T2.4}$$

Four conditions, so six amplitudes are left.

Second, the relabelling that is still allowed. A relabelling with $\Box\xi_\mu = 0$ does not spoil the Lorenz condition, because, using input 3 and its trace-reversed form,

$$\begin{aligned} \delta\bar{h}_{\mu\nu} &\overset{(1)}{=} -\partial_\mu\xi_\nu - \partial_\nu\xi_\mu + \eta_{\mu\nu}\,\partial^\alpha\xi_\alpha, \\ \partial^\mu\delta\bar{h}_{\mu\nu} &\overset{(2)}{=} -\Box\xi_\nu - \partial_\nu\partial^\mu\xi_\mu + \partial_\nu\partial^\alpha\xi_\alpha \overset{(3)}{=} -\Box\xi_\nu. \end{aligned} \tag{T2.5}$$

(1) The trace of $\delta h_{\mu\nu} = -\partial_\mu\xi_\nu - \partial_\nu\xi_\mu$ is $\delta h = -2\partial^\alpha\xi_\alpha$, so $\delta\bar{h}_{\mu\nu} = \delta h_{\mu\nu} - \frac{1}{2}\eta_{\mu\nu}\delta h$ gains the last term.
(2) Take ∂^μ of each term; $\partial^\mu\partial_\mu = \Box$, and $\partial^\mu\eta_{\mu\nu} = \partial_\nu$.
(3) The last two terms cancel.

So any ξ_μ that itself solves the wave equation is still allowed. Take one with the same wave vector, $\xi_\mu = \mathrm{Re}[-i\zeta_\mu e^{ik_\alpha x^\alpha}]$, with four constants ζ_μ; then $\partial_\nu\xi_\mu = \mathrm{Re}[k_\nu\zeta_\mu e^{ik\cdot x}]$, and the first line of (T2.5) shifts the amplitudes by

$$\delta\varepsilon_{\mu\nu} = -k_\mu\zeta_\nu - k_\nu\zeta_\mu + \eta_{\mu\nu}\,(k^\alpha\zeta_\alpha), \qquad k^\alpha\zeta_\alpha = K(\zeta_0 + \zeta_3). \tag{T2.6}$$

Look at what the four constants can do. $\delta\varepsilon_{01} = -k_0\zeta_1 = K\zeta_1$, so ζ_1 can cancel ε_{01}; likewise ζ_2 cancels ε_{02}. $\delta\varepsilon_{03} = -k_0\zeta_3 - k_3\zeta_0 = K(\zeta_3 - \zeta_0)$, and the trace changes by $\eta^{\mu\nu}\delta\varepsilon_{\mu\nu} = -2k\cdot\zeta + 4k\cdot\zeta = 2K(\zeta_0 + \zeta_3)$. The combinations $\zeta_3 - \zeta_0$ and $\zeta_3 + \zeta_0$ are independent, so they cancel ε_{03} and the trace. Four constants, four conditions: we can always reach

$$\varepsilon_{01} = \varepsilon_{02} = \varepsilon_{03} = 0, \qquad \eta^{\mu\nu}\varepsilon_{\mu\nu} = 0. \tag{T2.7}$$

What is left. Combine (T2.7) with the Lorenz condition (T2.4). For $\nu = 1, 2$: $\varepsilon_{31} = \varepsilon_{32} = 0$. For $\nu = 3$: $\varepsilon_{33} = -\varepsilon_{03} = 0$. For $\nu = 0$: $\varepsilon_{00} = -\varepsilon_{30} = 0$. Only ε_{11}, ε_{12} and ε_{22} survive, and the zero trace, $-\varepsilon_{00} + \varepsilon_{11} + \varepsilon_{22} + \varepsilon_{33} = 0$, forces $\varepsilon_{22} = -\varepsilon_{11}$. Two numbers remain. Since the trace is zero, $\bar{h}_{\mu\nu} = h_{\mu\nu}$, and the wave is

$$h_{ij}^{\mathrm{TT}}(t, z) = \begin{pmatrix} h_+ & h_\times & 0 \\ h_\times & -h_+ & 0 \\ 0 & 0 & 0 \end{pmatrix} \left(t - \frac{z}{c}\right), \qquad h_{0\mu}^{\mathrm{TT}} = 0. \tag{T2.8}$$

We wrote $h_+(t - z/c)$ and $h_\times(t - z/c)$ for arbitrary wave forms, because a superposition of plane waves with the same direction is any function of $t - z/c$. Ten amplitudes, minus four Lorenz conditions, minus four allowed relabellings, leave two: the **plus** and **cross polarisations** [81, Sec. 1.2].

Working definition T2.1 (transverse-traceless gauge)

For a wave travelling along a unit vector $\hat{\boldsymbol{n}}$, the **TT gauge** is the choice of coordinates in which $h_{0\mu} = 0$, $h_{ij}\hat{n}_j = 0$ [no component along the direction of travel] and $h_{ii} = 0$ [no overall expansion]. In it the wave is the 2×2 matrix $\left(\begin{smallmatrix} h_+ & h_\times \\ h_\times & -h_+ \end{smallmatrix}\right)$ in the plane perpendicular to $\hat{\boldsymbol{n}}$. *Operationally:* given any symmetric 3×3 matrix M and the transverse unit vectors $\boldsymbol{p}, \boldsymbol{q}$, its TT part has $h_+ = \frac{1}{2}(M_{pp} - M_{qq})$ and $h_\times = M_{pq}$, where $M_{pq} = p_i M_{ij} q_j$.

The operational recipe in the box is just (T2.8) read backwards: keep the block of M in the plane spanned by $\boldsymbol{p}$ and $\boldsymbol{q}$, and subtract half its trace from the diagonal so that it becomes traceless, which leaves $\frac{1}{2}(M_{pp} - M_{qq})$ on the diagonal and M_{pq} off it. We use it in Section T2.2.

Example: turning the axes. The labels + and × refer to the axes x, y we chose in the transverse plane. Suppose we turn those axes by an angle α about the direction of travel. The 2×2 block of (T2.8) is $h_+\sigma_3 + h_\times\sigma_1$, with $\sigma_3 = \left(\begin{smallmatrix}1 & 0\\ 0 & -1\end{smallmatrix}\right)$ and $\sigma_1 = \left(\begin{smallmatrix}0 & 1\\ 1 & 0\end{smallmatrix}\right)$. Its components in the turned axes are $R^{\mathsf T} h R$ with R the rotation matrix by α, and

$$R^{\mathsf T}\sigma_3 R \overset{(1)}{=} \cos 2\alpha\,\sigma_3 - \sin 2\alpha\,\sigma_1, \qquad R^{\mathsf T}\sigma_1 R \overset{(1)}{=} \sin 2\alpha\,\sigma_3 + \cos 2\alpha\,\sigma_1,$$
$$h'_+ \overset{(2)}{=} h_+\cos 2\alpha + h_\times \sin 2\alpha, \qquad h'_\times \overset{(2)}{=} -h_+\sin 2\alpha + h_\times\cos 2\alpha. \tag{T2.9}$$

(1) Multiply out with $R = \left(\begin{smallmatrix}\cos\alpha & -\sin\alpha\\ \sin\alpha & \cos\alpha\end{smallmatrix}\right)$ and use $\cos^2\alpha - \sin^2\alpha = \cos 2\alpha$, $2\sin\alpha\cos\alpha = \sin 2\alpha$.
(2) Collect the coefficients of σ_3 (the new h_+) and of σ_1 (the new $h_\times$).

The polarisations mix with *twice* the angle. A turn by 45° turns a pure h_+ into a pure $h_\times$, and a turn by 90° only flips the sign; a turn by 180° changes nothing. This is the signature of a field of spin two, and it is why the orientation of the detector and the orientation of the source on the sky enter the detected signal through angles 2ϕ and 2ψ (Section T2.3).

T2.1.3 What a wave does to free masses

Do the masses move? Place a free test mass at rest at some point and let the wave (T2.8) pass. The mass follows a geodesic, $\mathrm{d}^2x^\mu/\mathrm{d}\tau^2 = -\Gamma^\mu_{\alpha\beta}(\mathrm{d}x^\alpha/\mathrm{d}\tau)(\mathrm{d}x^\beta/\mathrm{d}\tau)$. While it is at rest only $\mathrm{d}x^0/\mathrm{d}\tau$ is nonzero, so its acceleration is set by Γ^i_{00} alone:

$$\frac{\mathrm{d}^2x^i}{\mathrm{d}\tau^2} \overset{(1)}{=} -\Gamma^i_{00}\left(\frac{\mathrm{d}x^0}{\mathrm{d}\tau}\right)^2, \qquad \Gamma^i_{00} \overset{(2)}{=} \tfrac12\left(2\partial_0 h_{i0} - \partial_i h_{00}\right) \overset{(3)}{=} 0. \tag{T2.10}$$

(1) Only the $\alpha=\beta=0$ term survives when the spatial velocity is zero.
(2) The Christoffel symbol to first order in h, $\Gamma^i_{\alpha\beta} = \tfrac12(\partial_\alpha h_{i\beta} + \partial_\beta h_{i\alpha} - \partial_i h_{\alpha\beta})$, at $\alpha=\beta=0$.
(3) In the TT gauge every $h_{0\mu}$ vanishes, (T2.8).

So the mass stays at rest, and keeps its coordinates, for as long as the wave lasts. The TT coordinates are, in effect, painted onto the free masses. Yet something does change: the distance between two masses, as a ruler or a light signal would measure it.

The distances do change. Take one mass at the origin and another at coordinates $\boldsymbol{x}_0$, a distance much smaller than the wavelength so that h_{ij} is the same all the way between them. The proper distance follows from $\mathrm{d}s^2 = (\delta_{ij} + h_{ij})\mathrm{d}x^i\mathrm{d}x^j$. To first order the matrix $\delta_{ij} + h_{ij}$ is a perfect square,

$$\mathrm{d}s^2 \overset{(1)}{=} \mathrm{d}\boldsymbol{x}^{\mathbb T}(\mathbb 1 + h)\,\mathrm{d}\boldsymbol{x} \overset{(2)}{=} \mathrm{d}\boldsymbol{x}^{\mathbb T}\left(\mathbb 1 + \tfrac12 h\right)^2\mathrm{d}\boldsymbol{x} \overset{(3)}{=} \left|\mathrm{d}\boldsymbol{X}\right|^2, \qquad \boldsymbol{X} \equiv \left(\mathbb 1 + \tfrac12 h\right)\boldsymbol{x}. \tag{T2.11}$$

(1) The spatial part of the TT metric, in matrix form.
(2) $(\mathbb 1 + \frac12 h)^2 = \mathbb 1 + h + \frac14 h^2$, and h^2 is second order.
(3) h is symmetric, so $\mathrm{d}\boldsymbol{x}^{\mathsf T}(\mathbb 1 + \frac12 h)^2\mathrm{d}\boldsymbol{x}$ is the squared length of the vector $(\mathbb 1 + \frac12 h)\mathrm{d}\boldsymbol{x}$.

So $\boldsymbol{X}$ are ordinary Cartesian coordinates, the ones a ruler would give, and the mass that sits at fixed $\boldsymbol{x}_0$ is found at

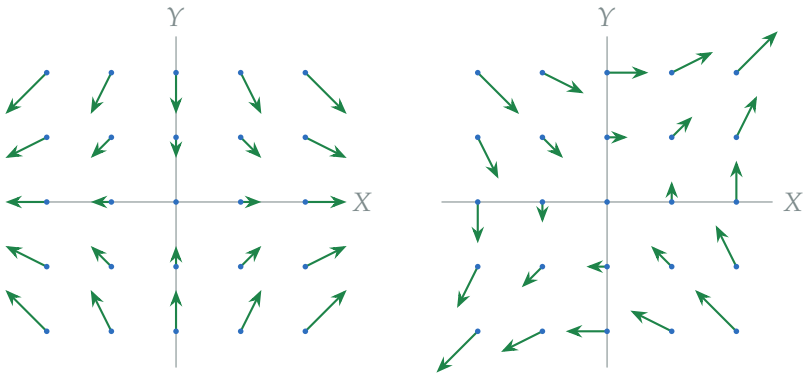
$$X_i(t) = x_{0,i} + \tfrac12\, h_{ij}(t)\, x_{0,j}, \qquad \frac{\delta L}{L} = \tfrac12\, h_{ij}\hat n_i \hat n_j, \tag{T2.12}$$

where $L = |\boldsymbol{x}_0|$ and $\hat{\boldsymbol{n}} = \boldsymbol{x}_0/L$. The second form is the fractional change of the distance, read off by projecting the displacement onto $\hat{\boldsymbol{n}}$. A mass on the x axis is pushed out by $\frac12 h_+ L$ while one on the y axis is pulled in by the same amount: the factor $\frac12$ in $\delta L/L = \frac12 h$ is the one the detector formulas of Section T2.3 start from.

The same thing as a tidal force. Differentiate (T2.12) twice in time. Since $\boldsymbol{x}_0$ is fixed,

$$\ddot{X}_i = \tfrac{1}{2}\,\ddot{h}_{ij}\,x_{0,j} = \tfrac{1}{2}\,\ddot{h}_{ij}\,X_j + O(h^2). \tag{T2.13}$$

Read as Newtonian mechanics in the ruler coordinates $\boldsymbol{X}$, each mass feels an acceleration proportional to its distance from the centre, $\frac{1}{2}\ddot{h}_{ij}X_j$. That is exactly the form of a tidal acceleration. The Moon's tidal field on the Earth is $-(\partial_i\partial_j\Phi)X_j$: it stretches the oceans along the line to the Moon and squeezes them across it, and its matrix has zero trace because $\nabla^2\Phi = 0$ in empty space. The wave plays the role of $\partial_i\partial_j\Phi$ with $\partial_i\partial_j\Phi \to -\frac{1}{2}\ddot{h}_{ij}$; this matrix is also traceless, also stretches one way while squeezing the other, but it oscillates, travels at c, and acts only across its direction of travel. In general relativity (T2.13) is the equation of geodesic deviation, $\ddot{\xi}^i = -c^2 R^i{}_{0j0}\xi^j$, with the curvature of the wave $R_{i0j0} = -\ddot{h}_{ij}/(2c^2)$ [81, Sec. 1.3].



$\ddot{h}_+ > 0$: stretch along X $\qquad$ $\ddot{h}_\times > 0$: stretch along the diagonal

Figure T2.2: The wave as a tidal field. Each dot is a free mass in the plane perpendicular to the direction of travel, and each arrow the acceleration $\frac{1}{2}\ddot{h}_{ij}X_j$ it feels at one instant. Left, a pure h_+ wave: masses are pushed out along X and pulled in along Y, with no push at the centre. Right, a pure $h_\times$ wave: the same pattern turned by 45°. Half a period later every arrow reverses.

Key idea: A gravitational wave changes distances, not positions

In the TT gauge free masses keep their coordinates, but the proper distance between two of them, separated by L along $\hat{\boldsymbol{n}}$, changes by $\delta L/L = \frac{1}{2}h_{ij}\hat{n}_i\hat{n}_j$. Equivalently, in ruler coordinates each mass feels the tidal acceleration $\frac{1}{2}\ddot{h}_{ij}X_j$: a traceless, transverse, oscillating squeeze.

Example: a ring of free masses. Put masses on a circle of radius R in the plane perpendicular to the wave, at $\boldsymbol{x}_0 = R(\cos\theta, \sin\theta)$, and let $h_+(t) = h\cos\omega t$ with $h_\times = 0$. By (T2.12),

$$X \overset{(1)}{=} R\cos\theta\,\big(1 + \tfrac{1}{2}h\cos\omega t\big), \qquad Y \overset{(1)}{=} R\sin\theta\,\big(1 - \tfrac{1}{2}h\cos\omega t\big),$$
$$\frac{X^2}{a^2} + \frac{Y^2}{b^2} \overset{(2)}{=} 1, \qquad a = R\big(1 + \tfrac{1}{2}h\cos\omega t\big), \quad b = R\big(1 - \tfrac{1}{2}h\cos\omega t\big). \tag{T2.14}$$

(1) $h_{xx} = h_+$, $h_{yy} = -h_+$, $h_{xy} = h_\times = 0$ in (T2.12).
(2) Divide X by a and Y by b; $\cos^2\theta + \sin^2\theta = 1$.

The ring becomes an ellipse with semi-axes along x and y, stretched along x and squeezed along y for half a period, then the other way round. To first order the area $\pi ab = \pi R^2(1 - \frac{1}{4}h^2\cos^2\omega t)$ does not change: the traceless squeeze moves masses around without compressing the region they enclose. With $h_\times$ instead, $X = R\cos\theta + \frac{1}{2}h_\times R\sin\theta$ and $Y = R\sin\theta + \frac{1}{2}h_\times R\cos\theta$, and the

ellipse has its axes along the diagonals. The script `12_ring_masses.py` draws both cases with the amplitude exaggerated to $h = 0.35$ (Figure T2.3), and checks that the $h_\times$ pattern is the h_+ pattern turned by 45°, as (T2.9) says.

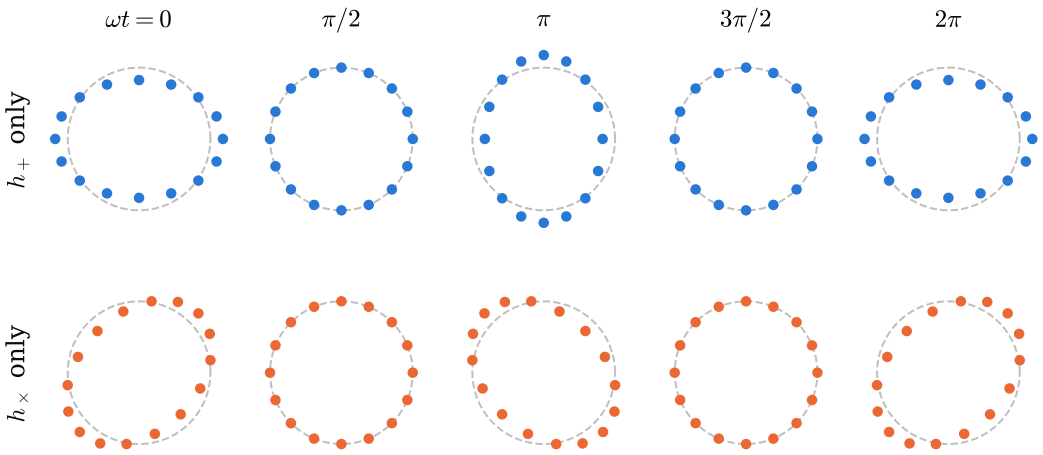


Figure T2.3: A ring of sixteen free masses (dots) as a wave travelling out of the page passes, at five phases of one period; the dashed circle is the ring without the wave. Top row, a pure plus wave: the ring is stretched along the horizontal and squeezed along the vertical, then the reverse. Bottom row, a pure cross wave: the same breathing along the diagonals. The real amplitude is some 10^{19} times smaller than drawn.

Example: how large is the effect? For a ground-based detector with arms of $L = 4\,\mathrm{km}$ and a wave of $h = 10^{-21}$, (T2.12) gives $\delta L = \frac{1}{2}hL = 2 \times 10^{-18}\,\mathrm{m}$, a few thousandths of the size of a proton. For LISA, with $L = 2.5 \times 10^9\,\mathrm{m}$ and $h = 10^{-20}$, $\delta L = \frac{1}{2} \times 10^{-20} \times 2.5 \times 10^9 \approx 10^{-11}\,\mathrm{m}$, about ten picometres. The longer the arm, the larger the displacement for the same strain, which is one reason to fly a detector in space. The other is that on the ground the seismic shaking of the Earth swamps everything below a few hertz.

Watch out: Coordinates that do not move

"In the TT gauge the masses do not move" and "the wave changes the distance between them" are both true. The first is a statement about labels, which the TT gauge attaches to the free masses themselves. The second is the physics. Every measurable effect must be computed as a proper distance or a light travel time, as in (T2.11), never read off from a change of coordinates.

T2.1.4 The energy a wave carries

A wave that can push masses around can do work, so it carries energy. Energy in general relativity is subtle, because the gravitational field has no energy density at a single point: at any one point a freely falling observer sees no field at all. What is well defined is the energy averaged over a region several wavelengths across. This is our fourth input, imported from general relativity [81, Sec. 1.4]: the energy crossing unit area per unit time, for a wave in the TT gauge, is

$$\mathcal{F} = \frac{c^3}{32\pi G}\left\langle \dot{h}^{\mathrm{TT}}_{ij}\dot{h}^{\mathrm{TT}}_{ij}\right\rangle = \frac{c^3}{16\pi G}\left\langle \dot{h}_+^2 + \dot{h}_\times^2\right\rangle, \tag{T2.15}$$

where the angle brackets average over several periods, and the second form follows from (T2.8), since $h_{ij}h_{ij} = h_+^2 + h_\times^2 + h_\times^2 + h_+^2$. Its form is that of the electromagnetic Poynting flux of a

radiation field, $\varepsilon_0 c\langle E^2\rangle$ with $E = -\dot{A}$. The flux is quadratic in the field because energy is, and it contains the time derivative because a static strain, like a static potential, carries nothing away.

Example: why a strain of 10^{-21} is a lot of energy. The constant in front of (T2.15) is enormous:

$$\frac{c^3}{16\pi G} \overset{(1)}{=} \frac{(3.0\times10^8)^3}{16\pi\times 6.67\times10^{-11}}\ \frac{\text{kg}}{\text{s}} \overset{(2)}{=} 8.0\times10^{33}\ \text{kg}\,\text{s}^{-1}.$$

(1) c in $\text{m}\,\text{s}^{-1}$ and G in $\text{m}^3\,\text{kg}^{-1}\,\text{s}^{-2}$; the units combine to $\text{kg}\,\text{s}^{-1}$, which times $\dot{h}^2$ in s^{-2} gives $\text{W}\,\text{m}^{-2}$.

(2) $2.7\times10^{25}/3.35\times10^{-9} = 8.0\times10^{33}$.

Take the first detected wave near its peak, $h_+ = h\cos 2\pi f t$ with $h \approx 10^{-21}$ and $f \approx 250\,\text{Hz}$. Then $\langle \dot{h}_+^2\rangle = \frac{1}{2}(2\pi f h)^2 = \frac{1}{2}(1.6\times10^{-18}\,\text{s}^{-1})^2 = 1.2\times10^{-36}\,\text{s}^{-2}$, and $\mathcal{F} \approx 8.0\times10^{33}\times1.2\times10^{-36} \approx 10^{-2}\,\text{W}\,\text{m}^{-2}$. For a fraction of a second, the Earth received more energy per square metre in gravitational waves than it receives in light from any star in the night sky. Yet the strain was 10^{-21}. Spacetime is stiff: the factor c^3/G means a large energy flux produces only a minute deformation, and, the other way round, a measurable deformation needs a violent source.

Where this is used. The two polarisations and the factor $\frac{1}{2}$ in $\delta L/L$ are what a detector turns into the single strain $h(t) = F_+h_+ + F_\times h_\times$ of Section T2.3. That strain is the signal part of the data stream $d(t) = h(t) + n(t)$ with which Section 10.8 begins, and the energy flux (T2.15) drives the chirp of Section T2.4, whose phase is what Section 10.9 measures.

T2.2 Why radiation starts at the quadrupole

A radio antenna radiates because electrons slosh back and forth along it: the electric dipole moment, charge times separation, oscillates. Two black holes on an orbit look at first sight like the same thing, two "charges" swinging round each other, so we might expect them to radiate like a rotating dipole, at the orbital frequency. They do not. The gravitational wave of a binary comes out at *twice* the orbital frequency, and it is weaker than a dipole estimate by a factor $(v/c)^2$. Both facts have one cause. The "charge" of gravity is mass, it has only one sign, and it is the same mass that sets the inertia of the body. As a result the two quantities that make up the monopole and the dipole of the field, the total mass and the position of the centre of mass, cannot oscillate: they are protected by conservation laws.

Starting from the field equation of Section T2.1, which part of a moving mass distribution produces a wave far away, how large is the wave, and how much power does it carry off? For a binary the answers are the waveform h_+, $h_\times$ that every template in Chapter 10 is built on, and the power P_{GW} that drives the chirp of Section T2.4.

T2.2.1 The field far from a slow source

Inputs. We keep the inputs of Section T2.1: the weak-field metric (T2.1) and the linearised field equation (T2.2), $\Box\bar{h}_{\mu\nu} = -(16\pi G/c^4)T_{\mu\nu}$. We add three about the source.

1. *Physical: small and slow.* The source has size d and internal speeds $v \ll c$. It then changes on a time scale d/v, so it emits at wavelengths $\lambda \sim c\,d/v \gg d$.
2. *Physical: energy and momentum are conserved,* $\partial_\mu T^{\mu\nu} = 0$, and the source is Newtonian, $T^{00} = \rho c^2$ with ρ the mass density. For a binary held together by its own gravity this is true only once the energy of the gravitational field is counted as well. The full theory shows that the leading-order result derived below is unchanged by that [81, Sec. 3.1].

3. *Mathematical: the retarded solution.* The solution of $\Box\psi = -4\pi s$ that contains only outgoing waves is $\psi(t,\boldsymbol{x}) = \int s(t-|\boldsymbol{x}-\boldsymbol{x}'|/c, \boldsymbol{x}')/|\boldsymbol{x}-\boldsymbol{x}'|\,\mathrm{d}^3x'$, the retarded potential familiar from electrodynamics. Applied to each component of (T2.2),

$$\bar{h}_{\mu\nu}(t,\boldsymbol{x}) = \frac{4G}{c^4}\int \frac{T_{\mu\nu}\big(t-|\boldsymbol{x}-\boldsymbol{x}'|/c,\ \boldsymbol{x}'\big)}{|\boldsymbol{x}-\boldsymbol{x}'|}\,\mathrm{d}^3x'. \tag{T2.16}$$

Far away. The detector sits at distance $r = |\boldsymbol{x}| \gg d$ in the direction $\hat{\boldsymbol{n}} = \boldsymbol{x}/r$. For a point $\boldsymbol{x}'$ of the source, $|\boldsymbol{x}-\boldsymbol{x}'| = r - \hat{\boldsymbol{n}}\cdot\boldsymbol{x}' + O(d^2/r)$. In the denominator of (T2.16) the correction is negligible and we write $1/r$. In the time argument it is not automatically negligible, because it multiplies a rapidly varying function:

$$T_{\mu\nu}\Big(t - \frac{r}{c} + \frac{\hat{\boldsymbol{n}}\cdot\boldsymbol{x}'}{c}, \boldsymbol{x}'\Big) = T_{\mu\nu}(t_r,\boldsymbol{x}') + \frac{\hat{n}_i x_i'}{c}\,\partial_t T_{\mu\nu}(t_r,\boldsymbol{x}') + \frac{\hat{n}_i\hat{n}_j x_i' x_j'}{2c^2}\,\partial_t^2 T_{\mu\nu}(t_r,\boldsymbol{x}') + \ldots, \tag{T2.17}$$

with $t_r = t - r/c$ the **retarded time**. Each term is smaller than the one before by about $(d/c)/(d/v) = v/c$: the time light takes to cross the source, divided by the time the source takes to change. This is the **multipole expansion**, and the slow-motion input says it converges quickly.

Which terms can change in time? Apply (T2.17) to the 00 component, with $T_{00} = \rho c^2$:

$$\bar{h}_{00} = \frac{4G}{c^2 r}\Big[M + \frac{\hat{n}_i}{c}\dot{M}_i + \frac{\hat{n}_i\hat{n}_j}{2c^2}\ddot{M}_{ij} + \ldots\Big]_{t_r}, \qquad M = \int\rho, \quad M_i = \int\rho\, x_i, \quad M_{ij} = \int\rho\, x_i x_j, \tag{T2.18}$$

where every integral is over the source at the retarded time. The three moments are the mass, the mass dipole and the second mass moment. Take them one at a time.

The monopole. Mass conservation, the $\nu = 0$ part of input 2, is the continuity equation $\partial_t\rho + \partial_k(\rho v_k) = 0$. Integrated over a volume enclosing the source, the divergence becomes a flux through a surface where there is no matter, so $\dot{M} = 0$. The first term is the static Newtonian potential $4GM/(c^2 r)$ and carries no wave.

The dipole. Using the continuity equation again,

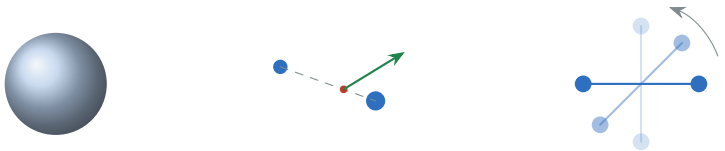
$$\dot{M}_i \overset{(1)}{=} \int x_i\,\partial_t\rho\,\mathrm{d}^3x \overset{(2)}{=} -\int x_i\,\partial_k(\rho v_k)\,\mathrm{d}^3x \overset{(3)}{=} \int \rho\, v_k\,\partial_k x_i\,\mathrm{d}^3x \overset{(4)}{=} \int \rho\, v_i\,\mathrm{d}^3x = P_i, \tag{T2.19}$$

(1) Differentiate under the integral; the integration region is fixed.
(2) Continuity equation.
(3) Integrate by parts; the surface term vanishes outside the source.
(4) $\partial_k x_i = \delta_{ik}$. The result is the total momentum P_i.

Momentum conservation, the $\nu = i$ part of input 2, gives $\dot{P}_i = 0$, so $\ddot{M}_i = 0$: $\dot{M}_i$ is constant. The second term of (T2.18) is the field of a mass whose centre moves at constant velocity, and that does not radiate either. In electromagnetism the dipole moment is $\sum q\,\boldsymbol{x}$, whose derivative $\sum q\,\boldsymbol{v}$ is not a conserved quantity, because different particles carry different charge per unit mass. In gravity every body's "charge" is its mass, the very quantity that sets its momentum, so the dipole is tied to the centre of mass and frozen.

The quadrupole. Nothing stops $\ddot{M}_{ij}$ from oscillating. The first time-dependent term, and so the first radiating one, is the second moment of the mass. A spinning symmetric top has a constant M_{ij} and stays silent; a dumbbell rotating about an axis through its middle does not, and radiates.

The wave itself. The radiation lives in the spatial components $\bar{h}_{ij}$. At leading order in v/c we keep only the first term of (T2.17), so $\bar{h}_{ij} = (4G/c^4 r)\int T_{ij}(t_r,\boldsymbol{x}')\,\mathrm{d}^3x'$. The integral of the



mass M conserved: no wave

centre of mass moves uniformly: no wave

shape M_{ij} changes twice per turn: waves

Figure T2.4: The three lowest moments of a mass distribution. The total mass (left) cannot change. The mass dipole is the total mass times the position of the centre of mass (middle, red dot), which can only move at constant velocity. The second moment (right) describes the shape; for a rotating dumbbell, drawn at three instants an eighth of a turn apart, it returns to the same value after half a turn, so it oscillates at twice the rotation frequency.

stresses T_{ij} is not obviously related to the mass distribution, but conservation turns it into one:

$$\frac{1}{c^2}\frac{\mathrm{d}^2}{\mathrm{d}t^2}\int T^{00}x_ix_j\,\mathrm{d}^3x \overset{(1)}{=} \int x_ix_j\,\partial_0\partial_0T^{00}\,\mathrm{d}^3x \overset{(2)}{=} \int x_ix_j\,\partial_k\partial_lT^{kl}\,\mathrm{d}^3x$$
$$\overset{(3)}{=} \int T^{kl}\,\partial_k\partial_l(x_ix_j)\,\mathrm{d}^3x \overset{(4)}{=} 2\int T^{ij}\,\mathrm{d}^3x. \tag{T2.20}$$

(1) $x^0 = ct$, so $\partial_0 = c^{-1}\partial_t$.
(2) From $\partial_\mu T^{\mu\nu} = 0$: $\partial_0 T^{00} = -\partial_k T^{k0}$ and $\partial_0 T^{0k} = -\partial_l T^{lk}$, so $\partial_0\partial_0 T^{00} = -\partial_k\partial_0 T^{0k} = \partial_k\partial_l T^{kl}$.
(3) Integrate by parts twice; both surface terms vanish outside the source.
(4) $\partial_l(x_ix_j) = \delta_{il}x_j + x_i\delta_{jl}$ and then ∂_k of that is $\delta_{il}\delta_{jk} + \delta_{ik}\delta_{jl}$; both terms give T^{ij} because T is symmetric.

With $T^{00} = \rho c^2$ the left side is $\ddot{M}_{ij}$, so $\int T_{ij} = \frac{1}{2}\ddot{M}_{ij}$ and $\bar{h}_{ij} = (2G/c^4r)\ddot{M}_{ij}(t_r)$. Far from the source the wave is locally a plane wave travelling along $\hat{\boldsymbol{n}}$, and its physical content is its TT part, obtained with the recipe of the working definition in Section T2.1: keep the components in the plane perpendicular to $\hat{\boldsymbol{n}}$ and remove their trace. Removing the trace means that M_{ij} may as well be replaced by its traceless part, the **mass quadrupole**

$$Q_{ij} = \int \rho\left(x_ix_j - \tfrac{1}{3}\delta_{ij}|\boldsymbol{x}|^2\right)\mathrm{d}^3x, \qquad h_{ij}^{\mathrm{TT}}(t,\boldsymbol{x}) = \frac{2G}{c^4r}\left[\ddot{Q}_{ij}(t-r/c)\right]^{\mathrm{TT}}. \tag{T2.21}$$

This is the **quadrupole formula** for the wave [81, Sec. 3.3]. In terms of the transverse unit vectors $\boldsymbol{p}$, $\boldsymbol{q}$ and the recipe of the working definition, $h_+ = \frac{G}{c^4r}\left(\ddot{Q}_{pp} - \ddot{Q}_{qq}\right)$ and $h_\times = \frac{2G}{c^4r}\ddot{Q}_{pq}$.

Key idea: The quadrupole formula

Far from a slow source, the wave is $h_{ij}^{\mathrm{TT}} = (2G/c^4r)\,[\ddot{Q}_{ij}(t-r/c)]^{\mathrm{TT}}$, with Q_{ij} the traceless second moment of the mass. Mass and momentum conservation forbid monopole and dipole radiation, so the quadrupole is the leading term, and each further multipole is suppressed by another factor v/c.

T2.2.2 The circular binary

The quadrupole of a binary. Two masses m_1, m_2 at positions $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, centre of mass at rest at the origin. With $M = m_1 + m_2$, the separation $\boldsymbol{x} = \boldsymbol{x}_1 - \boldsymbol{x}_2$ and the **reduced mass** $\mu = m_1m_2/M$, the positions are $\boldsymbol{x}_1 = (m_2/M)\boldsymbol{x}$ and $\boldsymbol{x}_2 = -(m_1/M)\boldsymbol{x}$, and

$$M_{ij} = m_1x_{1i}x_{1j} + m_2x_{2i}x_{2j} = \left(\frac{m_1m_2^2}{M^2} + \frac{m_2m_1^2}{M^2}\right)x_ix_j = \mu\,x_ix_j, \tag{T2.22}$$

so the binary radiates like a single body of mass μ on the relative orbit. For a circular orbit of radius r in the xy plane, $\boldsymbol{x} = r(\cos\omega t, \sin\omega t, 0)$, and Kepler's third law $\omega^2 = GM/r^3$ links ω and r. The components we need are

$$M_{xx} \overset{(1)}{=} \mu r^2 \cos^2\omega t \overset{(2)}{=} \tfrac{1}{2}\mu r^2(1+\cos 2\omega t), \qquad M_{yy} \overset{(2)}{=} \tfrac{1}{2}\mu r^2(1-\cos 2\omega t), \qquad M_{xy} \overset{(2)}{=} \tfrac{1}{2}\mu r^2 \sin 2\omega t,$$
$$\ddot{M}_{xx} \overset{(3)}{=} -2\mu r^2\omega^2 \cos 2\omega t = -\ddot{M}_{yy}, \qquad \ddot{M}_{xy} \overset{(3)}{=} -2\mu r^2\omega^2 \sin 2\omega t, \tag{T2.23}$$

and every component with a z is zero or constant.

(1) $x_x = r\cos\omega t$.
(2) $\cos^2\theta = \frac{1}{2}(1+\cos 2\theta)$, $\sin^2\theta = \frac{1}{2}(1-\cos 2\theta)$, $\sin\theta\cos\theta = \frac{1}{2}\sin 2\theta$.
(3) Two time derivatives of $\cos 2\omega t$ or $\sin 2\omega t$ bring down $-(2\omega)^2$; the constants drop out.

Everything oscillates at 2ω. The mass distribution of the pair, a dumbbell, looks the same after half a turn, so the wave frequency is twice the orbital one:

$$f_{\mathrm{GW}} = 2f_{\mathrm{orb}} = \frac{\omega}{\pi}. \tag{T2.24}$$

From here on f without a label means the wave frequency f_{GW}.

The two polarisations seen from an angle. Let the observer look at the binary from the direction $\hat{\boldsymbol{n}} = (0, \sin\iota, \cos\iota)$, where ι is the **inclination**, the angle between the orbital axis and the line of sight. Two transverse unit vectors are $\boldsymbol{p} = (1,0,0)$ and $\boldsymbol{q} = (0, \cos\iota, -\sin\iota)$. Then $\ddot{M}_{pp} = \ddot{M}_{xx}$, $\ddot{M}_{qq} = \cos^2\iota\,\ddot{M}_{yy}$ and $\ddot{M}_{pq} = \cos\iota\,\ddot{M}_{xy}$, because the z components do not contribute. By (T2.21) at distance D,

$$h_+ \overset{(1)}{=} \frac{G}{c^4 D}\left(\ddot{M}_{xx} - \cos^2\iota\,\ddot{M}_{yy}\right) \overset{(2)}{=} -\frac{4G\mu r^2\omega^2}{c^4 D}\,\frac{1+\cos^2\iota}{2}\,\cos 2\omega t_r,$$
$$h_\times \overset{(1)}{=} \frac{2G}{c^4 D}\cos\iota\,\ddot{M}_{xy} \overset{(2)}{=} -\frac{4G\mu r^2\omega^2}{c^4 D}\,\cos\iota\,\sin 2\omega t_r. \tag{T2.25}$$

(1) The recipe $h_+ = \frac{G}{c^4 D}(\ddot{M}_{pp} - \ddot{M}_{qq})$, $h_\times = \frac{2G}{c^4 D}\ddot{M}_{pq}$, with M in place of Q since their difference is a multiple of δ_{ij}, which the TT recipe removes.
(2) Insert (T2.23): $\ddot{M}_{xx} - \cos^2\iota\,\ddot{M}_{yy} = -2\mu r^2\omega^2(1+\cos^2\iota)\cos 2\omega t$.

Seen face on ($\iota = 0$) the two polarisations have equal size and are a quarter cycle apart: the wave is circularly polarised, as the rotating dumbbell suggests. Seen edge on ($\iota = 90°$) only h_+ is left, at half the face-on size: the dumbbell seen from the side only grows and shrinks.

The amplitude as a function of the wave frequency. Kepler's law gives $r^2\omega^2 = (GM)^{2/3}\omega^{2/3}$, and $\omega = \pi f$. So the common factor in (T2.25) is

$$A \equiv \frac{4G\mu r^2\omega^2}{c^4 D} \overset{(1)}{=} \frac{4G\mu (GM)^{2/3}(\pi f)^{2/3}}{c^4 D} \overset{(2)}{=} \frac{4(G\mathcal{M})^{5/3}(\pi f)^{2/3}}{c^4 D}, \qquad \mathcal{M} \equiv \mu^{3/5}M^{2/5} = \frac{(m_1 m_2)^{3/5}}{(m_1+m_2)^{1/5}}. \tag{T2.26}$$

(1) $r = (GM/\omega^2)^{1/3}$, so $r^2\omega^2 = (GM)^{2/3}\omega^{2/3}$.
(2) $G\mu\,G^{2/3}M^{2/3} = G^{5/3}\mu M^{2/3}$, and $\mu M^{2/3} = (\mu^{3/5}M^{2/5})^{5/3} = \mathcal{M}^{5/3}$.

The masses enter only through the **chirp mass** $\mathcal{M}$; Section T2.4 finds the same combination in the rate of the chirp. This amplitude is the one used for the Fourier-domain waveform in Section T2.5, where it becomes the constant $\mathcal{A}$ of (T2.56); it is the time-domain amplitude quoted by Eda et al. [82, eq. (21)].

Example: the size of the wave from a known source. For a GW150914-like binary, $\mathcal{M} = 28\,M_\odot$, so $G\mathcal{M}/c^3 = 28 \times 4.93\,\mu\text{s} = 1.38 \times 10^{-4}\,\text{s}$, at $D = 410\,\text{Mpc} = 1.27 \times 10^{25}\,\text{m}$ and $f = 250\,\text{Hz}$. Write (T2.26) with the masses as times, $A = 4(G\mathcal{M}/c^3)^{5/3}(\pi f)^{2/3}\, c/D$:

$$A \overset{(1)}{=} 4 \times \left(1.38 \times 10^{-4}\right)^{5/3} \times (785)^{2/3} \times \frac{3.0 \times 10^8}{1.27 \times 10^{25}} \overset{(2)}{=} 4 \times 3.7 \times 10^{-7} \times 85 \times 2.4 \times 10^{-17} \approx 3 \times 10^{-21}.$$

(1) $(G\mathcal{M})^{5/3}/c^4 = (G\mathcal{M}/c^3)^{5/3} c^5/c^4$, and $\pi \times 250 = 785\,\text{s}^{-1}$.
(2) $10^{-3.86\times 5/3} = 10^{-6.43} = 3.7 \times 10^{-7}$; $785^{2/3} = 85$; $c/D = 2.4 \times 10^{-17}\,\text{s}^{-1}$.

This is the strain of order 10^{-21} used in Section T2.1.4. For the intermediate mass-ratio benchmark of Section T2.4 ($\mathcal{M} \approx 19\,M_\odot$) at 75 Mpc and $f = 20\,\text{mHz}$ the same steps give $A \approx 1.6 \times 10^{-23}$. It is the millions of cycles, not the instantaneous size, that make such a source detectable.

T2.2.3 The power

From the flux. The power leaving the binary is the flux (T2.15) integrated over a sphere of radius D. Over one period, $\langle \sin^2 \rangle = \langle \cos^2 \rangle = \frac{1}{2}$, and each time derivative of (T2.25) brings down 2ω, so $\langle \dot{h}_+^2 + \dot{h}_\times^2 \rangle = \frac{1}{2}(2\omega A)^2 \left[\frac{1}{4}(1+\cos^2\iota)^2 + \cos^2\iota\right]$. The bracket is the angular pattern of the radiation (Figure T2.5). It does not depend on the observer's azimuth, because a circular orbit seen from a different azimuth is the same orbit seen at a different moment. With $x = \cos\iota$ and $\mathrm{d}\Omega = \mathrm{d}\phi\,\mathrm{d}x$,

$$\begin{aligned} P_{\text{GW}} &\overset{(1)}{=} \int \frac{c^3}{16\pi G} \langle \dot{h}_+^2 + \dot{h}_\times^2 \rangle D^2\,\mathrm{d}\Omega \overset{(2)}{=} \frac{c^3 D^2}{16\pi G} \cdot 2\omega^2 A^2 \cdot 2\pi \int_{-1}^{1} \left[\frac{(1+x^2)^2}{4} + x^2\right] \mathrm{d}x \\ &\overset{(3)}{=} \frac{c^3 D^2}{16\pi G} \cdot 2\omega^2 A^2 \cdot \frac{16\pi}{5} \overset{(4)}{=} \frac{2}{5} \frac{c^3}{G} \omega^2 (DA)^2 \overset{(5)}{=} \frac{32}{5} \frac{G}{c^5} \mu^2 r^4 \omega^6. \end{aligned} \tag{T2.27}$$

(1) Flux times area, summed over all directions.
(2) $\frac{1}{2}(2\omega A)^2 = 2\omega^2 A^2$; the ϕ integral gives 2π.
(3) $\int_{-1}^{1}(1+x^2)^2 \mathrm{d}x = 2 + \frac{4}{3} + \frac{2}{5} = \frac{56}{15}$ and $\int_{-1}^{1} x^2 \mathrm{d}x = \frac{2}{3}$, so the integral is $\frac{1}{4} \cdot \frac{56}{15} + \frac{2}{3} = \frac{14}{15} + \frac{10}{15} = \frac{8}{5}$, and $2\pi \cdot \frac{8}{5} = \frac{16\pi}{5}$.
(4) $\frac{2 \cdot 16\pi}{16\pi \cdot 5} = \frac{2}{5}$.
(5) $DA = 4G\mu r^2 \omega^2/c^4$ from (T2.26), so $(DA)^2 = 16G^2\mu^2 r^4 \omega^4/c^8$; then $\frac{2}{5} \cdot 16 = \frac{32}{5}$, $c^3/c^8 = c^{-5}$ and $G^2/G = G$.

The angular integral $\int [\frac{1}{4}(1+x^2)^2 + x^2]\mathrm{d}\Omega = 16\pi/5$ is worth remembering; divided by 4π it is the average $\frac{4}{5}$ that reappears in the sky-averaged waveform of Section T2.5.

From the quadrupole directly. The same integral can be done once and for all for any source. Insert (T2.21) into (T2.15). The TT recipe is a projection, applying it twice changes nothing, and for a traceless symmetric matrix it gives

$$\begin{aligned} [\ddot{Q}]_{ij}^{\text{TT}} [\ddot{Q}]_{ij}^{\text{TT}} &\overset{(1)}{=} \ddot{Q}_{ij}\ddot{Q}_{ij} - 2\,\hat{n}_i \ddot{Q}_{ik}\ddot{Q}_{kj}\hat{n}_j + \tfrac{1}{2}\left(\hat{n}_i \ddot{Q}_{ij}\hat{n}_j\right)^2, \\ \frac{1}{4\pi}\int [\ddot{Q}]_{ij}^{\text{TT}} [\ddot{Q}]_{ij}^{\text{TT}}\,\mathrm{d}\Omega &\overset{(2)}{=} \ddot{Q}_{ij}\ddot{Q}_{ij}\left(1 - \frac{2}{3} + \frac{2}{15} \cdot \frac{1}{2}\right) = \frac{2}{5}\ddot{Q}_{ij}\ddot{Q}_{ij}, \\ P_{\text{GW}} &\overset{(3)}{=} \frac{c^3 r^2}{32\pi G}\left(\frac{2G}{c^4 r}\right)^2 \cdot 4\pi \cdot \frac{2}{5}\left\langle \ddot{Q}_{ij}\ddot{Q}_{ij} \right\rangle = \frac{G}{5c^5}\left\langle \dddot{Q}_{ij}\dddot{Q}_{ij} \right\rangle. \end{aligned} \tag{T2.28}$$

(1) With the projector $P_{ij} = \delta_{ij} - \hat{n}_i\hat{n}_j$ onto the transverse plane, the recipe is $[Q]^{\text{TT}} = PQP - \frac{1}{2}P\,\text{tr}(PQ)$. Squaring and taking the trace, with $P^2 = P$ and $\text{tr}P = 2$, gives $\text{tr}(PQPQ) - \text{tr}(PQ)^2 + \frac{1}{4} \cdot 2\,\text{tr}(PQ)^2 = \text{tr}(PQPQ) - \frac{1}{2}\text{tr}(PQ)^2$. Multiplying out $P = \mathbb{1} - \hat{\boldsymbol{n}}\hat{\boldsymbol{n}}^{\mathsf{T}}$ gives $\text{tr}(PQPQ) = \text{tr}Q^2 - 2\,\hat{\boldsymbol{n}} \cdot Q^2\hat{\boldsymbol{n}} + (\hat{\boldsymbol{n}} \cdot Q\hat{\boldsymbol{n}})^2$, and $\text{tr}(PQ) = -\hat{\boldsymbol{n}} \cdot Q\hat{\boldsymbol{n}}$ because $\text{tr}\,Q = 0$.

(2) Average over directions with $\langle \hat{n}_i \hat{n}_j \rangle = \frac{1}{3}\delta_{ij}$ and $\langle \hat{n}_i \hat{n}_j \hat{n}_k \hat{n}_l \rangle = \frac{1}{15}(\delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk})$; for a traceless symmetric Q the second gives $\langle (\hat{\boldsymbol{n}} \cdot Q \hat{\boldsymbol{n}})^2 \rangle = \frac{2}{15}\mathrm{tr}Q^2$.

(3) The flux (T2.15) with $\dot{h}^{\mathrm{TT}} = (2G/c^4 r)[\ddot{Q}]^{\mathrm{TT}}$, times r^2, integrated over 4π; $\frac{c^3}{32\pi G} \cdot \frac{4G^2}{c^8} \cdot 4\pi \cdot \frac{2}{5} = \frac{G}{5c^5}$.

For the binary, the third derivatives of (T2.23) are $\dddot{Q}_{xx} = -\dddot{Q}_{yy} = 4\mu r^2 \omega^3 \sin 2\omega t$ and $\dddot{Q}_{xy} = -4\mu r^2 \omega^3 \cos 2\omega t$, so

$$\ddot{Q}_{ij}\ddot{Q}_{ij} = \ddot{Q}_{xx}^2 + \ddot{Q}_{yy}^2 + 2\ddot{Q}_{xy}^2 = (4\mu r^2\omega^3)^2\big(2\sin^2 + 2\cos^2\big)(2\omega t) = 32\,\mu^2 r^4 \omega^6, \tag{T2.29}$$

and (T2.28) gives $P_{\mathrm{GW}} = \frac{32}{5}(G/c^5)\mu^2 r^4 \omega^6$ again [81, Sec. 4.1]. Two routes, one answer: the first summed the radiation over directions by hand, the second did the same sum for a general traceless matrix.

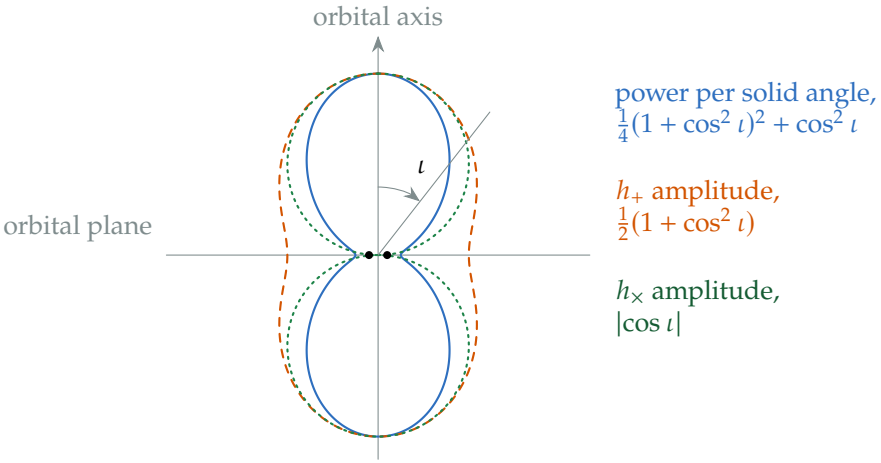


Figure T2.5: How a circular binary radiates in different directions, in units of the face-on value; the distance from the centre along a direction at angle ι from the orbital axis is the quantity named on the right. Along the axis both polarisations are full size and the power is largest. In the orbital plane $h_\times$ vanishes, h_+ is halved, and the power per solid angle drops from $\frac{1}{4}\cdot 2^2 + 1 = 2$ to $\frac{1}{4}$, an eighth of its value on the axis.

Example: how much weaker than a dipole? Suppose, against the conservation laws, that a binary could radiate like an electric dipole. Larmor's formula, $P = q^2 a^2/(6\pi\varepsilon_0 c^3)$, with $q^2/(4\pi\varepsilon_0)$ replaced by $G\mu^2$ and the acceleration $a = r\omega^2$, would give $P_{\mathrm{dip}} \sim G\mu^2 r^2 \omega^4/c^3$. The quadrupole power is $P_{\mathrm{GW}} \sim G\mu^2 r^4 \omega^6/c^5$, so

$$\frac{P_{\mathrm{GW}}}{P_{\mathrm{dip}}} \sim \frac{r^2\omega^2}{c^2} = \frac{v^2}{c^2}.$$

Quadrupole radiation is one power of $(v/c)^2$ weaker than dipole radiation would be. In theories of gravity in which bodies carry an extra "charge" that is not proportional to their mass, dipole radiation does appear, and because it is relatively stronger at low speeds, it shows up at low frequency. Section T2.6 shows that it enters the phase of the wave as the power $f^{-7/3}$.

Where this is used. The amplitude (T2.26) and its dependence on the inclination (T2.25) set the size of every template in Chapter 10: through the constant $\mathcal{A}$ of (T2.56) they fix the signal-to-noise ratio of Section 10.8.3, and through the factor $\frac{4}{5}$ they enter the sky-averaged forecasts of Section 10.11. The power (T2.27) drives the chirp of Section T2.4, whose phase is what Section 10.9 measures.

T2.3 How a detector turns a wave into one number

A wave arrives with two polarisations, $h_+(t)$ and $h_\times(t)$, defined with respect to axes in the plane perpendicular to its direction of travel (Section T2.1). A detector records one time series. LIGO records the difference between the lengths of its two arms, read from the interference of laser light that has run up and down each of them; LISA does the same with laser links between spacecraft millions of kilometres apart. Every likelihood in Chapter 10 compares that one recorded number with a template, so we need the exact rule that maps $(h_+, h_\times)$ onto it.

How does a change in the light travel time along one arm follow from h_{ij}? How do two arms combine into one strain? How strongly does the result depend on where the source sits on the sky and how its orbit is turned? And what changes when the arm is so long that the wave changes while the light is still in flight? The answers produce three numbers that appear without explanation in most papers: the $\frac{4}{5}$ of the inclination average, the $\frac{1}{5}$ of the sky average, and LISA's $\frac{3}{10}$, together with the transfer frequency $f_* = c/(2\pi L)$.

T2.3.1 One arm: the light travel time

The simplest case first. Put one free mass at the origin and one at $x = L$, the arm along x, and let a wave travel along z with a wavelength much longer than L, so that $h_{xx} = h_+$ is the same along the whole arm while the light is in flight. Light moves on $\mathrm{d}s^2 = 0$. Along the arm only $\mathrm{d}t$ and $\mathrm{d}x$ change, so with the TT metric

$$0 \overset{(1)}{=} -c^2\mathrm{d}t^2 + (1 + h_{xx})\,\mathrm{d}x^2 \;\Rightarrow\; c\,\mathrm{d}t \overset{(2)}{=} \sqrt{1 + h_{xx}}\,\mathrm{d}x \overset{(3)}{=} \Big(1 + \tfrac{1}{2}h_{xx}\Big)\mathrm{d}x,$$

$$T_{\text{round}} \overset{(4)}{=} \frac{2L}{c}\Big(1 + \tfrac{1}{2}h_{xx}\Big), \qquad \frac{\delta T_{\text{round}}}{2L/c} = \tfrac{1}{2}h_{xx}. \tag{T2.30}$$

(1) The TT line element (T2.8) along the x axis, with $\mathrm{d}y = \mathrm{d}z = 0$.
(2) Solve for $c\,\mathrm{d}t$.
(3) $\sqrt{1+\epsilon} = 1 + \frac{1}{2}\epsilon$ to first order.
(4) Integrate from 0 to L and back. The end masses keep their TT coordinates (Section T2.1.3), so the coordinate length is L both ways.

The light takes longer exactly as if the arm had grown by $\delta L/L = \frac{1}{2}h_{xx}$, which is the proper-distance result (T2.12) again: a light clock and a ruler agree. For an arm along any unit vector $\hat{\boldsymbol{a}}$, the same steps give $\delta L/L = \frac{1}{2}h_{ij}\hat{a}_i\hat{a}_j$, written $\frac{1}{2}h_{aa}$ for short.

Two arms. A Michelson interferometer sends light along two arms $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{v}}$ of equal length and compares the return times. Common changes cancel; the difference survives. The strain the instrument reports is the length difference divided by the arm length [83, eq. (5)], [84, eq. (11)]:

$$h(t) \equiv \frac{\delta L_u - \delta L_v}{L} = \tfrac{1}{2}\big(h_{uu} - h_{vv}\big) = D_{ij}h_{ij}, \qquad D_{ij} \equiv \tfrac{1}{2}\big(\hat{u}_i\hat{u}_j - \hat{v}_i\hat{v}_j\big). \tag{T2.31}$$

The matrix D, the **detector tensor**, is all the geometry of the instrument that matters while the arms are short compared with the wavelength. For a wave of plus polarisation coming straight down onto a right-angle detector with arms along x and y, $h_{uu} = h_+$ and $h_{vv} = -h_+$, so $h = h_+$: one arm grows while the other shrinks, and the two effects add in the difference. That is why the interferometer has two arms at right angles.

T2.3.2 Antenna patterns: where on the sky, and how turned

The general rule. Let the wave come from the direction $\hat{\boldsymbol{n}}$ (so it travels along $-\hat{\boldsymbol{n}}$), and let $\boldsymbol{p}$, $\boldsymbol{q}$ be the axes in the transverse plane with respect to which h_+ and $h_\times$ are defined. By (T2.8) the

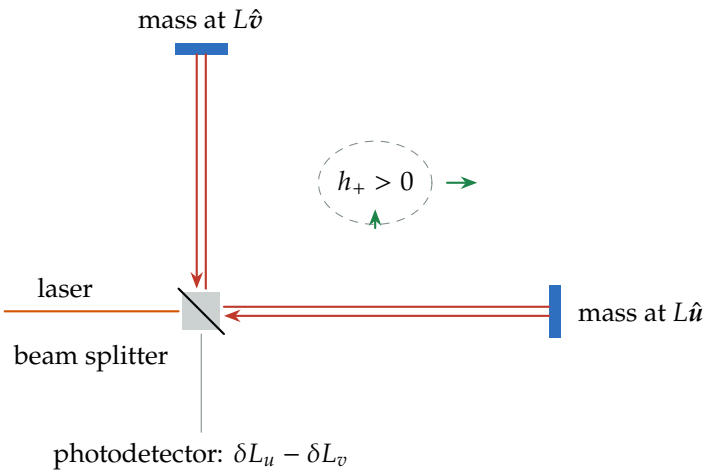


Figure T2.6: A Michelson interferometer. A beam splitter sends laser light along two perpendicular arms towards freely suspended test masses (blue). When the reflected beams return, their relative phase measures the difference in round-trip travel times. A plus wave arriving from above, at the instant drawn, lengthens the horizontal arm and shortens the vertical one (small ellipse), so the two changes add up in the difference.

wave is $h_{ij} = h_+ e^+_{ij} + h_\times e^\times_{ij}$ with $e^+ = \boldsymbol{p}\boldsymbol{p}^\mathsf{T} - \boldsymbol{q}\boldsymbol{q}^\mathsf{T}$ and $e^\times = \boldsymbol{p}\boldsymbol{q}^\mathsf{T} + \boldsymbol{q}\boldsymbol{p}^\mathsf{T}$. Inserting it in (T2.31),

$$h(t) = F_+\, h_+(t) + F_\times\, h_\times(t), \qquad F_A = D_{ij}\, e^A_{ij} \quad (A = +, \times), \tag{T2.32}$$

the form used for every detector [83, eq. (24)].

Working definition T2.2 (antenna pattern functions)

F_+ and $F_\times$ are the weights with which a detector adds the two polarisations of a wave into its one output, $h = F_+ h_+ + F_\times h_\times$ [numbers between −1 and 1, fixed by the arm directions, the source direction and the orientation of the polarisation axes]. *Operationally:* $F_A = D_{ij} e^A_{ij}$ with the detector tensor $D = \frac{1}{2}(\hat{\boldsymbol{u}}\hat{\boldsymbol{u}}^\mathsf{T} - \hat{\boldsymbol{v}}\hat{\boldsymbol{v}}^\mathsf{T})$.

A right-angle detector. Take arms along x and y and a source at polar angle θ and azimuth ϕ in the detector's frame. Choose first the natural transverse axes, $\boldsymbol{p}_0 = \hat{\boldsymbol{\theta}} = (\cos\theta\cos\phi, \cos\theta\sin\phi, -\sin\theta)$ and $\boldsymbol{q}_0 = \hat{\boldsymbol{\phi}} = (-\sin\phi, \cos\phi, 0)$. With $D = \frac{1}{2}\mathrm{diag}(1, -1, 0)$,

$$\begin{aligned} F^0_+ &\overset{(1)}{=} \tfrac{1}{2}\big(p_x^2 - q_x^2 - p_y^2 + q_y^2\big) \overset{(2)}{=} \tfrac{1}{2}\big(\cos^2\theta\cos^2\phi - \sin^2\phi - \cos^2\theta\sin^2\phi + \cos^2\phi\big) \\ &\overset{(3)}{=} \tfrac{1}{2}(1 + \cos^2\theta)\cos 2\phi, \\ F^0_\times &\overset{(1)}{=} p_x q_x - p_y q_y \overset{(2)}{=} -\cos\theta\cos\phi\sin\phi - \cos\theta\sin\phi\cos\phi \overset{(3)}{=} -\cos\theta\sin 2\phi. \end{aligned} \tag{T2.33}$$

(1) $D_{ij} e^A_{ij} = \frac{1}{2}(e^A_{xx} - e^A_{yy})$, with $e^+_{xx} = p_x^2 - q_x^2$ and $e^\times_{xx} = 2p_x q_x$.
(2) Insert the components of $\boldsymbol{p}_0$ and $\boldsymbol{q}_0$.
(3) $\cos^2\phi - \sin^2\phi = \cos 2\phi$ and $2\sin\phi\cos\phi = \sin 2\phi$.

If the wave's own axes are turned by an angle ψ, the **polarisation angle**, from $(\boldsymbol{p}_0, \boldsymbol{q}_0)$, the rule (T2.9) for turning axes rewrites the wave in the $(\boldsymbol{p}_0, \boldsymbol{q}_0)$ frame as $h_+\cos 2\psi - h_\times \sin 2\psi$ and $h_+ \sin 2\psi + h_\times \cos 2\psi$. Collecting the coefficients of h_+ and $h_\times$,

$$\begin{aligned} F_+ &= \tfrac{1}{2}(1 + \cos^2\theta)\cos 2\phi\cos 2\psi - \cos\theta\sin 2\phi\sin 2\psi, \\ F_\times &= -\tfrac{1}{2}(1 + \cos^2\theta)\cos 2\phi\sin 2\psi - \cos\theta\sin 2\phi\cos 2\psi, \end{aligned} \tag{T2.34}$$

which is [85, eq. (6)] and [84, eq. (15a)]; their $F_\times$ has the opposite overall sign, a matter of which way ψ is counted, which no average or signal-to-noise ratio can see. The detector is deaf in four directions in its own plane, $\theta = 90°$ and $\phi = 45°, 135°, \ldots$, along the bisectors of the arms, where both arms are stretched equally. It is most sensitive straight overhead. Figure T2.7 maps $F_+^2 + F_\times^2$, which does not depend on ψ.

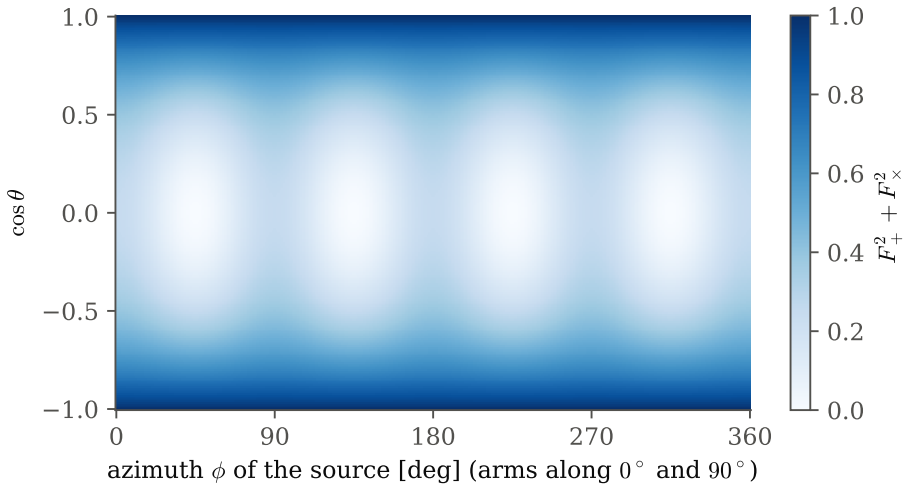


Figure T2.7: The sensitivity of a right-angle interferometer over the sky, $F_+^2 + F_\times^2$, as a function of the source's azimuth and of the cosine of its polar angle. It is largest above and below the detector ($\cos\theta = \pm 1$), and drops to zero at four points in the plane of the arms ($\cos\theta = 0$, $\phi = 45°$, $135°$, $225°$, $315°$), where the wave stretches both arms equally.

Example: a source overhead and a source in the plane of the arms. For a source straight overhead, $\theta = 0$, (T2.34) gives

$$F_+ \overset{(1)}{=} \cos 2\phi \cos 2\psi - \sin 2\phi \sin 2\psi \overset{(2)}{=} \cos(2\phi + 2\psi),$$
$$F_\times \overset{(1)}{=} -\cos 2\phi \sin 2\psi - \sin 2\phi \cos 2\psi \overset{(2)}{=} -\sin(2\phi + 2\psi),$$

(1) $\cos\theta = 1$, so $\frac{1}{2}(1 + \cos^2\theta) = 1$.
(2) The addition rules for $\cos(a + b)$ and $\sin(a + b)$.

so $F_+^2 + F_\times^2 = 1$ whatever the orientation: overhead, the detector catches the full wave, and turning the polarisation axes only moves the signal from one polarisation to the other. For a source in the plane of the arms, in the direction of the x arm ($\theta = 90°$, $\phi = 0$), we get $F_+ = \frac{1}{2}\cos 2\psi$ and $F_\times = -\frac{1}{2}\sin 2\psi$, so $F_+^2 + F_\times^2 = \frac{1}{4}$: the wave travels along the x arm, which it cannot stretch because the wave is transverse, and only the y arm responds, with its weight $\frac{1}{2}$ in the difference (T2.31).

The sky and polarisation average. For forecasts the source's position and orientation are not known, so the response is averaged over all of them, with the source direction uniform on the sphere and ψ uniform. Take $\langle F_+^2 \rangle$ in three steps, one angle at a time:

$$\langle F_+^2 \rangle_\psi \overset{(1)}{=} \tfrac{1}{2}\Big[\tfrac{1}{4}(1 + \cos^2\theta)^2 \cos^2 2\phi + \cos^2\theta \sin^2 2\phi\Big],$$
$$\langle F_+^2 \rangle_{\psi,\phi} \overset{(2)}{=} \tfrac{1}{4}\Big[\tfrac{1}{4}(1 + \cos^2\theta)^2 + \cos^2\theta\Big],$$
$$\langle F_+^2 \rangle \overset{(3)}{=} \tfrac{1}{4}\Big[\tfrac{1}{4} \cdot \tfrac{28}{15} + \tfrac{1}{3}\Big] = \tfrac{1}{4} \cdot \tfrac{12}{15} = \tfrac{1}{5}. \qquad \text{(T2.35)}$$

(1) Square (T2.34); over a full turn of ψ, $\langle\cos^2 2\psi\rangle = \langle\sin^2 2\psi\rangle = \frac{1}{2}$ and $\langle\cos 2\psi \sin 2\psi\rangle = 0$, so the cross term drops.
(2) The same averages over ϕ.
(3) A direction uniform on the sphere has $x = \cos\theta$ uniform on $[-1, 1]$, so $\langle(1 + x^2)^2\rangle = \frac{1}{2}\int_{-1}^{1}(1 + 2x^2 + x^4)\mathrm{d}x = 1 + \frac{1}{3} + \frac{1}{5} = \frac{28}{15}$ and $\langle x^2\rangle = \frac{1}{3}$.

The same steps give $\langle F_\times^2\rangle = \frac{1}{5}$, and $\langle F_+ F_\times\rangle = 0$. This is the $\frac{1}{5}$ of [85, eq. (7)]. The bracket in the second line is the same function $\frac{1}{4}(1 + \cos^2)^2 + \cos^2$ that described the radiation pattern of a binary in Section T2.2: a detector receives from each direction in the same pattern as a rotating quadrupole emits into it, because both are the TT part of a quadrupolar tensor.

The inclination average. The source has its own orientation. By (T2.25) a circular binary at inclination ι has $h_+ = A\frac{1}{2}(1 + \cos^2\iota)\cos\Phi$ and $h_\times = A\cos\iota\sin\Phi$ (up to an overall sign), where Φ is the wave's phase. Averaged over ψ, F_+^2 and $F_\times^2$ have the same mean and no correlation, so the mean squared signal is $\langle F_+^2\rangle A^2[\frac{1}{4}(1 + \cos^2\iota)^2 + \cos^2\iota]$, and averaging over orientations of the orbit, $x = \cos\iota$ uniform,

$$\left\langle \tfrac{1}{4}(1 + x^2)^2 + x^2 \right\rangle_x = \tfrac{1}{4}\cdot\tfrac{28}{15} + \tfrac{1}{3} = \tfrac{7}{15} + \tfrac{5}{15} = \tfrac{4}{5}. \tag{T2.36}$$

This is the $\frac{4}{5}$ in the sky-averaged waveform $\tilde{h} = \sqrt{4/5}\,\mathcal{A}f^{-7/6}e^{-i\Psi}$ of [86, eq. (44)] and [85, eq. (16)], and divided into $16\pi/5$ over 4π it is the same angular integral that gave the binary's total power. For a ground-based detector the two averages combine to $\frac{4}{5}\cdot\frac{1}{5} = \frac{4}{25}$ in power, $\frac{2}{5}$ in amplitude, the factor quoted in the LIGO literature.

T2.3.3 LISA at low frequency: 60° arms and two channels

LISA's three spacecraft form an equilateral triangle (Figure T2.8), so two arms meeting at a spacecraft make 60°, not 90°. *How much does that cost?* Let $\hat{\boldsymbol{e}}_1$ be the unit vector along the bisector of the two arms and $\hat{\boldsymbol{e}}_2$ the perpendicular one in their plane, so that with $\alpha = 60°$ the opening angle,

$$\hat{\boldsymbol{u}} = \cos\tfrac{\alpha}{2}\,\hat{\boldsymbol{e}}_1 + \sin\tfrac{\alpha}{2}\,\hat{\boldsymbol{e}}_2, \qquad \hat{\boldsymbol{v}} = \cos\tfrac{\alpha}{2}\,\hat{\boldsymbol{e}}_1 - \sin\tfrac{\alpha}{2}\,\hat{\boldsymbol{e}}_2,$$
$$2D = \hat{\boldsymbol{u}}\hat{\boldsymbol{u}}^{\mathsf{T}} - \hat{\boldsymbol{v}}\hat{\boldsymbol{v}}^{\mathsf{T}} \overset{(1)}{=} 2\cos\tfrac{\alpha}{2}\sin\tfrac{\alpha}{2}\left(\hat{\boldsymbol{e}}_1\hat{\boldsymbol{e}}_2^{\mathsf{T}} + \hat{\boldsymbol{e}}_2\hat{\boldsymbol{e}}_1^{\mathsf{T}}\right) \overset{(2)}{=} \sin\alpha\left(\hat{\boldsymbol{e}}_1\hat{\boldsymbol{e}}_2^{\mathsf{T}} + \hat{\boldsymbol{e}}_2\hat{\boldsymbol{e}}_1^{\mathsf{T}}\right). \tag{T2.37}$$

(1) Multiply out: the $\hat{\boldsymbol{e}}_1\hat{\boldsymbol{e}}_1^{\mathsf{T}}$ and $\hat{\boldsymbol{e}}_2\hat{\boldsymbol{e}}_2^{\mathsf{T}}$ terms are the same in both and cancel; the mixed terms add.
(2) $2\sin\frac{\alpha}{2}\cos\frac{\alpha}{2} = \sin\alpha$.

For $\alpha = 90°$ the bracket is the detector tensor of a right-angle interferometer whose arms lie along $(\hat{\boldsymbol{e}}_1 \pm \hat{\boldsymbol{e}}_2)/\sqrt{2}$. So a 60° interferometer is a right-angle interferometer, turned by 45° relative to its bisector, with every response multiplied by $\sin 60° = \sqrt{3}/2$: the factor in [84, eq. (14)]. Turning does not change a sky average, so each 60° channel has $\langle F_+^2\rangle = \frac{3}{4}\cdot\frac{1}{5} = \frac{3}{20}$. The three arms of the triangle provide two such independent channels at low frequency, the second equivalent to the first turned by 45° [84, eqs. (12), (15b)], and the sensitivity-curve literature sums them: $R = 2\times\frac{3}{20} = \frac{3}{10}$ [85, eqs. (8)–(9)]. The script `13_lisa_response.py` checks all of these averages by quadrature over the sky and ψ and finds 0.2 and 0.2 for a right-angle detector, 0.15 for one 60° channel, 0.8 for the inclination average and $3.2\,\pi$ for the angular integral of the binary's power.

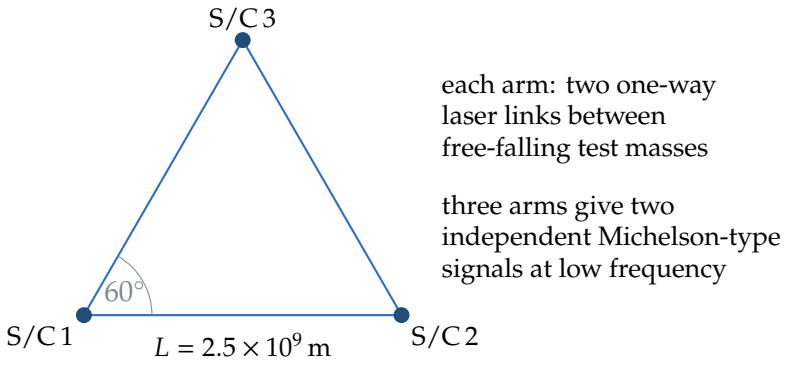


Figure T2.8: The LISA constellation. Three spacecraft form an equilateral triangle; each side carries laser links in both directions. Two arms meeting at a spacecraft act as a Michelson interferometer whose arms make 60° rather than 90°. At low frequency the three arms carry two independent combinations, written as detectors I and II in [84, eqs. (11)–(12)].

T2.3.4 Long arms: the transfer frequency

LISA's arms are $L = 2.5\times10^9$ m long, and light needs $L/c = 8.3$ s to cross one. A wave of 10 mHz has a period of 100 s, so it changes noticeably while the light is in flight, and (T2.30), which froze h during the trip, must be redone.

The simplest case: a wave arriving perpendicular to the arm. Then $h_{xx} = h(t)$ is the same at every point of the arm but changes in time. Light reflected at the far end at time t_m left at $t_m - L/c$ and returns at $t_m + L/c$. Following the steps of (T2.30) with h evaluated at the moment the light passes, for $h(t) = h_0 \cos 2\pi f t$,

$$\delta T_{\text{round}} \overset{(1)}{=} \frac{1}{2}\int_{t_m-L/c}^{t_m+L/c} h_0 \cos(2\pi f t')\,\mathrm{d}t' \overset{(2)}{=} \frac{h_0}{4\pi f}\left[\sin 2\pi f\left(t_m+\frac{L}{c}\right) - \sin 2\pi f\left(t_m-\frac{L}{c}\right)\right]$$
$$\overset{(3)}{=} \frac{h_0}{2\pi f}\cos(2\pi f t_m)\,\sin\frac{2\pi f L}{c} \overset{(4)}{=} \frac{L}{c}\,h(t_m)\,\mathrm{sinc}\left(\frac{f}{f_*}\right), \qquad f_* \equiv \frac{c}{2\pi L}, \tag{T2.38}$$

(1) Each element $\mathrm{d}x = c\,\mathrm{d}t'$ of the path adds $\frac{1}{2}h(t')\,\mathrm{d}t'$ to the travel time, by step (3) of (T2.30).
(2) The integral of cos is $\sin/(2\pi f)$.
(3) $\sin a - \sin b = 2\cos\frac{a+b}{2}\sin\frac{a-b}{2}$, with $\frac{a+b}{2} = 2\pi f t_m$ and $\frac{a-b}{2} = 2\pi f L/c$.
(4) Multiply and divide by $2\pi f L/c = f/f_*$, with $\mathrm{sinc}\,z \equiv \sin z/z$.

For $f \ll f_*$, sinc $\to 1$ and we are back at $\delta T = (L/c)h$, i.e. $\delta L/L = \frac{1}{2}h$. The **transfer frequency** f_* is where the wave's phase advances by one radian while the light crosses one arm:

$$f_* = \frac{c}{2\pi L} = \frac{3.0\times10^8\ \mathrm{m\,s^{-1}}}{2\pi\times 2.5\times10^9\ \mathrm{m}} = 19.1\ \mathrm{mHz}. \tag{T2.39}$$

At $f = \pi f_* = c/(2L) = 59.96$ mHz the wave completes exactly one cycle during the round trip, the stretching on the way out is undone by the squeezing on the way back, and the arm sees nothing. Above f_*, $|\mathrm{sinc}(f/f_*)| \le f_*/f$, so the response in power falls at least as $(f_*/f)^2$.

Example: how much is lost at 10 and at 100 millihertz. At $f = 10$ mHz, $f/f_* = 0.52$ and sinc(0.52) = sin(0.52)/0.52 = 0.497/0.52 = 0.96, so the response in power is $0.96^2 = 0.91$ of its low-frequency value: hardly any loss. At $f = 100$ mHz, $f/f_* = 5.24$ and sinc(5.24) = sin(5.24)/5.24 = −0.866/5.24 = −0.17, a power response of 0.027, below the bound $(f_*/f)^2 = 0.036$. A tenfold rise in frequency has cost a factor thirty in power, which the noise curve of Section T2.7 will show as a rising high-frequency wall.

Any direction. If the wave travels along $\hat{\boldsymbol{k}}$ and the arm points along $\hat{\boldsymbol{a}}$, the light moving outwards meets the wave at the position $\boldsymbol{x} = \hat{\boldsymbol{a}}\, c(t' - t_0)$, where the wave's phase is $2\pi f(t' - \hat{\boldsymbol{k}} \cdot \boldsymbol{x}/c) = 2\pi f[(1 - \hat{\boldsymbol{k}} \cdot \hat{\boldsymbol{a}})t' + \text{const}]$. The light rides with the wave if $\hat{\boldsymbol{k}} \cdot \hat{\boldsymbol{a}} = 1$ and meets it head on if -1, so the phase runs at the Doppler-shifted rate $2\pi f(1 - \hat{\boldsymbol{k}} \cdot \hat{\boldsymbol{a}})$ along the outward leg and $2\pi f(1 + \hat{\boldsymbol{k}} \cdot \hat{\boldsymbol{a}})$ on the way back. The integral over one leg of duration L/c is therefore

$$\int_{t_0}^{t_0 + L/c} e^{2\pi i f (1 \mp \hat{\boldsymbol{k}} \cdot \hat{\boldsymbol{a}}) t'} \, \mathrm{d}t' = \frac{L}{c}\, e^{2\pi i f (1 \mp \hat{\boldsymbol{k}} \cdot \hat{\boldsymbol{a}})(t_0 + L/2c)} \operatorname{sinc}\!\left[\frac{f}{2 f_*}\left(1 \mp \hat{\boldsymbol{k}} \cdot \hat{\boldsymbol{a}}\right)\right], \tag{T2.40}$$

the same calculation as (T2.38) for half the path and a shifted frequency. Putting both legs of both arms into (T2.31) gives the frequency-dependent $F_+(f)$ and $F_\times(f)$ of a LISA channel. The script averages $|F_+(f)|^2$ over the sky and polarisation for one 60° Michelson and doubles it for the two channels:

Listing T2.1: The light-travel integrals of one arm, and the sky-averaged response (`code/chT2/13_lisa_response.py`)

```
95
96  def leg(fr, kdota, start, sign):
97      """Integral of exp(2 pi i f [(1 - sign k.a) t' + const]) over one leg of length T = 1.
98
99      Outgoing leg (sign=+1): t' in [-2, -1], const = -2 k.a ; returning leg (sign=-1): t' in
        [-1, 0].
100     Closed form: int_A^{A+1} e^{i kappa t} dt = e^{i kappa (A + 1/2)} sinc(kappa/2).
101     """
102     alpha = 1 - sign * kdota
103     kappa = 2 * np.pi * fr * alpha
104     const = -2 * kdota if sign > 0 else 0.0
105     return np.exp(1j * kappa * (start + 0.5) + 2j * np.pi * fr * const) * sinc(kappa / 2)
106
107
108 def arm_transfer(fr, arm):
109     """Round-trip time perturbation of one arm divided by its low-frequency value."""
110     kda = k_hat @ arm
111     return (leg(fr, kda, -2.0, +1) + leg(fr, kda, -1.0, -1)) / 2.0
112
113
114 x_f = np.logspace(-2, 1.6, 120)                 # f / f*
115 R_X = np.empty_like(x_f)
116 hpu, hpv = proj(p, p, u) - proj(q, q, u), proj(p, p, v60) - proj(q, q, v60)   # e+ along
        each arm
117 for i, xf in enumerate(x_f):
118     fr = xf / (2 * np.pi)
119     Fp_f = 0.5 * (hpu * arm_transfer(fr, u) - hpv * arm_transfer(fr, v60))
120     R_X[i] = avg(np.abs(Fp_f) ** 2)
121 R2 = 2 * R_X                                    # two low-frequency channels
```

The function `leg` is (T2.40) with the light time L/c set to one, so that $2\pi f = f/f_*$. `arm_transfer` adds the outward and return legs and divides by the round-trip time, so it tends to one at low frequency. The loop forms $F_+(f)$ for each sky direction and polarisation angle and averages its squared modulus with the quadrature weights `W`.

Figure T2.9 shows the result. At low frequency it is 0.3, the $\frac{3}{10}$ found above. At f_* it has dropped to 0.708 of that (the fit $R(f) = \frac{3}{10}/[1 + 0.6(f/f_*)^2]$ of [85, eq. (9)] gives 0.625), and well above f_* it falls as $f^{-2.009}$, the f^{-2} of the single-arm argument. Between about $2f_*$ and $10f_*$ our single-Michelson response wiggles around the fit by up to 66.42%: these are the zeros of the individual arm responses, partly filled in by the sky average. The fit smooths them over, and the full response used in research analyses adds the third, high-frequency channel of the triangle [85].

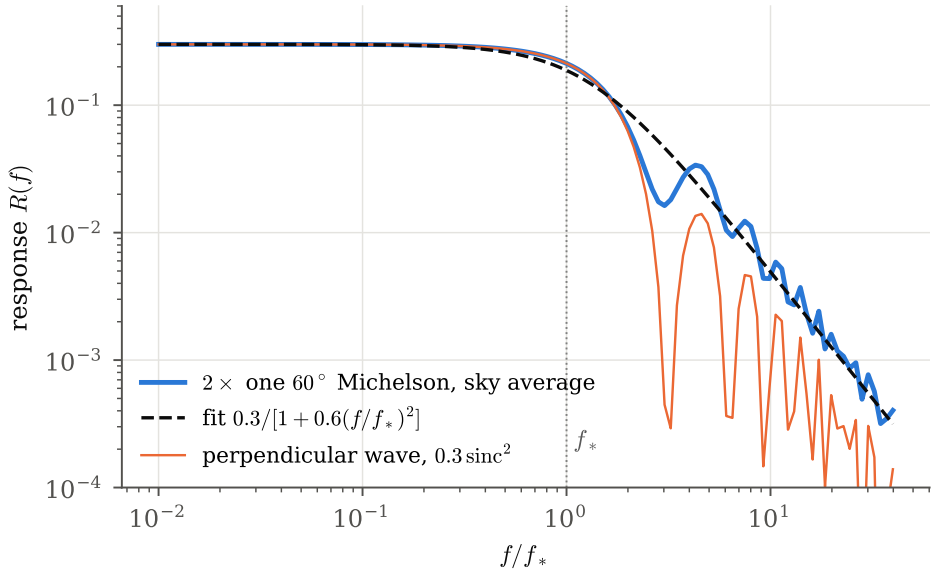


Figure T2.9: The sky- and polarisation-averaged response of LISA against frequency in units of $f_* = c/(2\pi L)$. Blue: computed from the light-travel integrals (T2.40) for one 60° Michelson channel and doubled for two channels. Dashed: the fit of Robson et al. Orange: a wave arriving perpendicular to the arms, whose response $\mathrm{sinc}^2(f/f_*)$ vanishes whenever a whole number of cycles fits into the round trip. Averaging over directions fills in those zeros, but the f^{-2} fall above f_* remains.

Key idea: The numbers that turn a source into an average source

A detector reports $h = F_+h_+ + F_\times h_\times$. Averaged over the sky and the polarisation angle, a right-angle interferometer has $\langle F_+^2\rangle = \langle F_\times^2\rangle = \frac{1}{5}$; a 60° channel has $\frac{3}{4}$ of that, $\frac{3}{20}$, and LISA's two channels together $\frac{3}{10}$. Averaging a circular binary over its inclination multiplies the signal power by $\frac{4}{5}$. Above the transfer frequency $f_* = c/(2\pi L)$ the wave changes while the light is in flight and the response falls as f^{-2}.

What the sky average leaves out. LISA orbits the Sun, so a source's signal is Doppler shifted back and forth over a year, and the triangle rotates, so the antenna patterns (T2.34) change slowly with time [84, eqs. (16)–(17)]. The phase is modulated by $2\pi f R \sin\theta_S \cos(2\pi t/\mathrm{yr} - \phi_S)$, with $R = 1\,\mathrm{AU}/c = 499\,\mathrm{s}$ and (θ_S, ϕ_S) the source's sky position [84, eq. (33)]. At $f = 10\,\mathrm{mHz}$ the amplitude of this modulation is $2\pi \times 0.01 \times 499 \approx 31$ rad, so it is easily measured, and it is how LISA locates sources on the sky. The sky-averaged response drops that information and replaces the real, moving detector by an average one. For forecasts of an environment, which acts on the phase evolution and not on the sky position, the averaged response is adequate, and the signal-to-noise ratios it gives are good to a factor of about two [85].

Where this is used. The response rule (T2.32) is what turns a template into the predicted data stream in the detection problem of Section 10.8. The averages $\frac{4}{5}$ and $\frac{3}{10}$ are the ones built into the sky-averaged waveform and the sensitivity curve of Section T2.7, which Section 10.11.1 checks against its source; f_* and the high-frequency fall of the response set the shape of that curve above about 20 mHz.

T2.4 Why a compact binary chirps

Take two black holes, masses m_1 and m_2, on a circular orbit of radius r. Section T2.2 showed that the pair radiates gravitational waves at twice its orbital frequency, with the power $P_{\mathrm{GW}} = \frac{32}{5}(G/c^5)\mu^2 r^4 \omega^6$ of (T2.27), and the energy flux of Section T2.1.4 says that this power is real energy leaving the system. The pair therefore has less and less orbital energy, so it must sit deeper in its own potential well, on a smaller orbit. A smaller orbit is a faster orbit, and since $P_{\mathrm{GW}} \propto r^4\omega^6 \propto r^{-5}$, a faster orbit radiates far more strongly. The process runs away: the orbit shrinks ever faster, and the frequency of the wave rises ever faster until the holes merge. Played as sound, the signal is a rising whistle, a **chirp**.

Every statistics calculation with such a signal starts from one question: *given the two masses and the frequency f of the wave we see today, how fast is f rising, and how long is left before the merger?* The answer is a single formula for $\dot{f}$. The time to merger, the number of cycles, the phase of the wave and the end of the inspiral all follow from it.

T2.4.1 The inputs

As with the Poisson process of Chapter 2, we first separate what goes in from what comes out. Four things go in, and only the first two are new.

1. *Physical: Newtonian circular orbits.* The two bodies obey Newton's gravity and move on a circle. The two-body problem reduces, as usual, to one body of reduced mass $\mu = m_1 m_2/M$ circling a fixed mass $M = m_1 + m_2$, with Kepler's third law $\omega^2 = GM/r^3$ for the orbital angular frequency ω.
2. *Physical: slow evolution.* One orbit takes much less time than the orbit takes to shrink noticeably. So at each moment the orbit is a circle of the current radius, and the orbital energy simply drains into the waves, $\mathrm{d}E/\mathrm{d}t = -P_{\mathrm{GW}}$. We check this with numbers at the end of the section.
3. *Derived earlier: the radiated power $P_{\mathrm{GW}} = \frac{32}{5}(G/c^5)\mu^2 r^4\omega^6$ of (T2.27), which came from the linearised field equation and the energy flux in Section T2.2, and the wave frequency $f = \omega/\pi$ of (T2.24).*
4. *Mathematical: $r(t)$ is smooth, so we may differentiate it.*

That only one combination of the masses matters is *not* an input; it comes out of the derivation.

T2.4.2 From the radiated power to the chirp

Energy and power as functions of the wave frequency. The frequency is what the detector measures, so we trade r for f using Kepler, $r = (GM/\omega^2)^{1/3}$ with $\omega = \pi f$:

$$E \overset{(1)}{=} -\frac{GM\mu}{2r} \overset{(2)}{=} -\frac{\mu}{2}(\pi GMf)^{2/3}, \tag{T2.41}$$

$$P_{\mathrm{GW}} \overset{(3)}{=} \frac{32}{5}\frac{G}{c^5}\mu^2(GM)^{4/3}(\pi f)^{10/3} \overset{(4)}{=} \frac{32}{5}\frac{c^5}{G}\left(\frac{\pi G\mathcal{M}f}{c^3}\right)^{10/3}, \qquad \mathcal{M} \equiv \mu^{3/5}M^{2/5} = \frac{(m_1m_2)^{3/5}}{(m_1+m_2)^{1/5}}. \tag{T2.42}$$

(1) Kinetic plus potential energy of a circular orbit: $\frac{1}{2}\mu v^2 - GM\mu/r$ with $v^2 = GM/r$.
(2) $1/r = (\omega^2/GM)^{1/3}$, so $GM/r = (GM)^{2/3}\omega^{2/3}$.
(3) $r^4\omega^6 = (GM)^{4/3}\omega^{-8/3}\omega^6 = (GM)^{4/3}\omega^{10/3}$.

(4) Collect the masses: $\mu^2 M^{4/3} = (\mu^{3/5} M^{2/5})^{10/3}$, and the powers of G and c rearrange as $G^{7/3}/c^5 = (c^5/G)(G/c^3)^{10/3}$.

The combination $\mathcal{M}$ is the chirp mass of (T2.26): it appeared there in the amplitude of the wave and appears here again, on its own, in step (4).

The chirp rate. Energy balance says $\mathrm{d}E/\mathrm{d}t = -P_{\mathrm{GW}}$. Since E depends on time only through f, $\mathrm{d}E/\mathrm{d}t = (\mathrm{d}E/\mathrm{d}f)\,\dot{f}$, so

$$\dot{f} \overset{(1)}{=} \frac{P_{\mathrm{GW}}}{|\mathrm{d}E/\mathrm{d}f|} \overset{(2)}{=} \frac{\frac{32}{5}\frac{G^{7/3}}{c^5}\mu^2 M^{4/3}(\pi f)^{10/3}}{\frac{\mu}{3}(\pi G M)^{2/3} f^{-1/3}} \overset{(3)}{=} \frac{96}{5}\,\pi^{8/3}\Big(\frac{G\mathcal{M}}{c^3}\Big)^{5/3} f^{11/3}. \tag{T2.43}$$

(1) Energy balance, with $\mathrm{d}E/\mathrm{d}f < 0$ (a higher frequency is a more bound orbit).
(2) P_{GW} from step (3) of (T2.42); $\mathrm{d}E/\mathrm{d}f = -\frac{\mu}{3}(\pi G M)^{2/3} f^{-1/3}$ from (T2.41).
(3) One μ cancels; $\mu M^{4/3}/M^{2/3} = \mu M^{2/3} = \mathcal{M}^{5/3}$, $G^{7/3}/G^{2/3} = G^{5/3}$ so $G^{5/3}/c^5 = (G/c^3)^{5/3}$, and $\pi^{10/3}/\pi^{2/3} \cdot f^{10/3} f^{1/3} = \pi^{8/3} f^{11/3}$.

Key idea: The Newtonian chirp

A circular binary radiating quadrupole waves sweeps up in frequency at the rate $\dot{f} = \frac{96}{5}\pi^{8/3}(G\mathcal{M}/c^3)^{5/3} f^{11/3}$. At this order the masses enter only through the chirp mass $\mathcal{M} = (m_1 m_2)^{3/5}/(m_1 + m_2)^{1/5}$, so the data measure $\mathcal{M}$ first and the separate masses only through smaller corrections.

The product $G\mathcal{M}/c^3$ has units of time. For the Sun it is $GM_\odot/c^3 = 4.93\,\mu\mathrm{s}$, which is the handiest number in gravitational-wave astronomy: a mass in solar units times $4.93\,\mu$s is that mass as a time.

Frequency against time, and the number of cycles. Since $\mathrm{d}t = \mathrm{d}f/\dot{f}$, the time left before the frequency runs off to infinity (the formal coalescence time t_c) is

$$\tau(f) \equiv t_c - t \overset{(1)}{=} \int_f^\infty \frac{\mathrm{d}f'}{\dot{f}(f')} \overset{(2)}{=} \frac{5}{96}\pi^{-8/3}\Big(\frac{G\mathcal{M}}{c^3}\Big)^{-5/3} \cdot \frac{3}{8} f^{-8/3} = \frac{5}{256}\Big(\frac{G\mathcal{M}}{c^3}\Big)^{-5/3}(\pi f)^{-8/3}, \tag{T2.44}$$

(1) Time is the integral of $\mathrm{d}f/\dot{f}$ along the sweep.
(2) $\int_f^\infty f'^{-11/3}\mathrm{d}f' = \frac{3}{8}f^{-8/3}$; then $\frac{5}{96}\cdot\frac{3}{8} = \frac{5}{256}$ and $\pi^{-8/3} f^{-8/3} = (\pi f)^{-8/3}$.

and inverting, $f(\tau) = \frac{1}{8\pi}\big(G\mathcal{M}/c^3\big)^{-5/8}\big(5/\tau\big)^{3/8}$ [85, eq. (24)]. The number of wave cycles between f_1 and f_2 is $N = \int f\,\mathrm{d}t = \int f\,\mathrm{d}f/\dot{f}$:

$$N(f_1, f_2) = \frac{1}{32\pi^{8/3}}\Big(\frac{G\mathcal{M}}{c^3}\Big)^{-5/3}\big(f_1^{-5/3} - f_2^{-5/3}\big), \tag{T2.45}$$

from $\int f^{-8/3}\mathrm{d}f = -\frac{3}{5}f^{-5/3}$ and $\frac{3}{5}\cdot\frac{5}{96} = \frac{1}{32}$. The low-frequency end dominates: most cycles are spent where the binary is slow.

Where the inspiral ends. In Newton's theory circular orbits exist all the way down. In general relativity they stop being stable inside the **innermost stable circular orbit** (ISCO), $r_{\mathrm{ISCO}} = 6GM/c^2$, three Schwarzschild radii; inside it the small body plunges. Kepler's law there gives $\omega^2 = c^6/(216\,G^2M^2)$, so

$$f_{\mathrm{ISCO}} = \frac{c^3}{6^{3/2}\pi G M} \approx 4.4\,\mathrm{kHz} \times \frac{M_\odot}{M}. \tag{T2.46}$$

Papers on dark-matter environments stop their inspirals here [82, 86], and so do we.

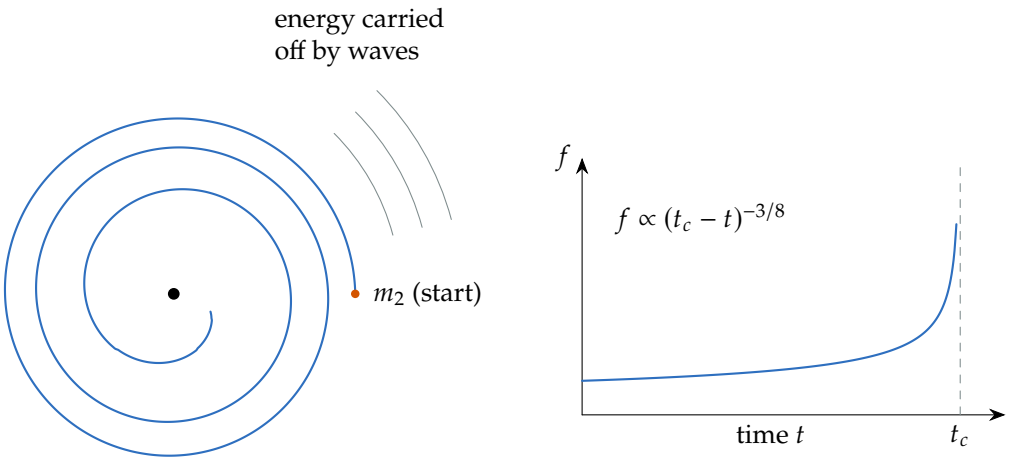


Figure T2.10: The chirp. Left: the light body's path around the heavy one (dot) for a Newtonian inspiral, the last three turns, drawn to scale in shape ($r \propto (\phi_c - \phi)^{2/5}$, which follows from $r \propto \tau^{1/4}$ and $\phi \propto \tau^{5/8}$); each turn is tighter and quicker than the last. Right: the wave frequency of (T2.44) inverted, rising slowly for a long time and then very fast just before t_c.

T2.4.3 The phase of the wave, and the two numbers that fix it

The detector does not record the frequency; it records the wave, $h(t) = A(t)\cos\phi(t)$ for one polarisation, with the amplitude $A \propto f^{2/3}$ of (T2.26) and a phase $\phi(t)$ whose rate of change is the angular frequency, $\dot\phi = 2\pi f$. So the phase is the integral of the frequency history. Write $T_{\mathcal{M}} \equiv G\mathcal{M}/c^3$, the chirp mass as a time, and $\tau = t_c - t$. The inverse of (T2.44) is $2\pi f(\tau) = \frac{1}{4}T_{\mathcal{M}}^{-5/8}5^{3/8}\tau^{-3/8}$, and since $\mathrm{d}\tau = -\mathrm{d}t$,

$$\phi(t) \overset{(1)}{=} \phi_c - \int_0^\tau 2\pi f(\tau')\,\mathrm{d}\tau' \overset{(2)}{=} \phi_c - \tfrac{1}{4}T_{\mathcal{M}}^{-5/8}5^{3/8}\cdot\tfrac{8}{5}\,\tau^{5/8} \overset{(3)}{=} \phi_c - 2\Big(\frac{\tau}{5T_{\mathcal{M}}}\Big)^{5/8}, \tag{T2.47}$$

(1) The phase grows by $2\pi f\,\mathrm{d}t$ in each moment. Integrating backwards from the end, where it reaches the value ϕ_c, the phase at time t is ϕ_c minus everything still to come.

(2) $\int_0^\tau \tau'^{-3/8}\mathrm{d}\tau' = \frac{8}{5}\tau^{5/8}$.

(3) $\frac{1}{4}\cdot\frac{8}{5}\cdot 5^{3/8} = \frac{2}{5}\cdot 5^{3/8} = 2\cdot 5^{-5/8}$, and $5^{-5/8}T_{\mathcal{M}}^{-5/8}\tau^{5/8} = (\tau/5T_{\mathcal{M}})^{5/8}$.

which is [87, eq. (17)]. Two integration constants appeared on the way, and both are now parameters of the signal.

The time of coalescence t_c is the moment at which the Newtonian frequency (T2.44) would formally run off to infinity. The real inspiral ends a little earlier, at the ISCO, but t_c is a convenient label for "when it all happens". Changing t_c by Δ slides the whole signal along the time axis, $h(t) \to h(t-\Delta)$, without changing its shape.

The phase at coalescence ϕ_c is the value of the phase at t_c. Since the wave phase is twice the orbital phase, changing ϕ_c by δ is the same as having started the two bodies at a different place on their orbit, turned by $\delta/2$ about the orbital axis. It does not change t_c, the masses or the frequency history; it only shifts the crests of the wave.

Why they are nuisance parameters. Neither number says anything about the physics of the source. t_c records when, on our clocks, the binary happened to merge; ϕ_c records where the two bodies happened to be on their orbit at that moment. Both are pure accident, both are unknown before the data arrive, and both must nevertheless be in the template: a template whose crests sit half a cycle away from the signal's crests correlates with it negatively, however right its masses. In the language of Chapter 5 they are **nuisance parameters**, unknowns we must fit or integrate over but do not care about. Section 10.8.4 shows that both can be searched

over almost for free, the phase with two filters and the arrival time with one Fourier transform, and Section T2.5 shows why: in the frequency domain they enter the phase only as the straight line $2\pi f t_c - \phi_c$.

T2.4.4 Examples: the benchmark binaries of the book

Two systems recur in the dark-matter literature and in Chapter 10. In both, a light black hole falls into a much heavier one. Binaries are named by their mass ratio $q = m_2/m_1$: comparable-mass binaries have $q \sim 1$, **intermediate mass-ratio inspirals** (IMRIs) roughly $q \sim 10^{-2}$–10^{-4}, and **extreme mass-ratio inspirals** (EMRIs) $q \lesssim 10^{-4}$; the boundary is a convention, and papers place it differently [86, 88]. A small q is what makes these systems good probes. By (T2.43), $\dot{f} \propto \mathcal{M}^{5/3} \approx m_2\, m_1^{2/3}$ when $m_2 \ll m_1$, so a light companion spirals in slowly and completes a huge number of orbits inside the environment.

The numbers. Every number quoted in the examples below comes from a few lines of `01_chirp.py`, which apply (T2.46), (T2.44), (T2.45) and (T2.43) to a pair of masses:

Listing T2.2: From two masses to the end of the inspiral (`code/chT2/01_chirp.py`)

```
28    Mc, M = chirp_mass(m1, m2), m1 + m2
29    fi = f_isco(M)
30    t_isco = tau_of_f(fi, Mc)                    # time left at the ISCO frequency
31    f4 = f_of_tau(t_isco + T4, Mc)               # frequency 4 years before the ISCO
32    N4 = cycles(f4, fi, Mc)
33    adiab = fdot_vac(fi, Mc) / fi**2             # fractional change of f per cycle
```

Line 29 is the ISCO frequency. Line 30 is the time left at that frequency, (T2.44). Line 31 adds four years to that time and inverts, giving the frequency four years before the ISCO. Line 32 counts the cycles between the two frequencies, and line 33 is the slow-evolution number checked in the last example. The helper functions (`tau_of_f`, `cycles`, and so on) sit in `lib_lisa.py` and are one-line transcriptions of the equations above.

Example: the EMRI of Kim et al. A $1\,M_\odot$ black hole falls into a $10^4\,M_\odot$ one [86, Sec. III F]. Kim et al. call it an IMRI; with $q = 10^{-4}$ it sits on the boundary, and we call it the EMRI. The chirp mass is

$$\mathcal{M} = \frac{(10^4 \cdot 1)^{3/5}}{(10^4 + 1)^{1/5}} \approx \frac{10^{2.4}}{10^{0.8}} = 10^{1.6} \approx 40\,M_\odot,$$

and the code gives $\mathcal{M} = 39.81\,M_\odot$. By (T2.46) the inspiral ends at $f_{\rm ISCO} = 4.4\,\text{kHz}/10^4 \approx 0.4397\,\text{Hz}$, inside the LISA band ($10^{-4}$–1 Hz). Four years earlier, from (T2.44), the frequency was 13.84 mHz, and by (T2.45) the binary completes $N = 2.787 \times 10^6$ wave cycles in those four years.

Example: the IMRI of Eda et al. and Coogan et al. A $1.4\,M_\odot$ object falls into a $10^3\,M_\odot$ black hole [82, 89]. Now $\mathcal{M} = 19.39\,M_\odot$ and $f_{\rm ISCO} = 4.391\,\text{Hz}$, above LISA's band: the last stretch of this inspiral is invisible to LISA. Four years before the ISCO the frequency is 21.7 mHz, and $N = 4.382 \times 10^6$ cycles follow. Figure T2.11 shows both frequency tracks.

Example: a GW150914-like binary passing through the LISA band. Robson, Cornish and Liu note that a binary like the first LIGO detection, five years before merger, sweeps from 16 to 29 mHz during a four-year mission [85, Sec. 2]. We can check this with (T2.44). Take $m_1 = 36$,

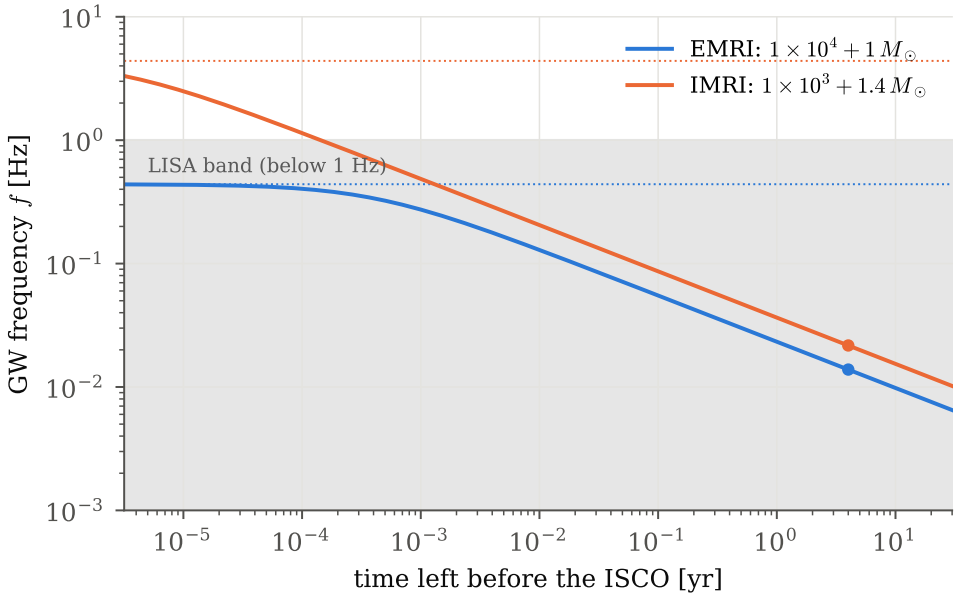


Figure T2.11: Frequency against the time left before the ISCO for the two benchmark binaries. Far from the end both tracks are straight lines of slope $-3/8$ on these log axes; close to the end they flatten onto $f_{\rm ISCO}$ (dotted). Dots mark the frequency four years before the ISCO. The shaded band is LISA's frequency range: the IMRI leaves it before its ISCO, the EMRI does not.

$m_2 = 29\,M_\odot$, so $\mathcal{M} = 28.1\,M_\odot$ and $G\mathcal{M}/c^3 = 28.1 \times 4.93\,\mu\text{s} = 1.38 \times 10^{-4}\,\text{s}$. Five years before merger, $\tau = 1.58 \times 10^8\,\text{s}$, and

$$f \overset{(1)}{=} \frac{1}{8\pi\, G\mathcal{M}/c^3}\left(\frac{5\,G\mathcal{M}/c^3}{\tau}\right)^{3/8} \overset{(2)}{=} \frac{1}{3.47 \times 10^{-3}\,\text{s}}\left(4.37 \times 10^{-12}\right)^{3/8}$$
$$\overset{(3)}{=} 288\,\text{Hz} \times 5.5 \times 10^{-5} \approx 16\,\text{mHz}.$$

(1) The inverse of (T2.44).
(2) $8\pi \times 1.38 \times 10^{-4} = 3.47 \times 10^{-3}$ and $5 \times 1.38 \times 10^{-4}/1.58 \times 10^8 = 4.37 \times 10^{-12}$.
(3) $(4.37 \times 10^{-12})^{3/8} = 10^{-11.36 \times 3/8} = 10^{-4.26}$.

The same steps with $\tau = 1\,\text{yr}$ give 29 mHz. The code returns 15.83 and 28.94 mHz, and the binary completes 2.534×10^6 cycles in between.

Example: checking the slow-evolution input. Input 2 assumed the orbit barely changes during one cycle. The fractional change of f during one cycle (duration $1/f$) is $\dot{f}/f^2 = \frac{96}{5}\pi^{8/3}(G\mathcal{M}f/c^3)^{5/3}$, largest at the ISCO. The code gives 6.839×10^{-5} for the EMRI and 9.550×10^{-4} for the IMRI. Even on the last orbit the frequency changes by less than a part in a thousand per cycle, so treating the orbit as a slowly shrinking circle is safe [87, eq. (25)].

What a million cycles buys. A template that drifts out of step with the signal by a single cycle, out of millions, no longer matches it. So the data pin down whatever sets the number of cycles very precisely [87, Sec. I]. By (T2.45), $N \propto \mathcal{M}^{-5/3}$, so a fractional change $\delta\mathcal{M}/\mathcal{M}$ shifts the count by $\delta N = \frac{5}{3} N\, \delta\mathcal{M}/\mathcal{M}$. Requiring $\delta N \lesssim 1$ gives $\delta\mathcal{M}/\mathcal{M} \lesssim 3/(5N)$, of order 10^{-7} for the EMRI. This is only a rough estimate: noise, a finite signal-to-noise ratio and correlations with other parameters weaken it, and the Fisher matrix of Chapter 10 does the calculation properly. For this EMRI with five years of data and signal-to-noise ratio 15, Kim et al. find $\mathcal{M} = 39.8^{+0.0012}_{-0.0006}\,M_\odot$ (95% range), a fractional error of about 2×10^{-5} [86, Fig. 4]. The parameters

that only set the amplitude, such as the distance, are measured with fractional errors of order one over the signal-to-noise ratio. This difference is the reason an environment is detectable at all: it acts on the phase, the most precisely measured part of the signal.

Watch out: Which mass is measured

The wave reaches us stretched by the cosmological redshift z, and a stretched chirp from mass $\mathcal{M}$ is indistinguishable from an unstretched chirp from mass $(1+z)\mathcal{M}$. The data measure the *detector-frame* (redshifted) chirp mass $(1+z)\mathcal{M}$, and (T2.43)–(T2.46) must be used with it [85, Sec. 2]. For the nearby benchmark binaries $z \ll 1$ and the difference is negligible.

Where this is used. The chirp rate (T2.43) and the binding energy (T2.41) are the vacuum baseline that Section 10.10.1 perturbs when an environment drains extra energy; the frequency four years before the ISCO and the cycle counts fix the band and the benchmark source of Section 10.11.2; and t_c, ϕ_c are the two nuisance parameters that Section 10.8.4 maximises over and that the Fisher matrix of Section 10.9 carries alongside the masses.

T2.5 The chirp in the frequency domain

The detector records a time series, but the likelihood of Section 10.8.2 is written in the frequency domain. There the stationary noise of the detector is simple: different Fourier components are independent, each with a variance set by the noise spectrum $S_n(f)$, and the log-likelihood is $-\frac{1}{2}(d-h\,|\,d-h)$ with the noise-weighted inner product $(a\,|\,b) = 4\,\mathrm{Re}\int_0^\infty \tilde{a}\tilde{b}^*/S_n\,\mathrm{d}f$. So the statistics needs the Fourier transform of the chirp of Section T2.4, as a formula in which the masses, the arrival time t_c and the phase ϕ_c appear explicitly.

A chirp makes this easy. At each moment it has one frequency, and it passes through each frequency once. So we ask in words: *which part of the time series builds the Fourier component at frequency f?* The answer is: essentially only the stretch during which the chirp's own frequency is close to f. At all other times the chirp and the probe wave $e^{-2\pi i f t}$ run at different rates, their product oscillates, and the oscillations cancel in the integral. Only near the moment t_f at which the chirp's frequency equals f do the two stay in step long enough to add up. Turning this sentence into a formula is the **stationary-phase approximation** (SPA) [87, eq. (21)].

T2.5.1 The Fourier convention, and why only positive frequencies count

The book's convention, used in every chapter, is

$$\tilde{h}(f) = \int_{-\infty}^{\infty} h(t)\, e^{-2\pi i f t}\,\mathrm{d}t, \qquad h(t) = \int_{-\infty}^{\infty} \tilde{h}(f)\, e^{2\pi i f t}\,\mathrm{d}f, \tag{T2.48}$$

with f in cycles per second, so that no factors of 2π stand in front of the integrals. A detector output is a real number at every instant, and that fixes the negative frequencies in terms of the positive ones:

$$\tilde{h}(-f) \overset{(1)}{=} \int h(t)\, e^{+2\pi i f t}\,\mathrm{d}t \overset{(2)}{=} \left[\int h(t)\, e^{-2\pi i f t}\,\mathrm{d}t\right]^* \overset{(3)}{=} \tilde{h}(f)^*. \tag{T2.49}$$

(1) The definition at $-f$.
(2) $h(t)$ is real, so complex-conjugating the integrand only flips the sign in the exponent.
(3) The definition at $+f$.

For example, the tone $\cos 2\pi f_0 t = \frac{1}{2}e^{2\pi i f_0 t} + \frac{1}{2}e^{-2\pi i f_0 t}$ has $\tilde{h}(f) = \frac{1}{2}\delta(f - f_0) + \frac{1}{2}\delta(f + f_0)$: half its weight sits at $+f_0$ and half, the complex conjugate, at $-f_0$. Since the negative half carries no new information, every quantity built from a real signal can be written with $f > 0$ only, at the price of a factor two. That is where the 4 in the inner product comes from: the 2 of the one-sided noise spectrum of Section 10.8.2 times the 2 from folding $-f$ onto $+f$. So it is enough to know $\tilde{h}(f)$ for $f > 0$.

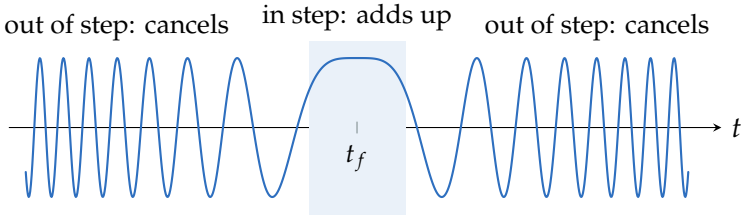


Figure T2.12: The real part of the integrand $e^{i[\phi(t)-2\pi f t]}$ near the moment t_f at which the chirp has frequency f, for a chirp whose frequency rises linearly: the phase difference grows like $(t - t_f)^2$. In the shaded window, of width about $1/\sqrt{\dot{f}}$, the integrand barely turns and its contributions add. Outside, positive and negative lobes alternate faster and faster and cancel.

T2.5.2 The stationary-phase formula

Inputs. The signal is $h(t) = A(t)\cos\phi(t)$, with an amplitude that changes slowly compared with the phase ($\dot{A}/A \ll \dot{\phi}$) and a frequency $\dot{\phi}/2\pi$ that changes slowly, $\ddot{\phi} \ll \dot{\phi}^2$. The second condition is $\dot{f}/f^2 \ll 1$, the slow-evolution number already checked in Section T2.4. We want $\tilde{h}(f)$ for $f > 0$.

Two exponentials. Writing the cosine as $\cos\phi = \frac{1}{2}(e^{i\phi} + e^{-i\phi})$,

$$\tilde{h}(f) = \frac{1}{2}\int A(t)\, e^{i[\phi(t)-2\pi f t]}\,\mathrm{d}t + \frac{1}{2}\int A(t)\, e^{-i[\phi(t)+2\pi f t]}\,\mathrm{d}t. \tag{T2.50}$$

The second integral is negligible. Its phase $\phi(t) + 2\pi f t$ always grows, at the rate $\dot{\phi} + 2\pi f \geq 2\pi f$, so its integrand turns steadily and its positive and negative lobes cancel. Integrating by parts once shows how well: $\int A e^{i\theta}\mathrm{d}t = [A e^{i\theta}/(i\dot{\theta})] - \int(\mathrm{d}/\mathrm{d}t)(A/i\dot{\theta})\, e^{i\theta}\mathrm{d}t$, so the integral is of order $A/(2\pi f)$, the amplitude times the time the integrand needs to turn once. We will see that the first integral is of order $A/\sqrt{\dot{f}}$; the ratio is $\sqrt{\dot{f}}/f = (\dot{f}/f^2)^{1/2} \ll 1$.

Where the first integral comes from. Call its phase $\vartheta(t) \equiv \phi(t) - 2\pi f t$. It is *stationary*, $\dot{\vartheta} = 0$, where $\dot{\phi} = 2\pi f$: at the moment t_f when the chirp's own frequency equals f. Near that moment Taylor-expand, $\vartheta(t) = \vartheta(t_f) + 0\cdot(t - t_f) + \frac{1}{2}\ddot{\phi}(t_f)(t - t_f)^2 + \ldots$, the linear term vanishing by the definition of t_f, and freeze the slowly varying amplitude at $A(t_f)$. Then

$$\begin{aligned}\tilde{h}(f) &\overset{(1)}{=} \frac{1}{2}A(t_f)\, e^{i\vartheta(t_f)}\int_{-\infty}^{\infty} e^{\frac{i}{2}\ddot{\phi}(t_f)\,s^2}\,\mathrm{d}s \overset{(2)}{=} \frac{1}{2}A(t_f)\, e^{i\vartheta(t_f)}\sqrt{\frac{2\pi}{\ddot{\phi}(t_f)}}\, e^{i\pi/4} \\ &\overset{(3)}{=} \frac{1}{2}A(t_f)\,\dot{f}(t_f)^{-1/2}\, e^{-i\Psi(f)}, \qquad \Psi(f) \equiv 2\pi f t_f - \phi(t_f) - \frac{\pi}{4}.\end{aligned} \tag{T2.51}$$

(1) The expansion with $s = t - t_f$. The limits can be taken to $\pm\infty$, because far from t_f the integrand oscillates ever faster and contributes nothing more (Figure T2.12).

(2) The Fresnel integral, derived just below: $\int e^{ias^2/2}\mathrm{d}s = \sqrt{2\pi/a}\, e^{i\pi/4}$ for $a > 0$, here with $a = \ddot{\phi}(t_f) > 0$ because the frequency rises.

(3) $\ddot{\phi} = 2\pi\dot{f}$, so $\sqrt{2\pi/\ddot{\phi}} = \dot{f}^{-1/2}$. Collecting the phases, $\vartheta(t_f) + \frac{\pi}{4} = \phi(t_f) - 2\pi f t_f + \frac{\pi}{4} = -\Psi(f)$.

The Fresnel integral. The integrand $e^{ias^2/2}$ has modulus one everywhere, so the integral converges only because of the cancellations. Make that explicit by adding a small damping $e^{-\epsilon s^2}$, $\epsilon > 0$, and letting $\epsilon \to 0$ at the end. With $b \equiv \epsilon - ia/2$,

$$\int_{-\infty}^{\infty} e^{-\epsilon s^2 + ias^2/2}\,\mathrm{d}s \overset{(1)}{=} \int_{-\infty}^{\infty} e^{-bs^2}\,\mathrm{d}s \overset{(2)}{=} \sqrt{\frac{\pi}{b}} \xrightarrow{\epsilon\to 0} \sqrt{\frac{\pi}{-ia/2}} \overset{(3)}{=} \sqrt{\frac{2\pi}{a}}\, e^{i\pi/4}. \tag{T2.52}$$

(1) Definition of b; its real part ϵ is positive, so the integral converges absolutely.

(2) The Gaussian integral $\int e^{-bs^2}\mathrm{d}s = \sqrt{\pi/b}$, familiar for real $b > 0$. Both sides are analytic functions of b in the half-plane $\mathrm{Re}\, b > 0$ (with the square root positive for real b), and two analytic functions that agree on the positive real axis agree everywhere in that half-plane.

(3) $-i = e^{-i\pi/2}$, so $1/(-i) = e^{i\pi/2}$ and its square root is $e^{i\pi/4}$; the branch is the one that is continuous from $\mathrm{Re}\, b > 0$, where $\arg b \to -\pi/2$ and $\arg\sqrt{1/b} \to +\pi/4$.

The real and imaginary parts are the classic Fresnel integrals, $\int \cos(as^2/2)\mathrm{d}s = \int \sin(as^2/2)\mathrm{d}s = \sqrt{\pi/a}$; the phase $e^{i\pi/4}$ says they are equal.

How accurate is it? The width of the window that contributes is set by the Gaussian factor in step (1): $|t - t_f| \lesssim 1/\sqrt{\ddot{\phi}} \sim 1/\sqrt{\dot{f}}$. During that window the chirp completes about $f/\sqrt{\dot{f}} = (\dot{f}/f^2)^{-1/2}$ cycles, which is large precisely when the slow-evolution number is small. The first neglected term of the Taylor expansion, $\frac{1}{6}\dddot{\phi}\,(t - t_f)^3$, is of size $\dddot{\phi}/\ddot{\phi}^{3/2}$ across the window; with $\ddot{\phi} = 2\pi\ddot{f}$ and $\ddot{f} = \frac{11}{3}\dot{f}^2/f$ for the Newtonian chirp, that is again of order $(\dot{f}/f^2)^{1/2}$. So the same condition makes the orbit adiabatic and the SPA accurate.[1]

What the phase knows. Differentiate Ψ and use $\dot{\phi}(t_f) = 2\pi f$:

$$\frac{\mathrm{d}\Psi}{\mathrm{d}f} \overset{(1)}{=} 2\pi t_f + 2\pi f\frac{\mathrm{d}t_f}{\mathrm{d}f} - \dot{\phi}(t_f)\frac{\mathrm{d}t_f}{\mathrm{d}f} \overset{(2)}{=} 2\pi t(f), \qquad \frac{\mathrm{d}^2\Psi}{\mathrm{d}f^2} \overset{(3)}{=} \frac{2\pi}{\dot{f}}. \tag{T2.53}$$

(1) Product rule on $2\pi f t_f$ and chain rule on $\phi(t_f)$, since t_f depends on f.

(2) The last two terms cancel because $\dot{\phi}(t_f) = 2\pi f$. We write $t(f) \equiv t_f$.

(3) $\mathrm{d}t/\mathrm{d}f = 1/\dot{f}$.

The slope of the Fourier phase with respect to f gives 2π times the arrival time of that frequency, and its curvature gives 2π divided by the chirp rate. Whatever changes how fast the binary sweeps, an environment included, changes Ψ through (T2.53). Section T2.10 applies exactly this equation to a binary that is losing energy to its surroundings as well as to waves.

The Newtonian waveform. Integrate (T2.53) with $t(f) = t_c - \tau(f)$ from (T2.44):

$$\begin{aligned}\Psi(f) &\overset{(1)}{=} 2\pi f t_c - 2\pi\int \tau(f)\,\mathrm{d}f + \text{const} \overset{(2)}{=} 2\pi f t_c + 2\pi\cdot\frac{3}{5}\cdot\frac{5}{256}\left(\frac{G\mathcal{M}}{c^3}\right)^{-5/3}\pi^{-8/3}f^{-5/3} + \text{const}\\ &\overset{(3)}{=} 2\pi f t_c - \phi_c - \frac{\pi}{4} + \frac{3}{128}\left(\frac{\pi G\mathcal{M}f}{c^3}\right)^{-5/3}.\end{aligned} \tag{T2.54}$$

(1) Integrate $\mathrm{d}\Psi/\mathrm{d}f = 2\pi t_c - 2\pi\tau(f)$ once.

[1] Cutler and Flanagan, Eda et al. and Kim et al. define the Fourier transform with $e^{+2\pi i f t}$ and therefore write $e^{+i\Psi(f)}$. The function $\Psi(f)$ is the same; the two conventions give complex-conjugate $\tilde{h}$, and the inner product, the signal-to-noise ratio and the Fisher matrix, which contain only $\mathrm{Re}\,\tilde{a}\tilde{b}^*$, are unchanged.

(2) $\tau = \frac{5}{256}(G\mathcal{M}/c^3)^{-5/3}\pi^{-8/3}f^{-8/3}$ and $\int f^{-8/3}\mathrm{d}f = -\frac{3}{5}f^{-5/3}$.
(3) $2\pi \cdot \frac{3}{256}\pi^{-8/3} = \frac{3}{128}\pi^{-5/3}$; the integration constant is named after the phase at coalescence, $-\phi_c - \pi/4$.

This agrees with [87, eq. (24)], where the last term is written $\frac{3}{4}(8\pi\mathcal{M}f)^{-5/3}$ ($G = c = 1$); $\frac{3}{4}\cdot 8^{-5/3} = \frac{3}{4}\cdot\frac{1}{32} = \frac{3}{128}$. The phase constant $-\phi_c - \pi/4$ and the term $2\pi f t_c$ are exactly the two nuisance parameters of Section T2.4.3, and in the frequency domain they enter only as a straight line in f: a shift of the arrival time is a phase that grows linearly with frequency.

The amplitude. For the amplitude we need $A(t_f)$, the face-on amplitude of (T2.26) at frequency f, with the distance now the luminosity distance D_L, and $\dot{f}$ from (T2.43). With $T_\mathcal{M} = G\mathcal{M}/c^3$,

$$\tfrac{1}{2}A\,\dot{f}^{-1/2} \overset{(1)}{=} \frac{1}{2}\cdot 4T_\mathcal{M}^{5/3}(\pi f)^{2/3}\frac{c}{D_L}\cdot\left(\frac{96}{5}\right)^{-1/2}\pi^{-4/3}T_\mathcal{M}^{-5/6}f^{-11/6} \overset{(2)}{=} \sqrt{\frac{5}{24}}\,\pi^{-2/3}\,T_\mathcal{M}^{5/6}\,\frac{c}{D_L}\,f^{-7/6}, \tag{T2.55}$$

(1) $A = 4(G\mathcal{M})^{5/3}(\pi f)^{2/3}/(c^4 D_L) = 4T_\mathcal{M}^{5/3}(\pi f)^{2/3}c/D_L$, and $\dot{f}^{-1/2} = (\frac{96}{5})^{-1/2}\pi^{-4/3}T_\mathcal{M}^{-5/6}f^{-11/6}$.
(2) $2(\frac{5}{96})^{1/2} = (\frac{20}{96})^{1/2} = (\frac{5}{24})^{1/2}$; $\pi^{2/3-4/3} = \pi^{-2/3}$; $T_\mathcal{M}^{5/3-5/6} = T_\mathcal{M}^{5/6}$; $f^{2/3-11/6} = f^{-7/6}$.

so that

$$\tilde{h}(f) = \mathcal{A}\,f^{-7/6}e^{-i\Psi(f)}, \qquad \mathcal{A} = \sqrt{\frac{5}{24}}\,\pi^{-2/3}\left(\frac{G\mathcal{M}}{c^3}\right)^{5/6}\frac{c}{D_L}, \tag{T2.56}$$

the amplitude of Robson et al. [85, eq. (20)] and of Kim et al. [86, eq. (45)], written out; the time-domain amplitude is that of Eda et al. [82, eq. (21)]. A real source is not face-on, and the detector does not see both polarisations equally. By (T2.25) the polarisations of a circular binary have amplitudes $\frac{1}{2}(1+x^2)$ and x times A, with $x = \cos\iota$, and Section T2.3 showed that averaging over the orientation of the orbit multiplies the signal power by $\frac{4}{5}$, (T2.36), while the average over sky position and polarisation is put into the noise curve (Section T2.7). Hence the sky-averaged waveform $\tilde{h} = \sqrt{4/5}\,\mathcal{A}f^{-7/6}e^{-i\Psi}$ used in the forecasts [86, eq. (44)].

Key idea: The signal model of the forecasts

In the stationary-phase approximation a slowly chirping binary has $\tilde{h}(f) = \frac{1}{2}A(t_f)\dot{f}^{-1/2}e^{-i\Psi(f)}$, with $\mathrm{d}\Psi/\mathrm{d}f = 2\pi t(f)$ and $\mathrm{d}^2\Psi/\mathrm{d}f^2 = 2\pi/\dot{f}$. For the Newtonian chirp this is $\mathcal{A}f^{-7/6}e^{-i\Psi}$ with $\Psi = 2\pi f t_c - \phi_c - \pi/4 + \frac{3}{128}(\pi G\mathcal{M}f/c^3)^{-5/3}$: four parameters, $\ln\mathcal{A}$, $\mathcal{M}$, t_c and ϕ_c.

T2.5.3 Checking the formula numerically

A stellar-mass chirp lasts seconds, so its Fourier transform can be computed directly and compared with (T2.51). The script `02_spa_check.py` builds a Newtonian chirp with $\mathcal{M} = 10\,M_\odot$ from 20 Hz to 1 kHz, 4.729 s long, sampled at 4096 Hz:

Listing T2.3: A Newtonian chirp in the time domain (`code/chT2/02_spa_check.py`)

```
t = np.arange(-tau0, -tau1, 1 / fs)            # time from coalescence (t_c = 0)
tau = -t
f_t = (1 / (8 * np.pi * m)) * (5 * m / tau) ** (3 / 8)     # f(t), Newtonian
phi_t = -2 * (tau / (5 * m)) ** (5 / 8)                     # GW phase, phi_c = 0
A_t = (f_t / f0) ** (2 / 3)                                 # amplitude ~ f^{2/3}
# cosine ramps: 0.3 s at the start (f near 20 Hz), 2 ms at the end (f > 360 Hz),
# so the band 30-150 Hz used for the comparison is untouched by the window
win = np.ones_like(t)
r0, r1 = t < t[0] + 0.3, t > t[-1] - 2e-3
```

```
win[r0] = 0.5 * (1 - np.cos(np.pi * (t[r0] - t[0]) / 0.3))
win[r1] = 0.5 * (1 - np.cos(np.pi * (t[-1] - t[r1]) / 2e-3))
h = A_t * np.cos(phi_t) * win
```

Lines 28–29 are the inverse of (T2.44) and its time integral, $\phi(t) = \phi_c - 2[\tau/(5G\mathcal{M}/c^3)]^{5/8}$ [87, eq. (22)]. The amplitude $\propto f^{2/3}$ follows the quadrupole formula (the overall scale is irrelevant here). The cosine ramps at the two ends switch the signal on and off gently. An abrupt start would add the transform of a step, which rings across all frequencies. The ramps are kept short and placed at the ends of the sweep, so the band we compare, 30–150 Hz, is untouched. The prediction is then four lines:

Listing T2.4: The stationary-phase prediction (`code/chT2/02_spa_check.py`)

```
# SPA prediction
band = (f > 30) & (f < 150)
fb = f[band]
tf = -tau_of_f(fb, Mc)
phif = -2 * (tau_of_f(fb, Mc) / (5 * m)) ** (5 / 8)
Psi = 2 * np.pi * fb * tf - phif - np.pi / 4
Hspa = 0.5 * (fb / f0) ** (2 / 3) / np.sqrt(fdot_vac(fb, Mc)) * np.exp(-1j * Psi)
```

Here `tf` is t_f from (T2.44), `phif` is $\phi(t_f)$, and line 50 is (T2.51). Figure T2.13 shows the result. In the comparison band the modulus agrees with the FFT to within 3.655%, and the phase difference has an rms of 0.01296 rad while Ψ itself changes by 168.4 rad across the band. The small ripples that remain are the leftovers of switching the signal on and off, not a failure of the approximation.

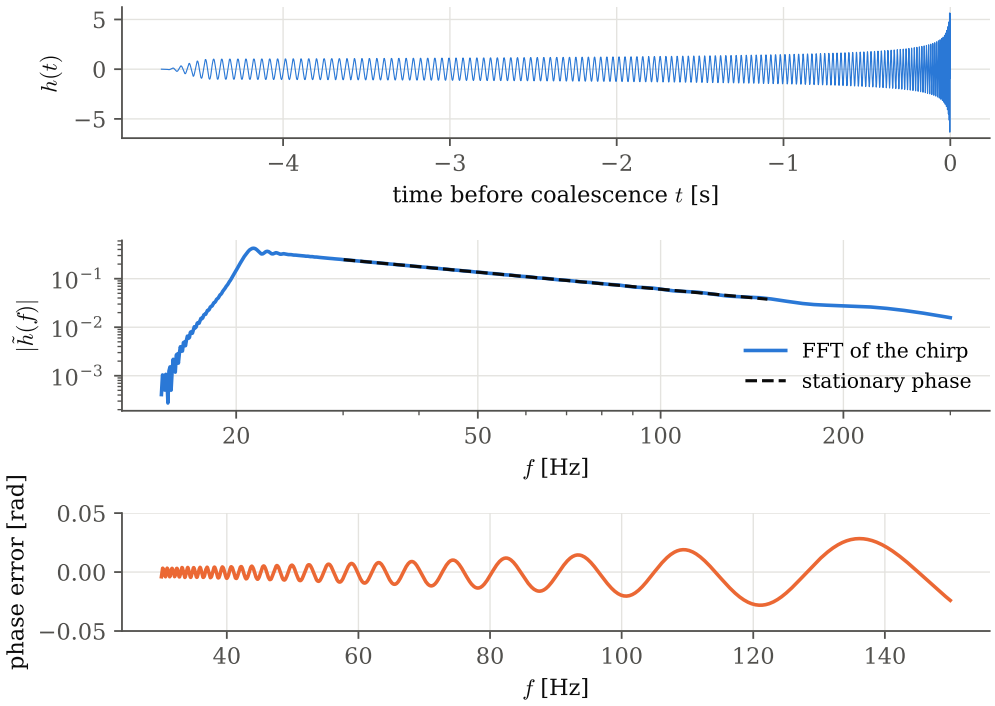


Figure T2.13: The stationary-phase formula against a direct FFT. Top: the chirp in time, its amplitude and frequency rising towards coalescence. Middle: $|\tilde{h}|$ from the FFT and from (T2.51), which fall together as $f^{-7/6}$ in the band; the FFT drops at 20 Hz, where the signal starts. Bottom: the phase of the FFT minus the predicted $-\Psi(f)$, a few hundredths of a radian.

Example: two phases that are easy to confuse. Papers on environments plot the *time-domain* phase left before merger, $\Phi(f) = 2\pi \int_t^{t_c} f \, \mathrm{d}t$, which is 2π times the number of cycles [89], [86, eq. (48)]. The Fisher matrix uses the Fourier phase $\Psi(f)$. For the Newtonian chirp, (T2.45) with $f_2 \to \infty$ gives $\Phi(f) = \frac{1}{16}x$, with $x \equiv (\pi G\mathcal{M}f/c^3)^{-5/3}$. With $t_c = \phi_c = 0$, the phase at time t_f is $\phi(t_f) = -\Phi(f)$, so

$$\Psi(f) + \frac{\pi}{4} \overset{(1)}{=} 2\pi f\, t_f - \phi(t_f) \overset{(2)}{=} -2\pi f\, \tau(f) + \Phi(f) \overset{(3)}{=} -\frac{10}{256}\, x + \frac{16}{256}\, x = \frac{3}{128}\, x.$$

(1) Definition of Ψ in (T2.51).

(2) $t_f = -\tau(f)$ and $\phi(t_f) = -\Phi(f)$, since the phase still to come is $\Phi(f)$ and the phase reaches $\phi_c = 0$ at t_c.

(3) $2\pi f\tau = 2\pi \cdot \frac{5}{256}(G\mathcal{M}/c^3)^{-5/3}\pi^{-8/3}f^{-5/3} = \frac{10}{256}x$, and $\frac{1}{16} = \frac{16}{256}$.

So Ψ carries $\frac{3}{8}$ of the time-domain phase, because the term $2\pi f t_f$ removes the part a shift of the arrival time can absorb. For the EMRI four years before the ISCO, $\Phi \approx 2\pi \times 2.787 \times 10^6$ while the f-dependent part of Ψ is about three eighths of that. Both phases are correct; they answer different questions, and a dephasing quoted in one cannot be compared with a phase error in the other without this conversion.

Where this is used. The waveform (T2.56) with the phase (T2.54) is the signal model of Section 10.9.2, extended there by the post-Newtonian terms derived in Section T2.6. The fact that t_c and ϕ_c enter Ψ only as $2\pi f t_c - \phi_c$ is what lets Section 10.8.4 search all arrival times with one inverse Fourier transform. The relations (T2.53) are the bridge from any change of the sweep rate to a change of the phase, used in Section 10.10.1, and Section 10.12.1 carries out the conversion between the two kinds of phase.

T2.6 Corrections to the phase: post-Newtonian orders and the ppE form

The chirp of Section T2.4 rests on Newton's law of gravity and on the quadrupole formula, both of which are the first term of an expansion in v/c. Far from merger that is excellent. Near the ISCO the orbital speed is $v = c/\sqrt{6} \approx 0.4c$, and the corrections are tens of per cent. Even far from merger, a correction of a few per cent to the rate of the inspiral is not small in the only currency that matters for a matched filter: over a million cycles, a few per cent is tens of thousands of cycles, and a template that slips by one cycle has lost the signal (Section T2.4). Other physics acts the same way. A cloud of dark matter, a disc of gas, or a departure from general relativity each changes how fast the binary loses energy, by a fraction that depends on the frequency.

So the question is: *if the binding energy or the radiated power differs from its Newtonian value by a small fractional term, how does that change the Fourier phase $\Psi(f)$, and which power of f does each kind of correction produce?* The answer is one rule. A correction that scales as v^p relative to the Newtonian term adds to Ψ a term proportional to $f^{(p-5)/3}$, with a coefficient we can compute once and for all. General-relativistic corrections have $p = 2, 3, 4, \ldots$; environments and some modifications of gravity have negative p. The rule turns the whole zoo of corrections into one family, $\delta\Psi = \beta u^b$, which is the form the forecast of Chapter 10 uses.

T2.6.1 The expansion parameter

Measure the speed of the orbit by the observable wave frequency. For a Newtonian circular orbit, Kepler's law gives the orbital speed $v_{\mathrm{orb}} = r\omega = (GM\omega)^{1/3}$, and with $\omega = \pi f$ from

(T2.24),

$$v \equiv \left(\frac{\pi G M f}{c^3}\right)^{1/3} = \frac{v_{\rm orb}}{c}, \qquad x \equiv v^2, \tag{T2.57}$$

with $M = m_1 + m_2$ the total mass. A correction of relative size v^{2k} is called **kPN** (post-Newtonian) order, so 1PN means v^2, 1.5PN means v^3. Using f rather than the orbital radius r is deliberate: r depends on the coordinates used to describe the strong-field region, while the frequency of the wave at the detector does not [87, Sec. III A]. Two mass combinations will appear: the **symmetric mass ratio** $\eta = m_1 m_2/M^2 = \mu/M$ [$\frac{1}{4}$ for equal masses, about m_2/m_1 when one body is much lighter], and the chirp mass, $\mathcal{M} = \eta^{3/5} M$. The reduced frequency of the Newtonian phase is then

$$u \equiv \frac{\pi G \mathcal{M} f}{c^3} = \eta^{3/5} v^3, \qquad u^{-5/3} = \eta^{-1} v^{-5}. \tag{T2.58}$$

T2.6.2 From the energy and the flux to the phase

Inputs. The orbit is still circular and evolves slowly, so energy balance $\mathrm{d}E/\mathrm{d}t = -\mathcal{F}$ holds at each moment, with $E(v)$ the binding energy and $\mathcal{F}(v)$ the power radiated, now allowed to carry corrections. In the Newtonian limit, by (T2.41) and (T2.42) written in terms of v,

$$E_{\rm N} = -\tfrac{1}{2}\mu c^2 v^2, \qquad \mathcal{F}_{\rm N} = \frac{32}{5}\frac{c^5}{G}\eta^2 v^{10}. \tag{T2.59}$$

(For E: $\frac{1}{2}\mu(\pi G M f)^{2/3} = \frac{1}{2}\mu c^2 v^2$. For the power: $u^{10/3} = \eta^2 v^{10}$ by (T2.58).)

The time. Energy balance gives $\mathrm{d}t = -\mathrm{d}E/\mathcal{F} = -(\mathrm{d}E/\mathrm{d}v)\,\mathrm{d}v/\mathcal{F}$. Write the corrections to $\mathrm{d}E/\mathrm{d}v$ and $\mathcal{F}$ together as one bracket,

$$\frac{\mathrm{d}t}{\mathrm{d}v} = -\frac{E'(v)}{\mathcal{F}(v)} = -\frac{E'_{\rm N}}{\mathcal{F}_{\rm N}}\left[1 + c_p v^p + \dots\right] = \frac{5GM}{32c^3\eta}\, v^{-9}\left[1 + c_p v^p + \dots\right], \tag{T2.60}$$

because $-E'_{\rm N}/\mathcal{F}_{\rm N} = \mu c^2 v \big/ \left(\frac{32}{5}\frac{c^5}{G}\eta^2 v^{10}\right) = \frac{5}{32}\frac{G\mu}{c^3\eta^2} v^{-9}$ and $\mu = \eta M$. The bracket may contain several terms; each acts separately at first order, so we follow one, $c_p v^p$.

The phase. From (T2.53), $\mathrm{d}\Psi/\mathrm{d}f = 2\pi t(f) = 2\pi t_c - 2\pi\tau(f)$, where $\tau = t_c - t$ is the time left. So $\Psi = 2\pi f t_c - \phi_c - \frac{\pi}{4} - 2\pi\int^f \tau\,\mathrm{d}f'$, with the integration constants named as in (T2.54). We need τ as a function of v, and $\mathrm{d}f$ in terms of $\mathrm{d}v$:

$$\begin{aligned}
\tau(v) &\overset{(1)}{=} \int_v^\infty \frac{\mathrm{d}t}{\mathrm{d}v'}\,\mathrm{d}v' \overset{(2)}{=} \frac{5GM}{32c^3\eta}\left[\frac{v^{-8}}{8} + c_p\frac{v^{p-8}}{8-p}\right], \qquad \mathrm{d}f \overset{(3)}{=} \frac{3c^3 v^2}{\pi G M}\,\mathrm{d}v, \\
-2\pi\int \tau\,\mathrm{d}f &\overset{(4)}{=} -\frac{15}{16\eta}\int\left[\frac{v^{-6}}{8} + c_p\frac{v^{p-6}}{8-p}\right]\mathrm{d}v \overset{(5)}{=} \frac{15}{16\eta}\left[\frac{v^{-5}}{40} + c_p\frac{v^{p-5}}{(8-p)(5-p)}\right] \\
&\overset{(6)}{=} \frac{3}{128\,\eta\,v^5}\left[1 + \frac{40\,c_p}{(8-p)(5-p)}\,v^p\right].
\end{aligned} \tag{T2.61}$$

(1) The time left is the integral of $\mathrm{d}t$ from the current speed to the end, where $v \to \infty$ formally, as in (T2.44).

(2) $\int_v^\infty v'^q \mathrm{d}v' = v^{q+1}/(-q-1)$ for $q < -1$; here $q = -9$ and $q = p - 9$, which needs $p < 8$.

(3) $f = c^3 v^3/(\pi G M)$ from (T2.57).

(4) Multiply: $2\pi \cdot \frac{5GM}{32c^3\eta} \cdot \frac{3c^3}{\pi G M} = \frac{15}{16\eta}$, and $v^{-8}v^2 = v^{-6}$.

(5) $\int v^q \mathrm{d}v = v^{q+1}/(q+1)$: for $q = -6$ this is $-v^{-5}/5$ and for $q = p-6$ it is $-v^{p-5}/(5-p)$, which needs $p \neq 5$. The two minus signs cancel the one in front; the constants of integration join ϕ_c.

(6) Factor out the Newtonian term: $\frac{15}{16}\cdot\frac{1}{40}=\frac{3}{128}$, and the ratio of the two coefficients is $\frac{1/((8-p)(5-p))}{1/40}$.

For $p=0$ only the Newtonian term is left, $\frac{3}{128}\eta^{-1}v^{-5}=\frac{3}{128}u^{-5/3}$, which is (T2.54) again. A correction $c_p v^p$ in the rate of the inspiral therefore becomes a phase term

$$\delta\Psi=\frac{3}{128}\,u^{-5/3}\,\frac{40\,c_p}{(8-p)(5-p)}\,v^p\;\propto\;f^{(p-5)/3}. \tag{T2.62}$$

At $p=5$ the integral in step (5) is a logarithm instead of a power; this is why the 2.5PN term of the phase contains $\ln f$.

T2.6.3 The 1PN and 1.5PN phase

The inputs from general relativity. The binding energy to 1PN order and the flux to 1.5PN order are standard results, which Cutler and Flanagan quote in terms of the orbital radius r in a particular coordinate system [87, eqs. (35)–(37)], [90]. In units $G=c=1$,

$$\Omega^2=\frac{M}{r^3}\Big[1-(3-\eta)\frac{M}{r}\Big],\qquad E=-\frac{\mu M}{2r}\Big[1-\frac{7-\eta}{4}\frac{M}{r}\Big], \tag{T2.63}$$

and the flux is $\frac{32}{5}\eta^2(M\Omega)^{10/3}[1-(\frac{1247}{336}+\frac{35}{12}\eta)\frac{M}{r}+4\pi(\frac{M}{r})^{3/2}]$. The first expression is the square of their eq. (35), to the same order. The 4π is the **tail**: part of the outgoing wave scatters off the curved spacetime around the binary on its way out, and the scattered part interferes with the rest.

Rewritten in v. With $v^2=(M\Omega)^{2/3}$,

$$\begin{aligned}
v^2 &\overset{(1)}{=}\frac{M}{r}\Big[1-(3-\eta)\frac{M}{r}\Big]^{1/3}\overset{(2)}{=}\frac{M}{r}\Big[1-\Big(1-\frac{\eta}{3}\Big)\frac{M}{r}\Big],\qquad \frac{M}{r}\overset{(3)}{=}v^2\Big[1+\Big(1-\frac{\eta}{3}\Big)v^2\Big],\\
E &\overset{(4)}{=}-\frac{\mu}{2}v^2\Big[1+\Big(1-\frac{\eta}{3}\Big)v^2-\frac{7-\eta}{4}v^2\Big]\overset{(5)}{=}-\frac{\mu}{2}v^2\Big[1-\frac{9+\eta}{12}v^2\Big].
\end{aligned} \tag{T2.64}$$

(1) $(M\Omega)^{2/3}=(M^2/r^3)^{1/3}[\dots]^{1/3}=(M/r)[\dots]^{1/3}$.
(2) $(1-\epsilon)^{1/3}=1-\epsilon/3$.
(3) Invert to first order: if $v^2=y(1-ky)$ then $y=v^2(1+kv^2)$.
(4) Insert into E; in the 1PN bracket $M/r=v^2$ is enough.
(5) $1-\frac{\eta}{3}-\frac{7}{4}+\frac{\eta}{4}=-\frac{3}{4}-\frac{\eta}{12}=-\frac{9+\eta}{12}$.

The flux is already written with $(M\Omega)^{10/3}=v^{10}$ in front, and inside its bracket $M/r=v^2$ to the order needed. So, restoring c,

$$E'(v)=-\mu c^2 v\Big[1-\frac{9+\eta}{6}v^2\Big],\qquad \mathcal{F}=\frac{32}{5}\frac{c^5}{G}\eta^2v^{10}\Big[1-\Big(\frac{1247}{336}+\frac{35}{12}\eta\Big)v^2+4\pi v^3\Big], \tag{T2.65}$$

where E' comes from differentiating $v^2-\frac{9+\eta}{12}v^4$ to $2v-\frac{9+\eta}{3}v^3$.

The bracket of (T2.60). Dividing,

$$\frac{E'/E'_{\rm N}}{\mathcal{F}/\mathcal{F}_{\rm N}}\overset{(1)}{=}\Big[1-\frac{9+\eta}{6}v^2\Big]\Big[1+\Big(\frac{1247}{336}+\frac{35}{12}\eta\Big)v^2-4\pi v^3\Big]\overset{(2)}{=}1+\Big(\frac{743}{336}+\frac{11}{4}\eta\Big)v^2-4\pi v^3, \tag{T2.66}$$

(1) $1/(1-\epsilon)=1+\epsilon$ to first order, applied to the flux bracket.
(2) Multiply out to order v^3: $\frac{1247}{336}-\frac{9}{6}=\frac{1247-504}{336}=\frac{743}{336}$ and $\frac{35}{12}-\frac{1}{6}=\frac{33}{12}=\frac{11}{4}$.

so $c_2 = \frac{743}{336} + \frac{11}{4}\eta$ and $c_3 = -4\pi$. The rule (T2.61) gives the factors $\frac{40}{6\cdot3} = \frac{20}{9}$ for $p = 2$ and $\frac{40}{5\cdot2} = 4$ for $p = 3$:

$$\Psi(f) = 2\pi f t_c - \phi_c - \frac{\pi}{4} + \frac{3}{128}\,u^{-5/3}\Big[1 + \frac{20}{9}\Big(\frac{743}{336} + \frac{11}{4}\eta\Big)x - 16\pi x^{3/2}\Big], \tag{T2.67}$$

with $x = v^2$. This is Cutler and Flanagan's restricted post-Newtonian phase [87, eq. (42)], in our sign convention, and it is the waveform used in Section 10.9.2. The same bracket inverted gives the chirp rate, $\dot f \propto 1 - (\frac{743}{336} + \frac{11}{4}\eta)x + 4\pi x^{3/2}$, their eq. (38). The symmetric mass ratio η appears for the first time at 1PN: it is the post-Newtonian terms that let the data separate m_1 from m_2 once $\mathcal{M}$ is known.

Example: how many cycles each term is worth. Count cycles instead of phase: the number of wave cycles is $N = \int f\,\mathrm{d}t = \int f(\mathrm{d}t/\mathrm{d}v)\,\mathrm{d}v$, and with (T2.60) and $f = c^3v^3/(\pi GM)$,

$$N(v_1, v_2) = \frac{5}{32\pi\eta}\int_{v_1}^{v_2} v^{-6}\big[1 + c_2v^2 + c_3v^3\big]\mathrm{d}v = \frac{5}{32\pi\eta}\Big[\frac{v^{-5}}{5} + c_2\frac{v^{-3}}{3} + c_3\frac{v^{-2}}{2}\Big]_{v_2}^{v_1}. \tag{T2.68}$$

The first term is (T2.45) again. The script `14_pn_cycles.py` evaluates the three terms separately (Figure T2.14). For a pair of $1.4\,M_\odot$ neutron stars observed from 10 Hz to the ISCO by a ground-based detector, the Newtonian term gives 16 034 cycles, the 1PN term 441 and the tail −211, in agreement with the counts Blanchet tabulates for the same band (about 1.6×10^4, 441 and −211) [90]. For the book's intermediate mass-ratio benchmark in its last four years (v from 0.06954 to 0.4082), the counts are 4.38×10^6, 77 847 and −44 949; for the EMRI, 2.79×10^6, 1.66×10^5 and -1.69×10^5. Tens of thousands of cycles from each correction: no template can leave them out. Even the GW150914-like binary crossing the LISA band years before merger, at $v \approx 0.03$, gains 5 521 cycles from the 1PN term.

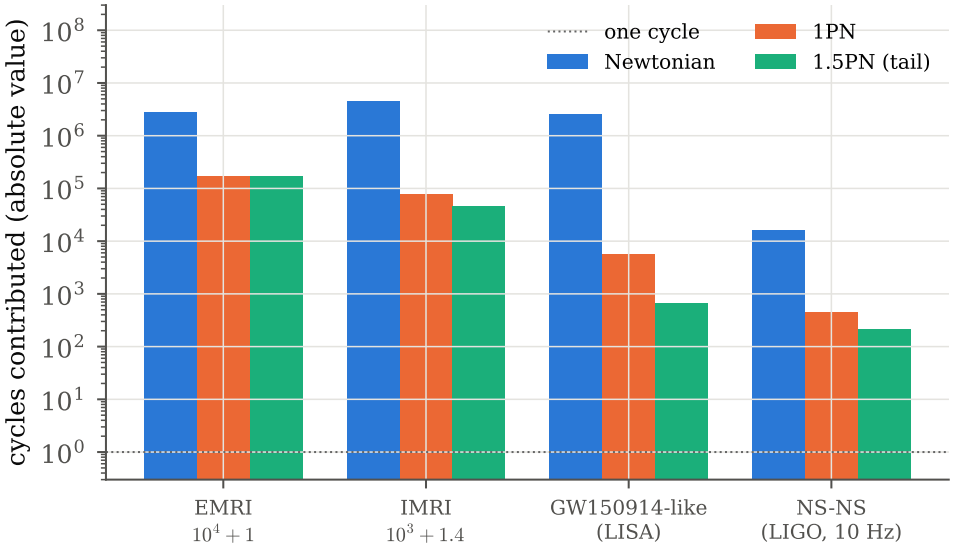


Figure T2.14: Number of wave cycles contributed by the Newtonian term and by the 1PN and 1.5PN (tail) corrections, for the four binaries named below the bars, each in its observation window. The dotted line marks one cycle, the slip that ruins a matched filter. Every correction is worth hundreds to hundreds of thousands of cycles.

Watch out: Post-Newtonian series for extreme mass ratios

For the EMRI the 1PN and 1.5PN counts are almost equal and opposite, each about six per cent of the total. Near the ISCO, $v \approx 0.4$, successive terms of the series shrink only slowly, so truncating it at 1.5PN leaves an error of many cycles. Research templates for extreme mass ratios therefore come from black-hole perturbation theory (the small body moving in the exact geometry of the large one, corrected by its own field) rather than from the post-Newtonian series; the power-law structure derived here is used to describe small *changes* to such a template, not the template itself.

T2.6.4 Every correction is a power of the frequency: the ppE form

The rule (T2.62) applies to any small correction that scales as a power of v, whatever its origin. In terms of the reduced frequency u of (T2.58), $v^p = \eta^{-p/5} u^{p/3}$, so a correction of relative order v^p becomes

$$\delta\Psi = \beta\, u^b, \qquad b = \frac{p-5}{3}, \tag{T2.69}$$

with all the constants, including the powers of η, in β. For the general-relativistic terms $p = 2k$ at kPN order, so $b = (2k-5)/3$: -1 at 1PN, $-2/3$ at 1.5PN, $-1/3$ at 2PN. A term that grows *at low frequency* faster than the Newtonian $-\frac{5}{3}$ must come from a correction with $p < 0$, one that is relatively more important when the binary is wide and slow. Yunes and Pretorius proposed to test general relativity with exactly this family: the **parameterised post-Einsteinian** (ppE) waveform

$$\tilde{h}(f) = \tilde{h}_{\mathrm{GR}}(f)\,\left(1 + \alpha u^a\right) e^{-i\beta u^b}, \tag{T2.70}$$

in our sign convention, where (α, a) change the amplitude and (β, b) the phase, and general relativity is $\alpha = \beta = 0$ [91, eq. (1)]. (Yunes and Pretorius write the powers of $(\pi G\mathcal{M}f/c^3)^{1/3}$, so each of their exponents is three times ours.) The environments of Section T2.10 fit into the same family: a dark-matter spike adds a drag whose power relative to the radiated one falls as f^{-n}, which by (T2.69) with $p = -3n$ is $b = -\frac{5}{3} - n$, (T2.94).

Example: dipole radiation, $b = -7/3$. In some alternatives to general relativity, such as the Brans–Dicke theory, a compact body carries a scalar "charge" that is not proportional to its mass, and the binary emits dipole radiation as well. By the estimate at the end of Section T2.2, dipole power relative to quadrupole power scales as $(v/c)^{-2}$. So $\mathcal{F} = \mathcal{F}_{\mathrm{N}}(1 + Bv^{-2})$, the bracket of (T2.60) gains $-Bv^{-2}$, $p = -2$, and $b = (-2-5)/3 = -\frac{7}{3}$: Yunes and Pretorius's Brans–Dicke case [91]. A "-1PN" term, it is largest at the lowest frequencies, which is why a long, slow inspiral in the LISA band is a sensitive test of it.

Example: a massive graviton, $b = -1$. If the graviton had a mass m_g, waves of energy hf would travel at the group velocity $v_g = c\sqrt{1 - m_g^2c^4/(hf)^2} \approx c\,[1 - m_g^2c^4/(2h^2f^2)]$, and low frequencies would arrive late. Over a distance D the delay is $\Delta t(f) = (D/c)\, m_g^2c^4/(2h^2f^2) = Dc/(2\lambda_g^2 f^2)$, with $\lambda_g = h/(m_g c)$ the graviton's Compton wavelength. The arrival time of frequency f is what the slope of the Fourier phase measures, $\mathrm{d}\Psi/\mathrm{d}f = 2\pi t(f)$ by (T2.53), so

$$\delta\Psi(f) \overset{(1)}{=} 2\pi \int \Delta t(f)\,\mathrm{d}f \overset{(2)}{=} 2\pi\,\frac{Dc}{2\lambda_g^2}\int f^{-2}\,\mathrm{d}f \overset{(3)}{=} -\frac{\pi Dc}{\lambda_g^2}\,f^{-1} + \text{const}, \tag{T2.71}$$

(1) Integrate the extra delay once, as in (T2.54).
(2) Insert $\Delta t(f)$.
(3) $\int f^{-2}\mathrm{d}f = -f^{-1}$; the constant joins ϕ_c.

a term $\propto f^{-1}$, so $b = -1$, the ppE massive-graviton case [91]. It has the same exponent as the 1PN term, which is why a graviton mass and the mass ratio are partly degenerate in a fit, a correlation Section 10.11.4 meets in another guise.

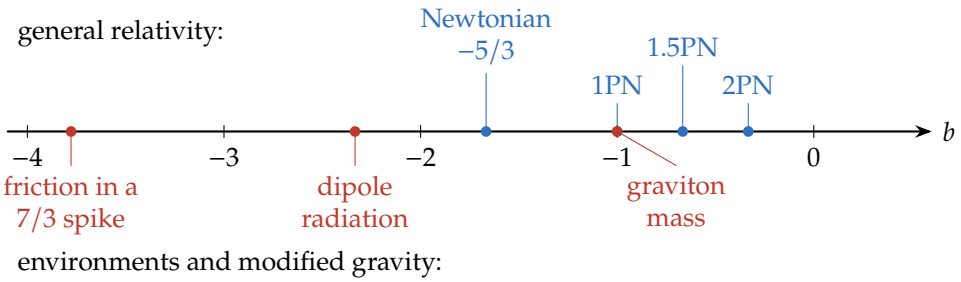


Figure T2.15: The exponent b of the term βu^b that each effect adds to the Fourier phase. General-relativistic corrections (blue, above the line) sit at $b = (2k - 5)/3$ and climb towards zero as the PN order k grows. Effects that matter most when the binary is wide and slow (red, below) have $b < -5/3$: dipole radiation at $-7/3$, dynamical friction in a dark-matter spike of slope $7/3$ at $-34/9$. A graviton mass lands on the same exponent as the 1PN term.

Key idea: A correction v^p becomes a phase term $f^{(p-5)/3}$

If the binding energy or the radiated power differs from its Newtonian value by a relative term $c_p v^p$, the Fourier phase gains $\delta\Psi = \frac{3}{128} u^{-5/3} \cdot \frac{40 c_p}{(8-p)(5-p)} v^p \propto u^{(p-5)/3}$. General relativity's kPN terms have $p = 2k$; environments and dipole radiation have $p < 0$. Any such effect is a ppE term βu^b, with the exponent b naming the physics and β its strength.

Where this is used. The phase (T2.67) is the signal model of Section 10.9.2, and the exponents of its terms are what make its Fisher matrix in Section 10.9.3 a set of weighted averages of powers of f. The rule (T2.69) is the general form behind Section 10.10.1, where an environment is a power law in the phase, and behind Section 10.11.3, which computes the error on β for every exponent b at once; the specific exponents of dark-matter environments are derived in Section T2.10.

T2.7 What LISA measures, and what hides the signal

LISA will be three spacecraft flying in a triangle $L = 2.5$ million kilometres on a side, trailing the Earth around the Sun (Figure T2.8). Each spacecraft shields test masses that fall freely, and laser links measure the distance between test masses on different spacecraft. By (T2.12) a gravitational wave of strain h changes an arm's length by about $\frac{1}{2}hL$. For $h \sim 10^{-20}$ that is about 10^{-11} m, ten picometres over millions of kilometres. The whole instrument is designed around measuring such changes, and its noise is the floor under which a signal can hide.

For the statistics, LISA is summarised by a single function. The data are the signal plus noise, $d = h(\boldsymbol{\theta}) + n$. For a given $\boldsymbol{\theta}$, the likelihood of Section 10.8.2 asks how plausible the particular noise realisation $n = d - h(\boldsymbol{\theta})$ is, and for stationary Gaussian noise that plausibility is set, one Fourier component at a time, by the noise power at that frequency. So the likelihood weights each frequency by the inverse of the noise power there, and the Fisher matrix of Chapter 10 inherits that weight. The question is therefore physical: *at each frequency, how large is the noise, where does it come from, and how large would a gravitational wave have to be to produce the same output?* The answer is the sensitivity curve $S_n(f)$ [85].

T2.7.1 A noise spectrum as a physical quantity

The statistics of a noise spectrum, how it describes the correlations of a stationary random process and how it is estimated from data, is developed in Sections 10.8.2 and 10.8.5. Here we need only what it means physically. Pass the noise $n(t)$ through a narrow filter that keeps only frequencies between f and $f + \Delta f$, and measure the mean square of what comes out. The **one-sided power spectral density** $S_n(f)$ is that mean square divided by Δf:

$$\left\langle n^2 \right\rangle_{[f,f+\Delta f]} = S_n(f)\,\Delta f, \qquad \left\langle n^2 \right\rangle = \int_0^\infty S_n(f)\,\mathrm{d}f. \tag{T2.72}$$

"One-sided" means that only positive frequencies are counted, the negative ones having been folded onto them as in Section T2.5. Its square root, the **amplitude spectral density** $\sqrt{S_n}$, is in units of the measured quantity per $\sqrt{\text{Hz}}$: a displacement noise of $15\,\text{pm}\,\text{Hz}^{-1/2}$ means that, averaged over one second (a band of about 1 Hz), the length reading scatters by about 15 pm. Two noises from independent causes add in power, $S = S_1 + S_2$, because the mean square of a sum of uncorrelated terms is the sum of their mean squares. And if a noise in one quantity is converted into another by a linear, frequency-dependent factor, as a force becomes a displacement, the spectrum is multiplied by the squared modulus of that factor.

T2.7.2 From test-mass jitter and photon counting to strain sensitivity

The two instrumental noises. Two processes limit the arm-length measurement [85, eqs. (10)–(11)]:

$$\begin{aligned} P_{\text{OMS}}(f) &= (1.5\times10^{-11}\,\text{m})^2\left[1+\left(\frac{2\,\text{mHz}}{f}\right)^4\right]\text{Hz}^{-1}, \\ P_{\text{acc}}(f) &= (3\times10^{-15}\,\text{m}\,\text{s}^{-2})^2\left[1+\left(\frac{0.4\,\text{mHz}}{f}\right)^2\right]\left[1+\left(\frac{f}{8\,\text{mHz}}\right)^4\right]\text{Hz}^{-1}. \end{aligned} \tag{T2.73}$$

The first is **optical metrology** noise, the error with which the laser interferometry reads the length of one link: about $15\,\text{pm}\,\text{Hz}^{-1/2}$, flat in frequency apart from a rise below 2 mHz. The second is **acceleration noise**: stray forces on the free-falling test masses, from impacts of residual gas molecules, electrostatic and magnetic forces on the slightly charged and magnetised masses, and the pressure of thermal radiation from their housing, push each of them around by about $3\times10^{-15}\,\text{m}\,\text{s}^{-2}\,\text{Hz}^{-1/2}$.

Example: where 15 picometres comes from. Most of the metrology noise is photon *shot noise*, the graininess of light, and its size follows from counting photons. A laser of about 2 W leaves one spacecraft through a telescope of diameter $D \approx 0.3\,\text{m}$. Diffraction spreads the beam by an angle $\theta \approx \lambda/D = 1.064\,\mu\text{m}/0.3\,\text{m} = 3.5\times10^{-6}$, so after $L = 2.5\times10^9\,\text{m}$ the spot has a radius of about $\theta L \approx 9\,\text{km}$, and the receiving telescope, of radius 0.15 m, catches a fraction of order $(0.15/9000)^2 \approx 3\times10^{-10}$ of it: a received power $P_r \approx 6\times10^{-10}\,\text{W}$. Each photon carries $hc/\lambda = 1.9\times10^{-19}\,\text{J}$, so about $\dot N = P_r\lambda/hc \approx 3\times10^9$ photons arrive per second. A phase measured by counting N photons is uncertain by $1/\sqrt{N}$ radians, the Poisson fluctuation of the count (Chapter 2), and a phase $\delta\varphi$ is a length $\lambda\,\delta\varphi/2\pi$. Per second of averaging,

$$\delta x \approx \frac{\lambda}{2\pi}\frac{1}{\sqrt{\dot N\times 1\,\text{s}}} = \frac{1.06\times10^{-6}\,\text{m}}{2\pi}\times\frac{1}{\sqrt{3\times10^9}} \approx 3\times10^{-12}\,\text{m},$$

a few picometres per $\sqrt{\text{Hz}}$, the same order as the $15\,\text{pm}\,\text{Hz}^{-1/2}$ allocated to all the metrology noise together. Counting statistics does not depend on how fast the signal varies, so shot noise is white in displacement. The arm length helps the signal ($\delta L \propto L$) but hurts the light budget ($P_r \propto L^{-2}$); 2.5 million kilometres is a compromise between the two.

Why acceleration noise is divided by $(2\pi f)^4$. A force noise becomes a displacement noise by integrating twice in time, and integration in time is division by frequency:

$$\ddot{x}(t) = a(t) \quad \Rightarrow \quad (2\pi i f)^2 \tilde{x}(f) \overset{(1)}{=} \tilde{a}(f) \quad \Rightarrow \quad P_x(f) \overset{(2)}{=} \frac{P_a(f)}{(2\pi f)^4}. \tag{T2.74}$$

(1) Fourier transform with the book's convention (T2.48): each time derivative becomes a factor $2\pi i f$.
(2) The spectrum is multiplied by the squared modulus of the conversion factor; $|(2\pi i f)^2|^2 = (2\pi f)^4$.

A random push of given strength moves a mass far if it keeps pushing the same way for a long time, that is at low frequency, and hardly at all if it reverses quickly. So acceleration noise is harmless at high frequency and dominates at low frequency, rising as f^{-4} in power.

Adding up the noise in one Michelson channel. Follow one arm, with test mass A at the vertex spacecraft and test mass B at the far one, and let $x_A(t)$, $x_B(t)$ be their random displacements along the arm (positive in the direction from A to B). Light leaves A at $t - 2L/c$, is sent back from B at $t - L/c$ and returns to A at t. The measured round-trip length changes by

$$\begin{aligned}
\delta\ell(t) &\overset{(1)}{=} \big[x_B(t - L/c) - x_A(t - 2L/c)\big] + \big[x_B(t - L/c) - x_A(t)\big], \\
\tilde{\delta\ell}(f) &\overset{(2)}{=} 2e^{-2\pi i f L/c}\, \tilde{x}_B - \big(1 + e^{-4\pi i f L/c}\big)\tilde{x}_A, \\
P_{\delta\ell} &\overset{(3)}{=} 4P_x + \big|1 + e^{-4\pi i f L/c}\big|^2 P_x \overset{(4)}{=} 4P_x + 4\cos^2\!\Big(\frac{2\pi f L}{c}\Big)P_x = 4\Big(1 + \cos^2\frac{f}{f_*}\Big)P_x.
\end{aligned} \tag{T2.75}$$

(1) The outward leg is lengthened by B's displacement when the light arrives and shortened by A's when it left; the return leg likewise, with A's displacement at the end.
(2) A delay by T multiplies a Fourier transform by $e^{-2\pi i f T}$, as in a time shift (Section 10.8.4).
(3) A and B jitter independently with the same spectrum P_x, so their contributions add in power; $|2e^{-i\cdot}|^2 = 4$.
(4) $|1 + e^{-i\theta}|^2 = 2 + 2\cos\theta = 4\cos^2(\theta/2)$ with $\theta = 4\pi f L/c$, and $2\pi f L/c = f/f_*$.

The far mass counts four times because the light touches it twice at the same moment; the vertex mass counts at two moments $2L/c$ apart, which add in step at low frequency and cancel when the round trip is half a period. A Michelson channel is the difference of two such arms, and in LISA each arm has its own test mass at the vertex, so the acceleration noise of the channel is $8(1 + \cos^2 f/f_*)P_{\rm acc}/(2\pi f)^4$: sixteen units at low frequency. The metrology noise enters once per one-way link, two links per arm, two arms: $4P_{\rm OMS}$. The signal in the same channel is the round-trip difference $2L\, D_{ij}h_{ij}$ of Section T2.3, so dividing the noise by $(2L)^2$ expresses it as an equivalent strain:

$$P_n(f) = \frac{4P_{\rm OMS} + 8\big(1 + \cos^2\frac{f}{f_*}\big)\frac{P_{\rm acc}}{(2\pi f)^4}}{4L^2} = \frac{1}{L^2}\Big[P_{\rm OMS} + 2\Big(1 + \cos^2\frac{f}{f_*}\Big)\frac{P_{\rm acc}}{(2\pi f)^4}\Big], \tag{T2.76}$$

which is Robson et al.'s eq. (12), with its "four contributions from the optical metrology noise and sixteen from the test mass acceleration noise" [85].

The noise that is cancelled before it is counted. There is a third noise, far larger than both: the laser's own frequency jitter. Even a stabilised laser wanders in frequency by about one part in 10^{13} per $\sqrt{\rm Hz}$, and a Michelson interferometer is immune to that only if its two arms are equally long, so that both beams carry the same old laser light back to be compared. LISA's

arms differ by about one per cent, some 2.5×10^{7} m, so the jitter would appear as a length noise of order $10^{-13}\times2.5\times10^{7}\,\mathrm{m}\approx2.5\,\mu\mathrm{m}$ per $\sqrt{\mathrm{Hz}}$, some 10^{5} times the signal. LISA therefore records every one-way link separately and, on the ground, adds the records with time shifts chosen so that the laser noise cancels exactly, synthesising an equal-arm Michelson after the fact. This is **time-delay interferometry**. What survives it is the test-mass and metrology noise counted above.

From the noise in the channel to the sensitivity curve. P_n is the noise of one channel expressed as a strain, but a real wave is not optimally oriented: on average over the sky and polarisation the channels respond to only a fraction $R(f)$ of the wave's power, $\frac{3}{10}$ at low frequency for the two channels together and falling above f_* (Section T2.3). It is conventional to move that factor from the signal onto the noise [85, eq. (2)]: the **sensitivity curve** is the noise divided by the response. With the fit $R(f)=\frac{3}{10}/[1+0.6(f/f_*)^2]$, checked in Figure T2.9,

$$S_n(f)=\frac{P_n(f)}{R(f)}=\frac{10}{3L^2}\left[P_{\mathrm{OMS}}(f)+2\left(1+\cos^2\frac{f}{f_*}\right)\frac{P_{\mathrm{acc}}(f)}{(2\pi f)^4}\right]\left[1+\frac{6}{10}\left(\frac{f}{f_*}\right)^2\right]+S_c(f). \tag{T2.77}$$

At $f\ll f_*$ the factor $2(1+\cos^2)$ is 4, the simpler form of Robson et al.'s eq. (1) [85]; the code uses the full factor.

The Galactic confusion noise. The last term in (T2.77) is not instrumental. Our Galaxy contains tens of millions of binaries of white dwarfs radiating at millihertz frequencies. A four-year observation divides the frequency axis into bins of width $1/T\approx8\,\mathrm{nHz}$, and below a few millihertz there are more binaries than bins. They cannot be separated from one another, and their sum acts as an extra noise [85, Sec. 1]:

$$S_c(f)=A\,f^{-7/3}\,e^{-f^{\alpha}+\beta f\sin(\kappa f)}\left[1+\tanh(\gamma(f_k-f))\right]\mathrm{Hz}^{-1}, \tag{T2.78}$$

with $A=9\times10^{-45}$ and, for four years, $(\alpha,\beta,\kappa,\gamma,f_k)=(0.138,-221,521,1680,1.13\,\mathrm{mHz})$ [85, eq. (14)].

Why $f^{-7/3}$? The power law follows from the chirp. Suppose white-dwarf binaries form at a steady rate at low frequency and then drift upwards in frequency at the rate $\dot f\propto f^{11/3}$ of (T2.43). In a steady state the number of binaries crossing any frequency per unit time is the same at every f, like a steady flow of traffic: $(\mathrm{d}N/\mathrm{d}f)\,\dot f=\text{const}$, so $\mathrm{d}N/\mathrm{d}f\propto f^{-11/3}$. Where binaries move quickly they are few. Each binary contributes a mean-square strain $\propto A^2\propto f^{4/3}$, by (T2.26). Summing the power of all binaries in a band, as (T2.72) sums the power of a noise,

$$S_c(f)\propto\frac{\mathrm{d}N}{\mathrm{d}f}\,A^2\propto f^{-11/3}\,f^{4/3}=f^{-7/3}. \tag{T2.79}$$

The other factors of the fit describe where this simple picture fails. Above the knee f_k there is less than about one binary per bin, so the binaries can be fitted and subtracted one by one, and the tanh switches the confusion off. The longer LISA observes, the finer the bins, the more binaries are resolved, and the lower the knee moves (it is 2.58 mHz after six months). The confusion noise is also not stationary: it rises and falls during the year as the antenna pattern of Section T2.3 sweeps across the Galactic centre. The fit is its average.

Working definition T2.3 (sensitivity curve)

The **sensitivity curve** $S_n(f)$ is the one-sided noise power spectral density of the detector divided by its sky- and polarisation-averaged response, plus any unresolved foreground [the noise expressed as the strain power a gravitational wave would need to produce the

same output]. *Operationally:* in the inner product $(a\,|\,b) = 4\,\mathrm{Re}\int \tilde{a}\tilde{b}^*/S_n\,\mathrm{d}f$ use $S_n(f)$ of (T2.77) together with the angle-averaged waveform $\tilde{h} = \sqrt{4/5}\,\mathcal{A} f^{-7/6} e^{-i\Psi}$.

Figure T2.16 shows the three pieces. Acceleration noise dominates below 3.077 mHz (by hand: setting $P_{\mathrm{OMS}} = 4P_{\mathrm{acc}}/(2\pi f)^4$ without the correction factors gives $(2\pi f)^4 = 4\times 9\times 10^{-30}/2.25\times10^{-22} = 1.6\times10^{-7}\,\mathrm{s}^{-4}$, i.e. $f = 3.2$ mHz). The confusion noise exceeds the instrument between 0.4101 and 1.908 mHz. Above $f_* = 19.09$ mHz the falling response makes the curve rise again. The best sensitivity, $\sqrt{S_n} = 1.178\times10^{-20}\,\mathrm{Hz}^{-1/2}$, is reached near 7.256 mHz.

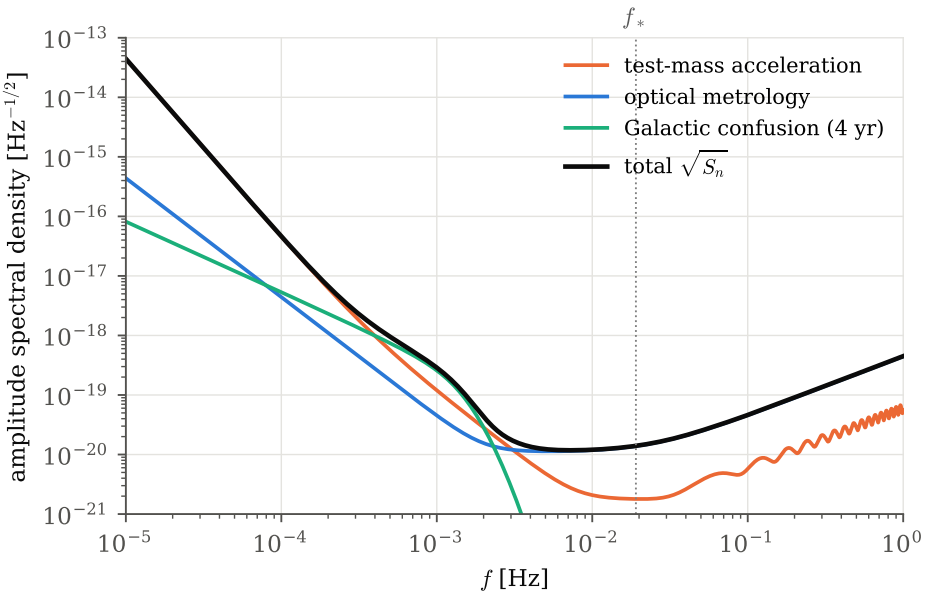


Figure T2.16: The LISA sensitivity curve and its three ingredients, as amplitude spectral densities $\sqrt{S}$. Test-mass acceleration noise falls steeply with frequency, optical metrology noise is flat and then rises above f_* (dotted) where the arms respond less, and the Galactic confusion noise fills the gap around a millihertz. The thick curve is the total used throughout the book.

The curve is a short function in `lib_lisa.py`:

Listing T2.5: The three pieces of $S_n(f)$ and their sum (`code/chT2/lib_lisa.py`)

```
100 def lisa_components(f):
101     """The three contributions to S_n(f) [1/Hz]: metrology, acceleration, confusion."""
102     resp = 10.0 / (3.0 * L_ARM**2) * (1 + 0.6 * (f / F_STAR) ** 2)
103     oms = resp * P_oms(f)
104     acc = resp * 2 * (1 + np.cos(f / F_STAR) ** 2) * P_acc(f) / (2 * np.pi * f) ** 4
105     return oms, acc, S_conf(f)
106
107
108 def Sn_lisa(f, confusion=True):
109     """Sky-averaged LISA S_n(f) [1/Hz]: instrument (their eq. 13) + confusion (eq. 14)."""
110     oms, acc, conf = lisa_components(f)
111     return oms + acc + (conf if confusion else 0.0)
```

The factor `resp` on line 102 is $\frac{10}{3L^2}[1 + 0.6(f/f_*)^2]$, the conversion from displacement noise to strain sensitivity. Line 103 is the metrology term of (T2.77), line 104 the acceleration term with its f^{-4} and the full factor $2(1+\cos^2)$, and line 105 hands back the confusion fit. `Sn_lisa` adds the three; with `confusion=False` it is the instrument alone.

T2.7.3 How loud is a source?

With the noise in hand, the signal-to-noise ratio of Chapter 0 is $\rho^2 = (h\,|\,h) = 4\int |\tilde h|^2/S_n\,\mathrm{d}f$. For the angle-averaged waveform of Section T2.5, $|\tilde h|^2 = \frac{4}{5}\mathcal{A}^2 f^{-7/3}$ (with $\mathcal{A}$ the amplitude of (T2.56)), and it is useful to rewrite the integral so that a plot shows it:

$$\rho^2 \overset{(1)}{=} \frac{16}{5}\int \frac{\mathcal{A}^2 f^{-7/3}}{S_n(f)}\,\mathrm{d}f \overset{(2)}{=} \int \left(\frac{h_c(f)}{h_n(f)}\right)^2 \mathrm{d}\ln f, \qquad h_c \equiv 2f|\tilde h(f)|, \quad h_n \equiv \sqrt{f S_n(f)}. \tag{T2.80}$$

(1) Insert $|\tilde h|^2 = \frac{4}{5}\mathcal{A}^2 f^{-7/3}$; $4\cdot\frac{4}{5} = \frac{16}{5}$ [85, eq. (18)].
(2) $4|\tilde h|^2\mathrm{d}f/S_n = (2f|\tilde h|)^2/(fS_n)\cdot \mathrm{d}f/f$, and $\mathrm{d}f/f = \mathrm{d}\ln f$.

Both h_c, the **characteristic strain**, and h_n are dimensionless. On log–log axes the height of a source's track above the noise curve shows how much signal-to-noise ratio each decade of frequency contributes. A track below the curve can still be detected if it is long enough, because the integral adds up many frequencies.

Example: the three benchmark sources. The script `03_lisa_noise.py` evaluates (T2.80) on the four-year stretch of each binary of Section T2.4. Kim et al. place the EMRI at $D_L = 203\,\mathrm{Mpc}$, where five years of data give a signal-to-noise ratio of about 15 [86, Sec. III F]. For our four years we place it where $\rho = 14$, which is $D_L = 199.8\,\mathrm{Mpc}$, close to their distance. The IMRI at 75 Mpc, the distance out to which Coogan et al. find dark-matter environments measurable [89], reaches $\rho = 12.1$ between 21.7 mHz and the 1 Hz end of the band. The GW150914-like binary at 410 Mpc reaches only $\rho = 3.89$: on its own LISA would not claim it, which is why such sources are studied together with ground-based detectors. Figure T2.17 shows the tracks. Since $h_c \propto f\cdot f^{-7/6} = f^{-1/6}$, each inspiral is an almost flat line that ends where the binary reaches its ISCO or leaves the band.

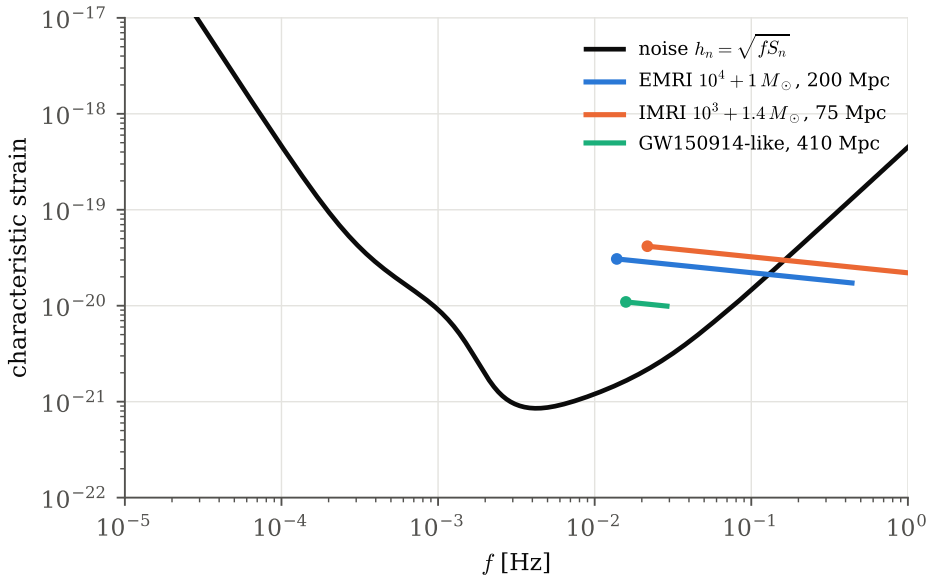


Figure T2.17: Characteristic strain of the three benchmark binaries during four years of observation (dots mark the start), over the LISA noise amplitude $h_n = \sqrt{fS_n}$. The squared ratio of the two curves, integrated over $\ln f$, is the squared signal-to-noise ratio (T2.80). The EMRI and the IMRI sit well above the noise over a decade or more of frequency; the GW150914-like binary is short and sits only several times above it.

Watch out: Three conventions for one curve

The literature plots $\sqrt{P_n}$ (noise only), $\sqrt{S_n} = \sqrt{P_n/R}$ (noise over response) and $h_n = \sqrt{fS_n}$ (dimensionless), and quotes the low-frequency response as 3/10 (two channels) or 3/20 (one) [85, Sec. 1]. Each is correct with the matching signal. Pair S_n with the angle-averaged waveform $\sqrt{4/5}\,\mathcal{A}f^{-7/6}$, or P_n with the full detector response, and never mix them: a mismatch changes every signal-to-noise ratio and Fisher error by a constant factor that nothing else in the calculation will reveal.

Where this is used. The curve (T2.77) is the $S_n(f)$ of the noise-weighted inner product of Section 10.8.2, and with it of every likelihood, matched filter and Fisher matrix in Chapter 10; Section 10.11.1 checks it against the published figure, and Section 10.11.2 uses the signal-to-noise ratio (T2.80) to place the benchmark source. In real data S_n is not known in advance but estimated, which is the subject of Sections 10.8.5 and 10.8.6; the non-stationary confusion noise is one reason why that estimate is hard.

T2.8 The symbols, where they come from, and where they are used

Every gravitational-wave object that Chapter 10 takes as given has a short derivation above. The table that follows gives each symbol in a few words, the equation or section where it was derived, and the part of Chapter 10 in which it does its work. The rows follow the chain of Figure T2.1: the wave, the source, the chirp, the frequency domain, the corrections, the detector and its noise.

The rows are links of one chain of reasoning. Einstein's equation for a weak field gave a wave with two polarisations that changes distances by $\frac{1}{2}h$. Mass and momentum conservation made the quadrupole the first source of such waves, and the energy they carry made a binary chirp. The stationary-phase approximation turned the chirp into a Fourier-domain template whose phase is a sum of powers of f, two of whose terms, $2\pi f t_c - \phi_c$, are nuisances, and whose other terms, Newtonian, post-Newtonian or environmental, carry the physics. The detector added its geometry, through F_+, $F_\times$ and the averages $\frac{4}{5}$ and $\frac{3}{10}$, and its noise, through $S_n(f)$. Nothing in the list is a convention to be taken on trust; each entry can be recomputed from the two physical inputs of Section T2.1 and Newton's law of gravity.

Recap

- A weak gravitational wave is a transverse, traceless ripple h_{ij} moving at c, with two polarisations h_+, $h_\times$ that mix under rotations with twice the angle; it changes the distance between free masses by $\delta L/L = \frac{1}{2}h_{ij}\hat{n}_i\hat{n}_j$, a travelling tidal squeeze.
- Conservation of mass and momentum forbids monopole and dipole radiation, so slow sources radiate through $\ddot{Q}_{ij}$: $h_{ij}^{\mathrm{TT}} = (2G/c^4 r)[\ddot{Q}_{ij}]^{\mathrm{TT}}$ and $P = (G/5c^5)\langle\dddot{Q}\dddot{Q}\rangle$. A circular binary radiates at $f = 2f_{\mathrm{orb}}$ with amplitude $4(G\mathcal{M})^{5/3}(\pi f)^{2/3}/(c^4 D)$ and power $\frac{32}{5}(G/c^5)\mu^2 r^4\omega^6$.
- A detector outputs $h = F_+h_+ + F_\times h_\times$. Averages: $\frac{1}{5}$ over sky and polarisation, $\frac{3}{4}$ for 60° arms, $\frac{3}{10}$ for LISA's two channels, $\frac{4}{5}$ over inclination. Above $f_* = c/(2\pi L) = 19\,\mathrm{mHz}$ the response falls as f^{-2}.
- Energy balance gives the chirp $\dot{f} \propto \mathcal{M}^{5/3}f^{11/3}$; integrating it gives the time to merger, the number of cycles and the phase, with the nuisance constants t_c and ϕ_c.

symbol	meaning	derived in	used in Chapter 10
h_+, $h_\times$	the two polarisations of a wave, in the transverse plane	(T2.8)	§10.8
$\delta L/L = \frac{1}{2}h_{aa}$	fractional change of an arm's length, read by light travel time	(T2.12), (T2.30)	§10.8
F_+, $F_\times$; D_{ij}	antenna patterns: weights of the polarisations in the detector output; detector tensor	(T2.32), (T2.34)	§10.8
$\frac{1}{5}$, $\frac{3}{20}$, $R = \frac{3}{10}$	sky- and polarisation-averaged response: right-angle, one 60° channel, two LISA channels	(T2.35), (T2.37)	§10.11.1
$\frac{4}{5}$	inclination average of a circular binary's signal power	(T2.36)	§10.11.1, §10.11.2
$f_* = c/2\pi L$	transfer frequency; response falls as f^{-2} above it	(T2.38)	§10.11.1
M, μ, η	total mass, reduced mass, symmetric mass ratio μ/M	§T2.2, (T2.57)	§10.9.2
$\mathcal{M}$	chirp mass $\mu^{3/5}M^{2/5}$, the mass the Newtonian signal depends on	(T2.26), (T2.42)	§10.9.2, §10.9.4
$f = 2f_{\rm orb}$	wave frequency, twice the orbital one	(T2.24)	throughout
A; $\mathcal{A}$	time-domain amplitude; Fourier amplitude of $\tilde{h} = \mathcal{A}f^{-7/6}e^{-i\Psi}$	(T2.26), (T2.56)	§10.8.3, §10.9.2
ι, ψ	inclination of the orbit; polarisation angle	(T2.25), (T2.34)	§10.11.1
$P_{\rm GW}$, $E(f)$	radiated power; orbital binding energy	(T2.27), (T2.41)	§10.10.1
$\dot{f}$	chirp rate $\frac{96}{5}\pi^{8/3}(G\mathcal{M}/c^3)^{5/3}f^{11/3}$	(T2.43)	§10.10.1
$\tau(f)$, N, $f_{\rm ISCO}$	time left to coalescence; number of cycles; end of the inspiral	(T2.44), (T2.45), (T2.46)	§10.11.2
t_c, ϕ_c	time and phase of coalescence: nuisance parameters	(T2.47), §T2.4.3	§10.8.4, §10.9.4
$\tilde{h}(f)$	Fourier transform, $\int h\, e^{-2\pi i f t}\,\mathrm{d}t$; $\tilde{h}(-f) = \tilde{h}(f)^*$	(T2.48), (T2.49)	§10.8.2
$\Psi(f)$	Fourier phase of the chirp, from the stationary-phase approximation	(T2.51), (T2.54)	§10.9.2
$\Psi' = 2\pi t(f)$, $\Psi'' = 2\pi/\dot{f}$	what the phase knows: arrival time and chirp rate	(T2.53)	§10.10.1
$\Phi(f)$	time-domain phase still to come; Ψ carries $\frac{3}{8}$ of it	§T2.5 (example)	§10.12.1
v, x, u	PN parameter $(\pi GMf/c^3)^{1/3}$, $x = v^2$, reduced frequency $\pi G\mathcal{M}f/c^3$	(T2.57), (T2.58)	§10.9.2
1PN, 1.5PN terms	corrections $\frac{20}{9}(\frac{743}{336}+\frac{11}{4}\eta)x$ and $-16\pi x^{3/2}$	(T2.67)	§10.9.2
β, b	ppE amplitude and exponent of a phase term βu^b; $b = (p-5)/3$	(T2.69), (T2.94)	§10.10, §10.11.3
$S_n(f)$	one-sided noise spectrum divided by the response: the sensitivity curve	(T2.72), (T2.77)	§10.8.2, §10.11.1
$P_{\rm OMS}$, $P_{\rm acc}$, S_c	metrology, test-mass acceleration and Galactic confusion noise	(T2.73), (T2.76)	§10.11.1
h_c, h_n, ρ	characteristic strain, noise amplitude, signal-to-noise ratio	(T2.80)	§10.11.2

Table T2.1: The gravitational-wave symbols of Chapter 10: what each one is, the equation or section that derives it, and the section of Chapter 10 that uses it.

- The stationary-phase approximation, a Taylor expansion of the phase plus a Fresnel integral, gives $\tilde{h} = \mathcal{A} f^{-7/6} e^{-i\Psi}$ with $\Psi' = 2\pi t(f)$ and $\Psi'' = 2\pi/\dot{f}$, checked against an FFT.
- A relative correction v^p to the energy or flux adds $\propto f^{(p-5)/3}$ to Ψ: the 1PN and 1.5PN terms of the standard template, dipole radiation at $b = -7/3$, a graviton mass at $b = -1$, environments at $b = -5/3 - n$, all of the ppE form βu^b.
- LISA's noise is shot noise (flat in displacement) and test-mass acceleration noise (divided by $(2\pi f)^4$), counted through the round trips of a Michelson channel and divided by the response to give $S_n(f)$, plus the $f^{-7/3}$ confusion noise of unresolved Galactic binaries.

T2.9 Dark matter around black holes

The signal model of Sections T2.5 and T2.6 describes a binary in empty space: a chirp whose Fourier phase $\Psi(f)$ (T2.54) is fixed by the chirp mass $\mathcal{M}$ (T2.42) and the coalescence constants t_c, ϕ_c (Section T2.4.3), and to which any extra physics adds a power of the frequency, βu^b (T2.69). This section and the next supply the extra physics that the forecasts of Chapter 10 test. Here we ask where dark matter around a black hole comes from and how dense it becomes; Section T2.10 finds the forces it exerts on the orbit and the exponent b they produce.

Near the Sun, dark matter has a density of about $0.01\,M_\odot\,\mathrm{pc}^{-3}$. Spread at that density over a sphere the size of the EMRI orbit of Section T2.4, about 30 Schwarzschild radii of a $10^4\,M_\odot$ hole or $3\times10^{-8}\,\mathrm{pc}$, it amounts to $\frac{4}{3}\pi(3\times10^{-8})^3\times0.01 \approx 10^{-24}\,M_\odot$. No orbit could feel that. A survey of environmental effects on inspirals concludes exactly this: for most sources, galactic dark matter at ordinary densities is irrelevant [92, Sec. IX]. If a gravitational wave is to show dark matter, something must have packed it around the black hole at densities some eighteen orders of magnitude higher. The black hole itself can do it. So the question is: *how does a black hole that grows at the centre of a dark-matter halo rearrange the dark matter around it, and how dense does it become where the companion orbits?*

T2.9.1 The halo far away

Far from any black hole, simulations of cold dark matter find the **Navarro–Frenk–White** (NFW) profile

$$\rho_{\mathrm{NFW}}(r) = \frac{\rho_s}{(r/r_s)(1+r/r_s)^2}, \tag{T2.81}$$

which falls as r^{-1} inside the scale radius r_s and as r^{-3} outside it [93], [94, eq. (1)]. The inner r^{-1} is called a **cusp**. A halo has no sharp edge, so it is labelled by the mass inside a conventional one. The reference density is the **critical density** $\rho_{\mathrm{crit}} = 3H_0^2/(8\pi G)$, the mean density at which a universe expanding at today's rate H_0 would be spatially flat, about $1.3\times10^{-7}\,M_\odot\,\mathrm{pc}^{-3}$ for $h = 0.674$. The edge r_{200} is the radius within which the mean density of the halo is $200\rho_{\mathrm{crit}}$, and M_{200} is the mass inside it. Since that mass is $200\rho_{\mathrm{crit}}$ times the volume $\frac{4}{3}\pi r_{200}^3$, the mass fixes the edge, $r_{200} = [3M_{200}/(800\pi\rho_{\mathrm{crit}})]^{1/3}$. The second label is the concentration $c = r_{200}/r_s$, which then fixes $r_s = r_{200}/c$. What remains is ρ_s, and it follows from the enclosed mass:

$$M(<r) \overset{(1)}{=} 4\pi\rho_s r_s^3 \int_0^{r/r_s} \frac{x\,\mathrm{d}x}{(1+x)^2} \overset{(2)}{=} 4\pi\rho_s r_s^3 \Big[\ln\Big(1+\frac{r}{r_s}\Big) - \frac{r/r_s}{1+r/r_s}\Big],$$
$$\rho_s \overset{(3)}{=} \rho_{\mathrm{crit}}\,\frac{200}{3}\,\frac{c^3}{\ln(1+c) - c/(1+c)} \equiv \rho_{\mathrm{crit}}\,\delta_c. \tag{T2.82}$$

(1) $M = \int_0^r 4\pi r'^2 \rho \, \mathrm{d}r'$ with $x = r'/r_s$; the factor x^2/x leaves x.
(2) $\frac{x}{(1+x)^2} = \frac{1}{1+x} - \frac{1}{(1+x)^2}$ integrates to $\ln(1+x) + \frac{1}{1+x} - 1 = \ln(1+x) - \frac{x}{1+x}$.
(3) Evaluate step (2) at $r = r_{200} = c r_s$ and set it equal to $200\rho_{\mathrm{crit}} \cdot \frac{4}{3}\pi r_{200}^3$: $4\pi\rho_s r_s^3 \left[\ln(1+c) - \frac{c}{1+c} \right] = \frac{800\pi}{3} \rho_{\mathrm{crit}} \, c^3 r_s^3$. The factor r_s^3 cancels, and dividing by $4\pi[\ldots]$ gives ρ_s [94, eq. (2)].

Take a dwarf-galaxy halo with $M_{200} = 10^8 \, M_\odot$. The concentration–mass relation of Correa et al. at redshift zero [95, eq. (19)] gives $c = 13.49$, and then $r_{200} = 9.82\,\mathrm{kpc}$, $r_s = 0.7279\,\mathrm{kpc}$, $\delta_c = 9.39 \times 10^4$ and $\rho_s = 0.01184 \, M_\odot \, \mathrm{pc}^{-3}$ (with $h = 0.674$). These are ordinary galactic numbers. The interesting physics is inside, where the black hole dominates.

T2.9.2 The spike: what slow growth of a black hole does

Suppose a black hole of mass M grows at the centre of the cusp slowly, over many orbital periods of the surrounding dark matter. Each dark-matter particle feels the potential deepen gradually and moves inward to a tighter orbit. Dark matter piles up around the hole. The simplest orbits, circles, already give the answer.

Inputs. (i) Physical: the growth is **adiabatic** [slow compared with the orbital period of the dark matter at the radius considered], so a particle on a circular orbit stays on a circular orbit and keeps its angular momentum, which is conserved in any spherical potential. (ii) Physical: orbits do not cross. Every circular orbit shrinks, but an orbit that started inside another stays inside it, so the shells keep their order, and the mass inside each orbit is the same before and after. (iii) Physical: before the growth the gravity at radius r_i comes from the dark matter inside it; afterwards the gravity at the new radius r_f comes from the hole alone. The last input needs a word. The compressed dark matter inside r_f weighs the same $M_i(r_i)$ it weighed before the growth, and the radii of interest are those where that is much less than M; there the hole's pull dominates and the dark matter's own pull is a small correction.

The mass inside the old orbit. For an initial cusp $\rho_i = \rho_0 (r/r_0)^{-\gamma_i}$ with $\gamma_i < 3$, adding up thin shells gives

$$M_i(r_i) = \int_0^{r_i} 4\pi r^2 \rho_0 \left(\frac{r}{r_0}\right)^{-\gamma_i} \mathrm{d}r = \frac{4\pi\rho_0 r_0^{\gamma_i}}{3-\gamma_i} \, r_i^{3-\gamma_i} \propto r_i^{3-\gamma_i}.$$

Same angular momentum, smaller orbit. Equating the angular momentum before and after, and then following the mass of a shell to its new radius,

$$\begin{aligned} G\,M_i(r_i)\,r_i &\overset{(1)}{=} G\,M\,r_f \quad\Rightarrow\quad r_f \overset{(2)}{=} \frac{M_i(r_i)\,r_i}{M} \propto r_i^{4-\gamma_i}, \\ \rho_f(r_f) &\overset{(3)}{=} \rho_i(r_i) \, \frac{r_i^2}{r_f^2} \, \frac{\mathrm{d}r_i}{\mathrm{d}r_f} \overset{(4)}{=} \frac{\rho_i(r_i)}{4-\gamma_i} \, \frac{r_i^3}{r_f^3} \propto r_i^{3-\gamma_i} \, r_f^{-3} \overset{(5)}{=} r_f^{\frac{3-\gamma_i}{4-\gamma_i}-3} = r_f^{-\frac{9-2\gamma_i}{4-\gamma_i}}. \end{aligned} \tag{T2.83}$$

(1) For a circular orbit $v^2 = GM_{\mathrm{enc}}/r$, so the squared angular momentum per unit mass is $L^2 = r^2 v^2 = GM_{\mathrm{enc}}\, r$. Before the growth $M_{\mathrm{enc}} = M_i(r_i)$; after it $M_{\mathrm{enc}} = M$ by input (iii). L^2 is the same by input (i).
(2) Solve for r_f and use $M_i(r_i) \propto r_i^{3-\gamma_i}$.
(3) By input (ii) the shell between r_i and $r_i + \mathrm{d}r_i$ becomes the shell between r_f and $r_f + \mathrm{d}r_f$ with the same mass: $4\pi r_i^2 \rho_i \, \mathrm{d}r_i = 4\pi r_f^2 \rho_f \, \mathrm{d}r_f$. Divide by $4\pi r_f^2 \mathrm{d}r_f$.
(4) From step (2), $\mathrm{d}r_f/\mathrm{d}r_i = (4-\gamma_i)\, r_f/r_i$, so $\mathrm{d}r_i/\mathrm{d}r_f = r_i/[(4-\gamma_i) r_f]$.
(5) Step (2) inverted, $r_i \propto r_f^{1/(4-\gamma_i)}$, and $\rho_i \propto r_i^{-\gamma_i}$. The exponent is $\frac{3-\gamma_i}{4-\gamma_i} - 3 = \frac{3-\gamma_i-12+3\gamma_i}{4-\gamma_i} = \frac{2\gamma_i-9}{4-\gamma_i}$.

Key idea: The dark-matter spike

A black hole that grows slowly inside a cusp $\rho \propto r^{-\gamma_i}$ compresses it into a **spike** $\rho_{\rm sp}(r) = \rho_{\rm sp}(r_{\rm sp}/r)^{\gamma_{\rm sp}}$ with $\gamma_{\rm sp} = (9-2\gamma_i)/(4-\gamma_i)$. An NFW cusp ($\gamma_i = 1$) gives $\gamma_{\rm sp} = 7/3$; the range $0 \le \gamma_i \le 2$ gives $2.25 \le \gamma_{\rm sp} \le 2.5$.

The circular-orbit argument is a sketch, but it gives the same exponent as the full calculation, which follows the whole phase-space distribution of the dark matter through the growth using all the adiabatic invariants [96], [86, eqs. (1)–(6)]. The full calculation changes two things that matter here.

The spike has a hole in the middle. General relativity adds one input to the Newtonian picture: a particle that falls in from far away with angular momentum per unit mass below $4GM/c$ has no turning point and crosses the horizon [97]. The circular orbit that carries exactly this angular momentum has radius $4GM/c^2 = 2r_h$ ($r_h = 2GM/c^2$), and no orbit is stable inside the innermost stable circular orbit (ISCO) at $6GM/c^2 = 3r_h$. A particle whose orbit dips below these radii is swallowed within one orbit, so the spike holds nothing there. The papers place the cutoff at $3r_h$ [82, Sec. II B], at $4GM/c^2$ [89], or smooth it with a factor $(1-2r_h/r)^{\gamma_{\rm sp}}$ [86, eq. (6)]. A fully relativistic calculation lets the spike reach closer to the hole and makes it denser there [97], [88, Sec. II].

The slope depends on how the dark matter moved before. Real orbits are not all circles. A particle on an eccentric orbit that started far out can spend part of its time deep inside, so the mass inside r_f collects contributions from many initial radii, and the one-to-one map $r_i \mapsto r_f$ of (T2.83) is lost. For an initially uniform core, Eda et al. quote the gentler $\rho \propto r^{-3/2}$ [82, Sec. II B]. The reason fits in one line. A core has nearly the same density of particles in phase space for all its slow particles, and slow growth does not change that phase-space density (Liouville's theorem with adiabatic invariants). After the growth, the particles bound at radius r have speeds up to the escape speed $v_{\rm esc} = \sqrt{2GM/r}$, so the density is that constant phase-space density times the volume $\frac{4}{3}\pi v_{\rm esc}^3$ of a ball in velocity space: $\rho \propto v_{\rm esc}^3 \propto r^{-3/2}$. The same exponent will appear for a wave falling freely into the hole, (T2.85) below, for a related reason: in both cases the speeds near the hole scale as $r^{-1/2}$. Finally, mergers with other black holes, or scattering off stars, can erase a spike altogether [89], which is why the slope $\gamma_{\rm sp}$ is treated as a parameter to measure rather than a prediction.

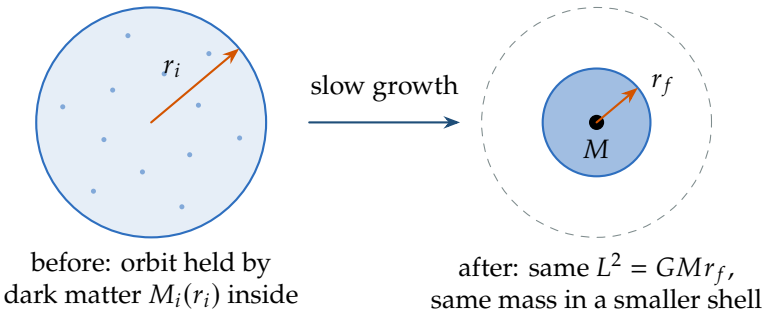


Figure T2.18: Adiabatic compression. Left: a dark-matter orbit of radius r_i in the original cusp, held by the dark matter inside it. Right: after the hole has grown, the same orbit, with the same angular momentum, sits at a smaller radius r_f (the dashed circle is the old one). The same mass now fills a smaller volume, so the density rises, more steeply the closer to the hole.

Fixing the normalisation. The slope says nothing about how high the spike is: (T2.83) is a proportionality. Eda et al. wanted a height fixed by the black-hole mass alone, so that a single number, M, would set the whole profile inside a given halo [82, Sec. II B]. They impose two

conditions. The first is continuity: the spike joins the NFW profile at the radius r_{sp} where it begins, so $\rho_{\mathrm{sp}} = \rho_{\mathrm{NFW}}(r_{\mathrm{sp}})$. The second needs the idea of a **radius of influence** r_{infl}, the radius out to which the black hole's gravity is comparable with that of the dark matter around it. The usual definition counts mass: r_{infl} is where the dark matter inside weighs twice the hole, $M_{\mathrm{DM}}(< r_{\mathrm{infl}}) = 2M$. Eda et al. then take the spike to occupy the inner fifth of that region, $r_{\mathrm{sp}} = 0.2\, r_{\mathrm{infl}}$ [82, eqs. (4)–(5)]. Both the factor 2 and the factor 0.2 are definitions, chosen to mark where the hole's influence ends; they are not consequences of a law, and a different choice moves the spike's height.

The condition in our notation. With $r_{\mathrm{infl}} = 5r_{\mathrm{sp}}$, the dark matter inside $5r_{\mathrm{sp}}$ is the spike from the inner cutoff r_{min} ($3r_h$ for Eda et al.) out to r_{sp}, plus the undisturbed NFW halo from r_{sp} to $5r_{\mathrm{sp}}$:

$$\underbrace{\frac{4\pi\rho_{\mathrm{sp}} r_{\mathrm{sp}}^{\gamma_{\mathrm{sp}}}}{3-\gamma_{\mathrm{sp}}}\left(r_{\mathrm{sp}}^{3-\gamma_{\mathrm{sp}}} - r_{\mathrm{min}}^{3-\gamma_{\mathrm{sp}}}\right)}_{\text{spike}} + \underbrace{M_{\mathrm{NFW}}(<5r_{\mathrm{sp}}) - M_{\mathrm{NFW}}(<r_{\mathrm{sp}})}_{\text{halo, from (T2.82)}} = 2M. \tag{T2.84}$$

The spike term is $\int_{r_{\mathrm{min}}}^{r_{\mathrm{sp}}} 4\pi r^2 \rho_{\mathrm{sp}} (r_{\mathrm{sp}}/r)^{\gamma_{\mathrm{sp}}} \mathrm{d}r$, done as for M_i above; the halo terms are step (2) of (T2.82) at the two radii. Since continuity fixes ρ_{sp} once r_{sp} is known, (T2.84) is one equation for one unknown, r_{sp}. The script `04_dm_profile.py` solves it by root finding:

Listing T2.6: The spike radius from the mass condition (`code/chT2/04_dm_profile.py`)

```
49 def mass_condition(rsp):
50     m_in = 4*np.pi * nfw(rsp) * rsp**gsp * (rsp**(3-gsp) - rmin**(3-gsp)) / (3-gsp)
51     return m_in + (M_nfw(5 * rsp) - M_nfw(rsp)) - 2 * M
52 rsp = brentq(mass_condition, 1e-4, 1e3)
53 rho_sp = nfw(rsp)
```

Line 50 is the spike term of (T2.84), with ρ_{sp} replaced by `nfw(rsp)` (continuity). Line 51 adds the halo term and subtracts $2M$, so the root of `mass_condition` is the r_{sp} we want. For the $10^4\, M_\odot$ hole in the $10^8\, M_\odot$ halo the root is $r_{\mathrm{sp}} = 3.759\,\mathrm{pc}$, and then $\rho_{\mathrm{sp}} = 2.269\, M_\odot\,\mathrm{pc}^{-3}$.

Coogan et al.'s shortcut. Coogan et al. keep $r_{\mathrm{sp}} = 0.2\, r_{\mathrm{infl}}$ and $M_{\mathrm{DM}}(< r_{\mathrm{infl}}) = 2M$ but count the mass with the spike power law alone, from the centre, and treat ρ_{sp} as a free parameter instead of tying it to a halo:

$$2M \overset{(1)}{=} \frac{4\pi\rho_{\mathrm{sp}} r_{\mathrm{sp}}^{\gamma_{\mathrm{sp}}}}{3-\gamma_{\mathrm{sp}}} r_{\mathrm{infl}}^{3-\gamma_{\mathrm{sp}}} \overset{(2)}{=} \frac{4\pi\rho_{\mathrm{sp}} r_{\mathrm{sp}}^{\gamma_{\mathrm{sp}}}}{3-\gamma_{\mathrm{sp}}} 5^{3-\gamma_{\mathrm{sp}}} r_{\mathrm{sp}}^{3-\gamma_{\mathrm{sp}}} \quad\Rightarrow\quad r_{\mathrm{sp}} \overset{(3)}{=} \left[\frac{(3-\gamma_{\mathrm{sp}})\, 0.2^{3-\gamma_{\mathrm{sp}}} M}{2\pi\rho_{\mathrm{sp}}}\right]^{1/3}.$$

(1) The spike mass inside r_{infl}, the spike term of (T2.84) with $r_{\mathrm{min}} \to 0$ and upper limit r_{infl}.
(2) $r_{\mathrm{infl}} = 5r_{\mathrm{sp}}$.
(3) The powers of r_{sp} add to r_{sp}^3; divide by $4\pi\rho_{\mathrm{sp}} 5^{3-\gamma_{\mathrm{sp}}}/(3-\gamma_{\mathrm{sp}})$, and $1/5 = 0.2$.

This is their formula for r_{sp} [89, eq. (2)]. The two rules side by side:

	Eda et al.	Coogan et al.
height ρ_{sp}	$\rho_{\mathrm{NFW}}(r_{\mathrm{sp}})$ (continuity)	free parameter
mass inside $5r_{\mathrm{sp}}$	spike above $3r_h$ + NFW, (T2.84)	spike power law from $r = 0$
result	$r_{\mathrm{sp}} = 3.759\,\mathrm{pc}$ ($M = 10^4\, M_\odot$, our halo)	$r_{\mathrm{sp}} = 0.54\,\mathrm{pc}$ ($M = 10^3\, M_\odot$, $\rho_{\mathrm{sp}} = 226\, M_\odot\,\mathrm{pc}^{-3}$)

Example: why Coogan et al. use ρ_6. With their rule, $\rho_{\rm sp}$ is a poor parameter for a measurement. Write $r_{\rm sp} = (kM/\rho_{\rm sp})^{1/3}$ with $k = (3-\gamma_{\rm sp})0.2^{3-\gamma_{\rm sp}}/2\pi$. The density at a fixed radius r is

$$\rho(r) \overset{(1)}{=} \rho_{\rm sp}\Big(\frac{r_{\rm sp}}{r}\Big)^{\gamma_{\rm sp}} \overset{(2)}{=} \rho_{\rm sp}\Big(\frac{kM}{\rho_{\rm sp}}\Big)^{\gamma_{\rm sp}/3} r^{-\gamma_{\rm sp}} \overset{(3)}{=} \rho_{\rm sp}^{1-\gamma_{\rm sp}/3}(kM)^{\gamma_{\rm sp}/3}\, r^{-\gamma_{\rm sp}}.$$

(1) The spike profile.
(2) Insert $r_{\rm sp}$.
(3) Collect powers of $\rho_{\rm sp}$: $\rho_{\rm sp}^{1}\cdot\rho_{\rm sp}^{-\gamma_{\rm sp}/3}$.

For $\gamma_{\rm sp} = 7/3$ the exponent is $1-\frac{7}{9} = \frac{2}{9}$: the density felt by the binary grows only as $\rho_{\rm sp}^{2/9}$, so a factor of ten in $\rho_{\rm sp}$ moves the density at the orbit by $10^{2/9} \approx 1.7$. The gravitational wave only senses the density near the orbit, at 10^{-8}–10^{-7} pc for these systems. So Coogan et al. parametrise the spike by ρ_6, the density at the fixed radius $r_6 = 10^{-6}$ pc close to the orbits, $\rho = \rho_6(r_6/r)^{\gamma_{\rm sp}}$ [89, eqs. (3)–(4)].

Their benchmark in the new parameter. For $m_1 = 10^3\,M_\odot$ and $\rho_{\rm sp} = 226\,M_\odot\,{\rm pc}^{-3}$, their rule gives $r_{\rm sp} = [\frac{2}{3}\times 0.2^{2/3}\times 10^3/(2\pi\times 226)]^{1/3}\,{\rm pc} = 0.544\,{\rm pc}$, and then

$$\rho_6 = \rho_{\rm sp}\Big(\frac{r_{\rm sp}}{r_6}\Big)^{\gamma_{\rm sp}} = 226\times\big(5.44\times10^5\big)^{7/3}\,M_\odot\,{\rm pc}^{-3} = 5.45\times10^{15}\,M_\odot\,{\rm pc}^{-3}.$$

Why a pivot radius helps. Suppose the data fix the density $\rho(r_o)$ at the orbit, $r_o \approx 10^{-8}$ pc, and nothing else. Write the profile through an amplitude ρ_A at some radius r_A, $\ln\rho(r_o) = \ln\rho_A + \gamma_{\rm sp}\ln(r_A/r_o)$. Holding $\ln\rho(r_o)$ fixed, a change $\delta\gamma$ of the slope must be compensated by $\delta\ln\rho_A = -\ln(r_A/r_o)\,\delta\gamma$. With the amplitude at $r_A = r_{\rm sp} \approx 0.5$ pc the lever arm is $\ln(5\times10^7) \approx 18$: a slope uncertainty of 0.01 drags the amplitude by 18%, and the two are almost perfectly correlated. With the amplitude at $r_A = r_6$ the lever arm is $\ln 100 \approx 4.6$, four times shorter, and the correlation is correspondingly weaker. Moving r_A all the way to the place where the data carry their information would remove it entirely. This is the choice of a pivot scale in cosmology, and the straight-line example of Section 10.4.3 shows that the decorrelating pivot sits at the information-weighted mean of the variable along which the line is drawn; Chapter 10 reads these correlations off a Fisher matrix.

T2.9.3 Ultralight dark matter: a wave falling into the hole

If dark matter is a very light boson of mass μ, its quantum nature can show on astrophysical scales. Two comparisons of lengths decide when.

Particle or wave? A particle of momentum μv is also a wave of de Broglie wavelength $\lambda_{\rm dB} = 2\pi\hbar/(\mu v)$, and a wave cannot be squeezed into a region smaller than its wavelength. If $\lambda_{\rm dB}$ is much shorter than the region r we look at, the wave can be built into packets much smaller than r that move like particles along orbits; if it is longer, there is no way to say where along the orbit the dark matter is, and only the wave description makes sense. Up to the factor 2π, the dividing line is $\mu v r/\hbar \sim 1$: particles for $\mu v r/\hbar \gg 1$, a wave otherwise.

Does the wave see the horizon? The second length is the Compton wavelength $\hbar/(\mu c)$, the wavelength of the field's oscillation in time, to be compared with the size r_h of the hole. A wave meets the hole roughly as light meets an obstacle. If its wavelength is much shorter than the obstacle, it behaves like a ray and simply falls in. If its wavelength is much longer, it hardly resolves the horizon: most of it is reflected by the gravitational potential around the hole, and only a small part is absorbed per encounter. The reflected wave interferes with the incoming one, and the density shows standing-wave nodes.

For the speed we take $v = 1000\,\mathrm{km\,s^{-1}}$, which is $c/300$. Near the hole the typical speed is not fixed; it grows inward as the orbital speed $\sqrt{GM/r}$ does, from the $\sim 10\,\mathrm{km\,s^{-1}}$ dispersion of a dwarf galaxy to near c at the horizon. A value of $c/300$ is near the top of the range of dispersions that Hui et al. consider around black holes [98, eq. (1.5)]; a smaller v only raises the boson mass at which the wave regime ends ($\mu \propto 1/v$ below) and moves the self-gravity radius outward ($r_{\mathrm{sg}} \propto 1/v^2$ below). The hole is the $10^4\,M_\odot$ one, for which $r_h = 9.571 \times 10^{-10}\,\mathrm{pc}$.

Example: where the wave regime begins. With $\hbar c = 1.97 \times 10^{-7}\,\mathrm{eV\,m}$ and $r_h = 2.95 \times 10^7\,\mathrm{m}$, the condition $\mu v r_h/\hbar = 1$ reads

$$\mu c^2 = \frac{\hbar c}{(v/c)\,r_h} = \frac{1.97 \times 10^{-7}\,\mathrm{eV\,m}}{3.3 \times 10^{-3} \times 2.95 \times 10^7\,\mathrm{m}} \approx 2 \times 10^{-12}\,\mathrm{eV},$$

and the code gives $2.003 \times 10^{-12}\,\mathrm{eV}$. Heavier dark matter is particle-like everywhere outside the hole. Dropping the factor v/c gives the mass at which the Compton wavelength equals the horizon, $\mu c^2 = \hbar c/r_h = 6.681 \times 10^{-15}\,\mathrm{eV}$. Hui et al. sort boson masses by this comparison [98, Sec. 1]. Above it the field falls straight into the hole. Below it, part of the wave is reflected near the hole and the profile develops nodes.

How far the black hole rules. Close to the hole its gravity dominates; far away the dark matter's own gravity does. The boundary, the **self-gravity radius** r_{sg}, is where the hole stops controlling the motion. The idea is to compare two energies per unit mass: the kinetic energy of the dark matter, of order v^2, and the depth of the hole's potential, GM/r. Far out $v^2 > GM/r$: the dark matter moves too fast for the hole to hold it, and its motion is set by the halo. Close in $GM/r > v^2$: the hole binds it. The boundary is where the two are equal. The virial theorem for dark matter of mass M_{sg} moving with dispersion v in the hole's potential says exactly this: $2\langle T\rangle + \langle V\rangle = M_{\mathrm{sg}}v^2 - GMM_{\mathrm{sg}}/r_{\mathrm{sg}} = 0$, so

$$r_{\mathrm{sg}} = \frac{GM}{v^2} = \frac{r_h}{2}\left(\frac{c}{v}\right)^2 = 4.5 \times 10^4\,r_h \qquad (v = 1000\,\mathrm{km\,s^{-1}}).$$

Hui et al. define the radius by $r_h/r_{\mathrm{sg}} = (v/c)^2$ [98, eq. (1.5)], that is $r_{\mathrm{sg}} = 2GM/v^2$, twice as large, about $10^5 r_h$. The factor 2 is the choice between the circular speed, $v^2 = GM/r$, and the escape speed, $v^2 = 2GM/r$, as the speed that marks the boundary; neither is more correct. We use the round value $r_{\mathrm{sg}} = 10^5 r_h$. How much does the choice matter? Inside r_{sg} the density follows $r^{-3/2}$ (below), and at r_{sg} it joins the spike $r^{-7/3}$. The density at an orbit inside r_{sg} is then $\rho_{\mathrm{sp}}(r_{\mathrm{sp}}/r_{\mathrm{sg}})^{7/3}(r_{\mathrm{sg}}/r)^{3/2} \propto r_{\mathrm{sg}}^{-5/6}$, so moving r_{sg} by a factor 2.2 changes the density at the orbit by $2.2^{5/6} \approx 1.9$. That is small next to the eighteen orders of magnitude we are explaining, and in the forecasts the overall density is a free parameter anyway. Inside r_{sg} the dark matter is a test field on the black-hole spacetime.

The field falling into the hole. *The question.* Hui et al. asked what a field of mass μ does once it is inside r_{sg}, where only the hole's gravity acts on it and the horizon swallows whatever reaches it [98, Sec. 2]. Hancock and Witek asked the same for a massive vector (Proca) field [99, Sec. V]. Their answer, for the heavier masses $\mu r_h \gtrsim 1$ ($\hbar = c = 1$ from here to the end of the paragraph), is [98, eq. (2.17)]

$$\chi(t, r) \propto r^{1/4}\, e^{-i\mu t}\, e^{-2i\mu\sqrt{r r_h}}, \tag{T2.85}$$

where $\chi = r\phi$ is r times the field ϕ. We now derive it far from the horizon, where gravity is weak.

Symbols. The field is complex, ϕ, and the physical (real) field is $\phi + \phi^*$. A field that oscillates as $e^{-i\mu t}$ is an oscillator of frequency μ, whose energy density is $\frac{1}{2}(\dot\phi^2_{\text{real}} + \mu^2\phi^2_{\text{real}}) = \frac{1}{2}\mu^2(2|\phi|)^2$, so $\rho \approx 2\mu^2|\phi|^2$ [98, eq. (2.12)], and $\rho \propto |\chi|^2/r^2$.

The equation. Far from the horizon the field's equation (the Klein–Gordon equation on the Schwarzschild spacetime) reduces to the Schrödinger equation for a particle of mass μ in the Newtonian potential $-GM/r = -r_h/(2r)$, with the fast factor $e^{-i\mu t}$ split off. Dark matter that starts at rest far away has zero energy in that equation, and for the spherically symmetric part of the field the usual substitution $\chi = r\phi$ removes the first derivative. The radial equation is

$$\chi'' + k(r)^2\chi = 0, \qquad k(r)^2 = 2\mu^2\,\frac{GM}{r} = \mu^2\,\frac{r_h}{r}. \tag{T2.86}$$

In words: $k(r) = \mu v_{\text{ff}}$, with $v_{\text{ff}} = \sqrt{2GM/r} = \sqrt{r_h/r}$ the speed of a particle dropped from rest far away. The local wavenumber is the de Broglie wavenumber of freely falling dark matter.

A slowly changing wave. Where k changes little over one wavelength, try a wave with a slowly varying amplitude $A(r)$ and a phase that accumulates the local wavenumber, $\chi = A\,e^{-iS}$ with $S' = k$:

$$\chi'' + k^2\chi \overset{(1)}{=} \big(A'' - 2iA'k - iAk' - Ak^2\big)e^{-iS} + k^2Ae^{-iS} \overset{(2)}{=} \big[A'' - i(2A'k + Ak')\big]e^{-iS} = 0,$$

$$\text{so}\quad 2A'k + Ak' \overset{(3)}{=} 0 \quad\Leftrightarrow\quad (A^2k)' = 0 \quad\Rightarrow\quad A \propto k^{-1/2} \overset{(4)}{=} r^{1/4}.$$

(1) Differentiate twice: $\chi' = (A' - iAk)e^{-iS}$, $\chi'' = (A'' - 2iA'k - iAk' - Ak^2)e^{-iS}$.
(2) The Ak^2 terms cancel.
(3) Drop A'', which is smaller than the kept terms by a factor of order $1/(kr)^2 = 1/(\mu^2 r r_h)$, small when $\mu r_h \gtrsim 1$ and $r \gg r_h$; the imaginary part must vanish. Multiplying by A shows $2AA'k + A^2k' = (A^2k)'$.
(4) $k \propto r^{-1/2}$, so $k^{-1/2} \propto r^{1/4}$.

The phase is the one the free-fall picture predicts:

$$S(r) = \int^r k(r')\,\mathrm{d}r' \overset{(1)}{=} \mu\sqrt{r_h}\int^r \frac{\mathrm{d}r'}{\sqrt{r'}} \overset{(2)}{=} 2\mu\sqrt{r r_h},$$

(1) $k = \mu v_{\text{ff}}$ with $v_{\text{ff}} = \sqrt{r_h/r'}$.
(2) $\int r'^{-1/2}\mathrm{d}r' = 2r^{1/2}$.

and with $e^{-i\mu t}$ restored, $e^{-i(\mu t + S)}$ has constant phase along $S = -\mu t + \text{const}$: as t grows, r shrinks. The wave moves *inward*, which is the only choice allowed when the horizon absorbs everything and nothing comes back out. This is (T2.85): amplitude $r^{1/4}$, phase $2\mu\sqrt{r r_h}$. So the field is dark matter falling freely into the hole.

The density, and a check. $\rho \propto |\chi|^2/r^2 \propto r^{1/2}/r^2 = r^{-3/2}$. The amplitude condition $(A^2k)' = 0$ has a direct meaning: $A^2k \propto |\chi|^2 v_{\text{ff}} \propto r^2\rho\, v_{\text{ff}}$ is the mass crossing a sphere per unit time, divided by 4π. So the same exponent follows from steady inflow without any wave equation: the flux $4\pi r^2\rho\, v_{\text{ff}}$ must be the same at every radius, and $v_{\text{ff}} \propto r^{-1/2}$ forces $\rho \propto r^{-3/2}$ [98, Sec. 4]. Hancock and Witek find the same $r^{-3/2}$ for the vector field, for every mode [99, Sec. VI A].

Joining the pieces. At r_{sg} the infall profile must take over from the spike with the same density, $\rho_{\text{sp}}(r_{\text{sp}}/r_{\text{sg}})^{7/3}$. Inside, it falls as $r^{-3/2}$:

$$\rho(r) = \rho_{\text{sp}}\Big(\frac{r_{\text{sp}}}{r_{\text{sg}}}\Big)^{7/3}\Big(\frac{r_{\text{sg}}}{r}\Big)^{3/2} \qquad (r < r_{\text{sg}}).$$

The resulting profile is piecewise: infall $\rho \propto r^{-3/2}$ inside r_{sg}, spike $r^{-7/3}$ out to r_{sp}, NFW beyond. For $\mu \lesssim 6.681 \times 10^{-15}$ eV the WKB step fails near the hole and the field must be found

numerically. Its density then oscillates on the scale of the de Broglie wavelength, and Hancock and Witek smooth it over a few radial periods before matching [99, eq. (69)]. The smoothed profile is spread out near the hole instead of sharply peaked [99].

Figure T2.19 shows the whole construction. At r_{sg} the spike has reached $1.190\times10^{11}\,M_\odot\,\mathrm{pc}^{-3}$. On the EMRI's orbit four years before the ISCO, at $30.09\,r_h$, the density is $2.279\times10^{16}\,M_\odot\,\mathrm{pc}^{-3}$, about 2.279×10^{18} times the local value. That is the enhancement the opening estimate called for.

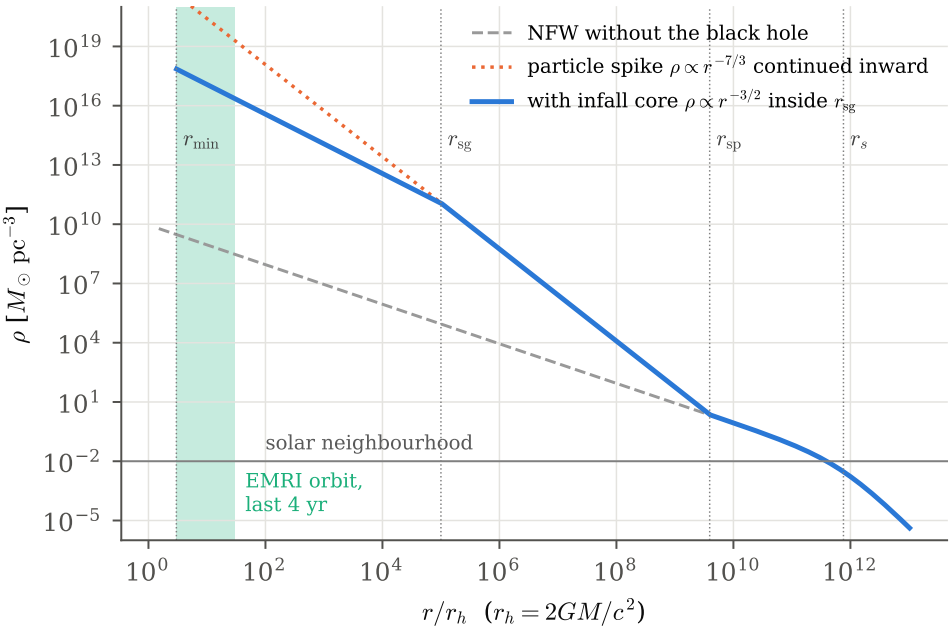


Figure T2.19: Dark matter around a $10^4\,M_\odot$ black hole in a $10^8\,M_\odot$ NFW halo. Without the hole the density would follow the grey dashed NFW profile. Slow growth of the hole turns the cusp inside r_{sp} into a $r^{-7/3}$ spike (dotted where continued inward). Inside the self-gravity radius $r_{sg} = 10^5 r_h$ an ultralight field falls freely into the hole and the profile flattens to $r^{-3/2}$ (solid). The curves stop at $r_{min} = 3r_h$: inside it nothing survives, the hole in the middle of the spike, and the density is zero. The green band is where the EMRI orbits during its last four years; the horizontal line is the density near the Sun.

Two further wave-dark-matter pictures. Kim et al. follow the adiabatic compression of a wave halo quantum mechanically [86]. Far from the hole the compressed wave halo looks like the particle spike, as expected when the wavelength is short. Close to it the picture changes, and the reason is the capture rule used above for particles. A wave component with angular-momentum quantum number l carries angular momentum $l\hbar$, that is $l\hbar/\mu$ per unit mass. A particle with less than $4GM/c$ per unit mass is swallowed, so the components with

$$l \lesssim l_{\mathrm{crit}} = \frac{4GM}{c}\cdot\frac{\mu}{\hbar} = \frac{2\mu c\, r_h}{\hbar}$$

fall into the hole, and over the age of the halo they are gone. A component of angular momentum l cannot come closer to the centre than the radius where its centrifugal barrier turns it back, a radius that grows with l. Once every component below l_{crit} has been absorbed, nothing is left to fill the region inside the turning radius of l_{crit}, and the density is cut off there [86, eq. (40)]. In the particle limit ($\mu r_h c/\hbar \gg 1$) l_{crit} is huge and the cutoff is the particle cutoff near 2–$3r_h$; for light bosons it moves out.

Chase et al. consider instead a virialised halo of ultralight *vector* dark matter made of many random waves [100]. The field oscillates at the frequency $\mu c^2/\hbar$, and its stress, which is quadratic in the field, oscillates at twice that, $\Omega = 2\mu c^2/\hbar$. Through Einstein's equations the stress makes the gravitational potential oscillate at Ω too. What the binary feels depends on the shape of that oscillation. For a scalar the oscillating stress is a pressure, the same in every direction; the oscillating pull on the companion is radial and the same at every point of a circular orbit, so it exerts no torque, and over many field oscillations its effect on the orbit averages to zero. A vector field points along a direction, its polarisation, so its stress is stronger along that axis. The oscillating tidal pull then depends on where the companion is on its orbit: a pattern with two maxima per turn, which exerts a torque proportional to $\sin 2(\omega t - \varphi_0)$ for orbital angular frequency ω and polarisation angle φ_0, and whose strength oscillates as $\cos \Omega t$. The product

$$\cos \Omega t \, \sin 2(\omega t - \varphi_0) = \tfrac{1}{2}\big[\sin\big((2\omega + \Omega)t - 2\varphi_0\big) + \sin\big((2\omega - \Omega)t - 2\varphi_0\big)\big]$$

averages to zero over many orbits unless $2\omega \approx \Omega$, when the second term stops oscillating. Away from that condition the pushes cancel; at it they add up, the way a swing responds only to pushes timed to its own period. This is the **resonance**, and the binary's phase changes mainly while its orbital frequency sweeps through it [100, Sec. V]. The effect is concentrated near one frequency, so, unlike friction and accretion (Section T2.10), it cannot be written as a single power of f.

T2.10 Environmental forces and the power law in the phase

Put the light companion of Section T2.4 (mass m_2, circling the heavy hole m_1, with $M = m_1 + m_2$) on its orbit inside the spike of Section T2.9. It moves at $v = \sqrt{GM/r}$, about $0.13c$ at $30r_h$, through dark matter of density $\rho(r)$. Three things happen. Its gravity focuses the dark matter it passes into a trailing wake, and the wake pulls it back: **dynamical friction**. It swallows the dark matter that comes close enough, and the swallowed matter has to be brought up to the companion's speed: **accretion**. And the dark matter inside the orbit adds to the central pull. The first two drain orbital energy on top of the gravitational waves, so the binary sweeps through frequency faster than in vacuum.

The question for the statistics is narrow: *for each effect, by how much does it speed up the sweep at frequency f, and what does that do to the Fourier phase $\Psi(f)$?* (Ψ is the phase of the stationary-phase signal of Section T2.5, $\tilde{h} = \mathcal{A} f^{-7/6} e^{-i\Psi}$.) The answer turns out to have the same form for all of them, a power of f added to the phase, and that is what makes a single, general forecast possible in Chapter 10.

T2.10.1 The forces

Dynamical friction. For a body of mass m_2 moving at speed v through collisionless matter of density ρ, Chandrasekhar's formula gives a drag

$$F_{\rm DF} = \frac{4\pi G^2 m_2^2 \, \rho \, \xi \ln \Lambda}{v^2} \tag{T2.87}$$

[101], [82, eq. (16)], [89, eq. (7)]. Every factor can be derived from one deflection.

One particle passing. Work in the companion's frame, where the dark matter streams past at speed v. Take first the simplest stream, every particle moving at exactly v. A particle of mass $m_{\rm dm} \ll m_2$ that would miss the companion by the **impact parameter** s (we write s, not the usual b, which is reserved for the ppE exponent below) is bent through an angle θ given by the

Kepler hyperbola, $\tan(\theta/2) = Gm_2/(sv^2)$. Its speed far away is unchanged, only its direction turns, so it loses forward momentum $m_{\rm dm}v(1-\cos\theta)$. That momentum goes to the companion and points along the stream, that is, against the companion's motion through the dark matter. The sideways kicks cancel between particles passing on opposite sides.

All the particles. Particles with impact parameter between s and $s + \mathrm{d}s$ arrive at the rate $(\rho/m_{\rm dm})\, v\, 2\pi s\, \mathrm{d}s$. With $s_{90} \equiv Gm_2/v^2$, the impact parameter that turns a particle through 90°,

$$F \overset{(1)}{=} \int \frac{\rho}{m_{\rm dm}}\, v\, 2\pi s\, \mathrm{d}s\; m_{\rm dm} v\,(1-\cos\theta) \overset{(2)}{=} 4\pi\rho v^2 s_{90}^2 \int_0^{s_{\rm max}} \frac{s\,\mathrm{d}s}{s^2+s_{90}^2}$$
$$\overset{(3)}{=} \frac{4\pi G^2 m_2^2 \rho}{v^2}\cdot\frac{1}{2}\ln\left(1+\frac{s_{\rm max}^2}{s_{90}^2}\right) \overset{(4)}{=} \frac{4\pi G^2 m_2^2\rho}{v^2}\,\ln\Lambda, \qquad \Lambda \equiv \frac{s_{\rm max}}{s_{90}}. \tag{T2.88}$$

(1) Momentum per particle times the rate of particles, summed over impact parameters. The mass $m_{\rm dm}$ cancels: only the density matters.
(2) $1-\cos\theta = 2\sin^2(\theta/2) = 2\tan^2(\theta/2)/[1+\tan^2(\theta/2)] = 2s_{90}^2/(s^2+s_{90}^2)$, using $\tan(\theta/2) = s_{90}/s$.
(3) $\int s\,\mathrm{d}s/(s^2+s_{90}^2) = \frac{1}{2}\ln(s^2+s_{90}^2)$; and $v^2 s_{90}^2 = G^2m_2^2/v^2$.
(4) For $s_{\rm max} \gg s_{90}$, $\frac{1}{2}\ln(1+\Lambda^2) \approx \ln\Lambda$.

So each factor has a reason. One m_2 comes from the strength of each deflection and the other from the size s_{90} of the region where deflections are strong; the $1/v^2$ says a slow body has more time to bend each particle. The logarithm counts the e-folds of impact parameter between s_{90} and $s_{\rm max}$, each adding the same amount, so distant encounters matter as much as close ones and the integral must be cut off by hand. The lower end is $s_{\rm min} = s_{90}$, below which particles are turned right round and nothing more is gained; the upper end $s_{\rm max}$ is the size of the region in which the stream is a uniform flow, the orbit itself, $s_{\rm max} \sim r$.

Particles of all speeds. A real spike has particles with a spread of speeds. Chandrasekhar showed that for an isotropic distribution the particles faster than the body pull it, on average, equally forwards and backwards, and only those slower than v contribute. The factor ξ in (T2.87) is that fraction, $\xi = 4\pi\int_0^v f(u)u^2\,\mathrm{d}u$ for the distribution $f(u)$ of particle speeds. For the 7/3 spike, whose speeds are set by the hole's potential, it is about 0.58 [89]; papers that ignore the spread set it to one.

The size of the logarithm. On a circular orbit $v^2 = GM/r$, so $s_{90} = Gm_2/v^2 = r\,m_2/M$. With $s_{\rm max} = r$, $\Lambda \approx M/m_2 \approx m_1/m_2$. Coogan et al. take a more cautious $s_{\rm max} = \sqrt{r s_{90}}$, the geometric mean of the two ends, which halves the logarithm: $\ln\Lambda = \ln\sqrt{m_1/m_2}$ [89]. For the $10^3 + 1.4\,M_\odot$ IMRI that is 3.286 against 6.6; others take about 3 [82]. Only the logarithm enters, so this factor of two is the size of the uncertainty in the prescription. *Check at* $30r_h$. There $v^2/c^2 = r_h/(2r) = 1/60$, so $s_{90} = Gm_2/v^2 = \frac{1}{2}(2Gm_2/c^2)(c/v)^2 = 30$ Schwarzschild radii of the companion: the deflections that build the logarithm happen well outside the companion's own horizon, where the Newtonian hyperbola is a fair description. The power the friction removes is $P_{\rm DF} = F_{\rm DF}\,v$.

Friction on a wave. If the dark matter is a wave (Section T2.9), a wake smaller than its wavelength cannot form. The force keeps the form of (T2.87), with $\xi\ln\Lambda$ replaced by a function of kr, where $k = \mu v/\hbar$ is the dark matter's wavenumber [102, eq. (D14)], [86, eq. (42)]:

$$F_{\rm DF}^{\rm wave} = \frac{4\pi G^2 m_2^2 \rho\, C_w(kr)}{v^2}, \qquad C_w(kr) = \mathrm{Cin}(2kr) - 1 + \frac{\sin 2kr}{2kr}, \qquad \mathrm{Cin}(z) = \int_0^z \frac{1-\cos t}{t}\,\mathrm{d}t. \tag{T2.89}$$

Hui and Ostriker obtain it by solving the scattering of the wave by the companion and summing the momentum carried off by its partial waves. The integral shows why it is a softened Coulomb logarithm. Its variable t measures distance in units of the reduced wavelength. Far beyond a wavelength, $\cos t$ oscillates and averages away, and the integrand is $1/t$, the integrand $\mathrm{d}s/s$ of $\ln\Lambda$. Within a wavelength the interference of the wave ($1-\cos t \approx t^2/2$) kills the integrand. So Cin is a Coulomb logarithm whose lower cutoff is the wavelength instead of s_{90}, and whose upper cutoff is the orbit, $z = 2kr$; the terms $-1 + \sin z/z$ come from the finite size of the region.

Its two limits tell the story. For a short wavelength, $kr \gg 1$, split off the logarithm: $\mathrm{Cin}(z) = \gamma_E + \ln z - \mathrm{Ci}(z)$, where $\gamma_E = 0.577$ is Euler's constant and $\mathrm{Ci}(z) = -\int_z^\infty \cos t\,\mathrm{d}t/t$ is a decaying oscillation, of size $|\sin z|/z$ for large z. Both $\mathrm{Ci}(z)$ and $\sin z/z$ go to zero, so $C_w \to \ln(2kr) + \gamma_E - 1$: a Coulomb logarithm with $s_{\max} \sim r$ and $s_{\min} \sim 1/k$, the particle result. For a long wavelength, $z = 2kr \ll 1$,

$$C_w \overset{(1)}{=} \left(\frac{z^2}{4} - \frac{z^4}{96}\right) - 1 + \left(1 - \frac{z^2}{6} + \frac{z^4}{120}\right) + O(z^6) \overset{(2)}{=} \frac{z^2}{12} + O(z^4) = \frac{(kr)^2}{3} + O\big((kr)^4\big),$$

(1) Expand $1-\cos t = t^2/2 - t^4/24$, so $\mathrm{Cin}(z) = z^2/4 - z^4/96$; and $\sin z/z = 1 - z^2/6 + z^4/120$.
(2) The constants cancel and $\frac{1}{4} - \frac{1}{6} = \frac{1}{12}$; with $z = 2kr$, $z^2/12 = (kr)^2/3$.

so friction is suppressed as $(kr)^2$ when the wavelength exceeds the orbit. Figure T2.20 shows both limits; at $kr = 0.1$ the small-kr form is already accurate to 0.8254%. This suppression, together with the lower density near the hole, is why wave-like dark matter dephases the inspiral less than particle dark matter of the same mass density [86, Sec. III D].

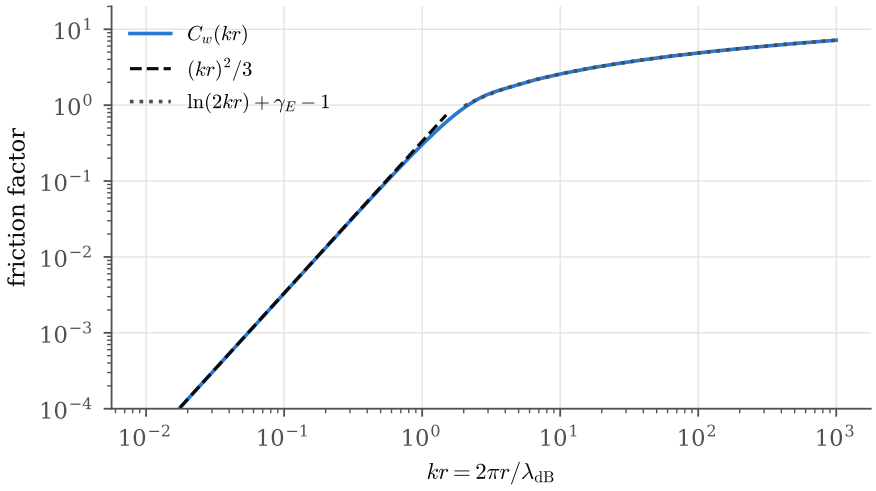


Figure T2.20: The friction factor of a wave-like medium as a function of $kr = 2\pi r/\lambda_{\mathrm{dB}}$. When the de Broglie wavelength is longer than the orbit ($kr \ll 1$) friction is suppressed as $(kr)^2/3$ (dashed); when it is shorter, the factor grows like the Coulomb logarithm of particle dark matter (dotted).

Accretion. How much dark matter does the companion swallow? In its own rest frame the dark matter streams in at speed v, each particle with angular momentum per unit mass $L = sv$ about the companion. General relativity supplies one fact: a particle arriving from far away with L below $L_{\mathrm{crit}} = 4Gm_2/c$ meets no centrifugal barrier and crosses the horizon (the same capture rule that cuts off the spike near the central hole, Section T2.9). Setting $sv = L_{\mathrm{crit}}$ gives the capture radius and the cross-section,

$$s_{\mathrm{cap}} = \frac{4Gm_2}{cv}, \qquad \sigma = \pi s_{\mathrm{cap}}^2 = \frac{16\pi G^2 m_2^2}{c^2 v^2}.$$

The mass swallowed per unit time is the mass flux through that area, $\dot{m} = \rho v \sigma = 16\pi G^2 m_2^2 \rho/(c^2 v)$. Now go back to the frame of the spike. There the dark matter is (on average) at rest, and after capture it moves with the companion at speed v. The companion must supply that momentum, $\dot{m}\, v$ per unit time: a drag force $F_{\rm acc} = \dot{m}\, v$ and a power $P_{\rm acc} = \dot{m} v^2$. The ratio $s_{\rm cap}/s_{90} = 4v/c$ is about 0.5 at $30r_h$: the particles swallowed are a subset of those strongly deflected, which is why accretion comes out as a correction to friction. Compared with friction,

$$\frac{P_{\rm acc}}{P_{\rm DF}} = \frac{16\pi G^2 m_2^2 \rho\, v/c^2}{4\pi G^2 m_2^2 \rho \ln\Lambda/v} = \frac{4v^2}{c^2 \ln\Lambda} = \frac{2r_h}{r\ln\Lambda}, \tag{T2.90}$$

using $v^2 = GM/r = c^2 r_h/(2r)$ and $\xi = 1$. With $\ln\Lambda = 3$ this is 0.02222 at $30r_h$ and 0.2222 at the ISCO. Accretion of collisionless dark matter is a relativistic correction to friction. Kim et al. check this for their binaries: over the last five years the companion's mass grows by only about one part in 10^4 [86, Sec. IV A]. For gas, which is collisional, the Bondi–Hoyle–Lyttleton rate $\dot{m} \propto \rho/v^3$ applies instead, and accretion then scales exactly like friction [92, Sec. X].

The gravitational pull. The dark matter inside the orbit adds to the central mass. For the spike $\rho = \rho_6 (r_6/r)^\gamma$ (with $\gamma < 3$) it weighs

$$M_{\rm DM}(<r) = \int_0^r 4\pi r'^2 \rho_6 \left(\frac{r_6}{r'}\right)^\gamma {\rm d}r' = \frac{4\pi\rho_6 r_6^\gamma}{3-\gamma}\, r^{3-\gamma}.$$

By the shell theorem, a spherical mass inside the orbit pulls like a point mass at the centre, so Kepler's law holds with $M_{\rm tot} = M + M_{\rm DM}(<r)$ in place of M [88, eqs. (9)–(11)]. This removes no energy, but it changes the relation between the frequency and the orbit. How much does it change the chirp? Hold the gravitational-wave frequency f fixed, with $\omega = \pi f$, and treat $M_{\rm DM}$ as a constant addition to the central mass over the small range of radii near the current orbit. Each quantity in the chirp rate is a power of $M_{\rm tot}$:

$$r \overset{(1)}{=} \left(\frac{GM_{\rm tot}}{\pi^2 f^2}\right)^{1/3} \propto M_{\rm tot}^{1/3}, \qquad E \overset{(2)}{=} -\frac{GM_{\rm tot}\mu}{2r} = -\frac{\mu}{2}(\pi G M_{\rm tot} f)^{2/3} \propto M_{\rm tot}^{2/3},$$

$$P_{\rm GW} \overset{(3)}{=} \frac{32}{5}\frac{G}{c^5}\mu^2 r^4 \omega^6 \propto M_{\rm tot}^{4/3}, \qquad \frac{\dot{f}}{\dot{f}_{\rm V}} \overset{(4)}{=} \frac{P_{\rm GW}/|{\rm d}E/{\rm d}f|}{(\dots)_{M_{\rm tot}=M}} = \left(1+\frac{M_{\rm DM}}{M}\right)^{2/3} \approx 1 + \frac{2}{3}\frac{M_{\rm DM}(<r)}{M}.$$

(1) Kepler's law, $\omega^2 = GM_{\rm tot}/r^3$.
(2) The binding energy (T2.41) with $M \to M_{\rm tot}$; ${\rm d}E/{\rm d}f$ scales the same way.
(3) The quadrupole power at fixed ω depends on $M_{\rm tot}$ only through r^4.
(4) The chirp rate of (T2.43) is $P_{\rm GW}/|{\rm d}E/{\rm d}f| \propto M_{\rm tot}^{4/3-2/3}$; expand for $M_{\rm DM} \ll M$.

So the pull speeds up the chirp by the fraction

$$\epsilon_{\rm pull} = \frac{2}{3}\frac{M_{\rm DM}(<r)}{M} = \frac{8\pi\rho_6 r_6^\gamma}{3(3-\gamma)M}\, r^{3-\gamma} \propto f^{-2(3-\gamma)/3},$$

using $r \propto f^{-2/3}$. The coefficient $\frac{2}{3}$ is leading order only, because $M_{\rm DM}(<r)$ itself changes as the orbit shrinks; the exponent does not depend on that. Friction usually matters more [92, Sec. IX].

T2.10.2 From an extra power to a power law in the phase

The fractional extra power. Throughout, f is the gravitational-wave frequency, twice the orbital one, $f = \omega/\pi$ (T2.24). Friction drains energy slowly, so at each moment the orbit is

still a circle obeying Kepler's law, and its energy at frequency f is the vacuum $E(f)$ of (T2.41); only the rate at which it moves through these energies changes. Energy balance now reads $\mathrm{d}E/\mathrm{d}t = -(P_{\mathrm{GW}} + P_{\mathrm{DF}})$, with P_{GW} the quadrupole power of a circular binary (T2.27). As in (T2.43),

$$\dot f \overset{(1)}{=} -\frac{P_{\mathrm{GW}}+P_{\mathrm{DF}}}{\mathrm{d}E/\mathrm{d}f} \overset{(2)}{=} \frac{P_{\mathrm{GW}}}{|\mathrm{d}E/\mathrm{d}f|}\left(1+\frac{P_{\mathrm{DF}}}{P_{\mathrm{GW}}}\right) \overset{(3)}{=} \dot f_{\mathrm{V}}\left[1+\epsilon(f)\right], \qquad \epsilon \equiv \frac{P_{\mathrm{DF}}}{P_{\mathrm{GW}}}, \tag{T2.91}$$

(1) $\mathrm{d}E/\mathrm{d}t = (\mathrm{d}E/\mathrm{d}f)\,\dot f$, since E depends on time only through f.
(2) $\mathrm{d}E/\mathrm{d}f < 0$; factor out P_{GW}.
(3) The first factor is the vacuum rate $\dot f_{\mathrm{V}}$ of (T2.43).

This is the energy balance of Kim et al. [86, eq. (36)]. For friction in a spike $\rho = \rho_6(r_6/r)^\gamma$ (here and below γ is the spike slope γ_{sp} of Section T2.9, and ρ_6 the density at $r_6 = 10^{-6}$ pc),

$$\epsilon \overset{(1)}{=} \frac{4\pi G^2 m_2^2 \rho\, \xi \ln\Lambda/v}{\frac{32}{5}G^4 M(m_1m_2)^2/(c^5 r^5)} \overset{(2)}{=} \frac{5\pi c^5 \xi \ln\Lambda}{8G^{5/2}M^{3/2}m_1^2}\,\rho(r)\,r^{11/2} \overset{(3)}{=} c_f\, f^{-n}, \qquad n = \frac{11-2\gamma}{3}. \tag{T2.92}$$

(1) $P_{\mathrm{DF}} = F_{\mathrm{DF}}v$ from (T2.87); $P_{\mathrm{GW}} = \frac{32}{5}(G/c^5)\mu^2 r^4\omega^6 = \frac{32}{5}G^4M^3\mu^2/(c^5r^5)$ by Kepler, and $M^3\mu^2 = M(m_1m_2)^2$ [89, eq. (6)].
(2) m_2^2 cancels, and $1/v = r^{1/2}(GM)^{-1/2}$. The result $\epsilon \propto \rho r^{11/2}$ is [88, eq. (16)].
(3) Insert $\rho = \rho_6 r_6^\gamma r^{-\gamma}$, so $\rho r^{11/2} = \rho_6 r_6^\gamma\, r^{11/2-\gamma}$, and trade the radius for the gravitational-wave frequency with Kepler's law, $r = (GM/\pi^2 f^2)^{1/3}$: $r^{11/2-\gamma} = (GM/\pi^2)^{(11/2-\gamma)/3} f^{-\frac{2}{3}(\frac{11}{2}-\gamma)}$, and $\frac{2}{3}(\frac{11}{2}-\gamma) = \frac{11-2\gamma}{3} = n$.

Collecting the constants of step (3),

$$c_f = \frac{5\pi c^5 \xi \ln\Lambda}{8G^{5/2}M^{3/2}m_1^2}\,\rho_6 r_6^\gamma \left(\frac{GM}{\pi^2}\right)^{(11/2-\gamma)/3}, \tag{T2.93}$$

with units of Hz^n, so that $c_f f^{-n}$ is a pure number; the f in $c_f f^{-n}$ is the gravitational-wave frequency. Friction and radiation remove energy equally fast where $\epsilon = 1$, at $f_{\mathrm{eq}} \equiv c_f^{1/n}$. The same step works for any effect whose power, relative to the gravitational-wave power, varies as a power of the radius: $P_{\mathrm{env}}/P_{\mathrm{GW}} \propto r^p$ gives $\epsilon \propto f^{-2p/3}$. Since $P_{\mathrm{GW}} \propto r^{-5}$ falls so steeply with radius, nearly every environmental effect has $p > 0$: it matters most early in the inspiral, at low frequency.

The phase. Recall why the chirp rate controls the Fourier phase. The slope of the phase is the arrival time of each frequency, $\mathrm{d}\Psi/\mathrm{d}f = 2\pi t(f)$, and differentiating once more turns $\mathrm{d}t/\mathrm{d}f$ into $1/\dot f$: the curvature of the phase is $2\pi/\dot f$ (T2.53), derived with the stationary-phase approximation in Section T2.5. A binary that sweeps faster spends less time at each frequency, and the phase bends less. With (T2.91), and $\dot f_{\mathrm{V}} = Kf^{11/3}$, $K = \frac{96}{5}\pi^{8/3}(G\mathcal{M}/c^3)^{5/3}$,

$$\begin{aligned}
\frac{\mathrm{d}^2\Psi}{\mathrm{d}f^2} &\overset{(1)}{=} \frac{2\pi}{\dot f_{\mathrm{V}}(1+\epsilon)} \overset{(2)}{=} \frac{2\pi}{\dot f_{\mathrm{V}}} - \frac{2\pi c_f}{K} f^{-11/3-n} + O(\epsilon^2),\\
\delta\Psi(f) &\overset{(3)}{=} -\frac{2\pi c_f}{K\,(n+\frac{8}{3})(n+\frac{5}{3})} f^{-5/3-n} + C_0 + C_1 f \overset{(4)}{=} \beta\, u^b, \qquad u \equiv \frac{\pi G\mathcal{M} f}{c^3}, \quad b = -\frac{5}{3} - n.
\end{aligned} \tag{T2.94}$$

(1) (T2.53) with (T2.91).
(2) $1/(1+\epsilon) = 1 - \epsilon + \ldots$ for $\epsilon \ll 1$; the first term gives the vacuum phase (T2.54), the second is the change, with $\epsilon/f^{11/3} = c_f f^{-11/3-n}$.

(3) Integrate twice: $f^{-11/3-n} \to f^{-8/3-n}/(-\frac{8}{3}-n) \to f^{-5/3-n}/[(\frac{8}{3}+n)(\frac{5}{3}+n)]$. The two integration constants $C_0 + C_1 f$ have the form of a shift of ϕ_c and of t_c in (T2.54), so they are absorbed there (Section T2.4.3).
(4) Rewrite $f = u\,c^3/(\pi G\mathcal{M})$, so $f^{-5/3-n} = u^b(\pi G\mathcal{M}/c^3)^{5/3+n}$; all constants go into β. u is the reduced frequency of (T2.58).

The amplitude, written out, is

$$\beta = -\frac{2\pi c_f}{K\,(n+\frac{8}{3})(n+\frac{5}{3})}\Big(\frac{\pi G\mathcal{M}}{c^3}\Big)^{5/3+n} = -\frac{5}{48}\,\frac{c_f\,(\pi G\mathcal{M}/c^3)^n}{(n+\frac{8}{3})(n+\frac{5}{3})},$$

using $2\pi(\pi G\mathcal{M}/c^3)^{5/3}/K = \frac{10\pi}{96}\pi^{5/3-8/3} = \frac{5}{48}$. It is proportional to c_f, hence to the density ρ_6 of the spike, and the combination $c_f(\pi G\mathcal{M}/c^3)^n$ is a pure number. A check: for $n = 0$ the environment speeds the chirp up by a constant fraction c_f, which by step (2) multiplies the Newtonian term $\frac{3}{128}u^{-5/3}$ of (T2.54) by $1 - c_f$; and indeed $-\frac{5}{48}c_f/(\frac{8}{3}\cdot\frac{5}{3}) = -\frac{3}{128}c_f$.

Key idea: Environments are power laws in the phase

If an environment adds a power $P_{\rm env}$ with $P_{\rm env}/P_{\rm GW} = c_f f^{-n} \ll 1$, the Fourier phase gains $\delta\Psi = \beta u^b$ with $u = \pi G\mathcal{M}f/c^3$ and $b = -5/3 - n$; the amplitude β carries the strength of the effect and the exponent b its physics. For dynamical friction in a spike $\rho \propto r^{-\gamma}$, $b = (2\gamma - 16)/3$; for $\gamma = 7/3$, $b = -34/9 \approx -3.78$.

This is the parametrised post-Einsteinian (ppE) phase term βu^b of (T2.70), now with a definite exponent: the general rule (T2.69) with $p = -3n$ gives the same $b = -\frac{5}{3} - n$. In the post-Newtonian counting of Section T2.6, a correction of relative size $(v/c)^{2k}$ is of order kPN, and on a circular orbit $(v/c)^2 = (\pi GMf/c^3)^{2/3} \propto f^{2/3}$. So $f^{-n} \propto [(v/c)^2]^{-3n/2}$, and $\epsilon \propto f^{-n}$ is a correction of order $-\frac{3}{2}n$ PN, so friction in a uniform medium ($\gamma = 0$, $n = 11/3$) is a -5.5PN effect. The dissipative effects, friction and accretion in any spike with $\gamma < 3$, have $b < -7/3$: they act more strongly at low frequency than the dipole radiation ($b = -7/3$) and the massive graviton ($b = -1$) of Section T2.6. In the Fisher matrix of Chapter 10 this is what separates them from the chirp mass, whose own phase term goes as $u^{-5/3}$: two terms with different powers of f are hard to mimic with one another.

The time-domain dephasing. Papers on environments usually plot a different phase: the number of radians the wave still has to go through before coalescence. Its phase advances at the rate $2\pi f$, so between frequency f and coalescence at f_c

$$\Phi(f) = 2\pi\int_t^{t_c} f(t')\,\mathrm{d}t' = 2\pi\int_f^{f_c}\frac{f'}{\dot f(f')}\,\mathrm{d}f',$$

changing the variable from time to frequency with $\mathrm{d}t' = \mathrm{d}f'/\dot f$. The **dephasing** is the vacuum phase still to come minus the phase still to come in the environment, $\Delta\Phi(f) = \Phi_{\rm V}(f) - \Phi(f)$ [86, eq. (49)]; $\Phi_{\rm V}$ is the Φ of Section T2.5 for the Newtonian chirp. A faster sweep leaves fewer cycles, so $\Delta\Phi > 0$.

Why two phases? $\Delta\Phi$ is the size that is easy to feel: it counts the radians of wave the environment removes, and once it reaches about a radian a vacuum template is visibly out of step with the signal. The Fisher matrix and the likelihood work in the frequency domain, with the phase Ψ of the template; there part of $\Delta\Phi$ is absorbed by adjusting the arrival time t_c and the phase ϕ_c, which are fitted anyway. So the two answer different questions, and we compute both.

With (T2.91) and $\dot f_{\rm V} = K f^{11/3}$,

$$\Delta\Phi(f) \overset{(1)}{=} 2\pi \int_f^\infty \frac{f'}{K f'^{11/3}}\Big(1 - \frac{1}{1 + c_f f'^{-n}}\Big)\mathrm{d}f' \overset{(2)}{=} \frac{2\pi c_f}{K}\int_f^\infty f'^{-8/3-n}\mathrm{d}f'$$
$$\overset{(3)}{=} \frac{2\pi c_f}{K}\,\frac{f^{-5/3-n}}{\frac{5}{3}+n} \overset{(4)}{=} \frac{5\,c_f f^{-n}}{2(8-\gamma)}\,\Phi_{\rm V}(f). \tag{T2.95}$$

(1) The difference of the two phase integrals, $1/\dot f_{\rm V} - 1/\dot f = \frac{1}{\dot f_{\rm V}}\big(1 - \frac{1}{1+\epsilon}\big)$, with the upper limit f_c sent to infinity (justified below).

(2) $1 - 1/(1+\epsilon) \approx \epsilon = c_f f'^{-n}$ for $\epsilon \ll 1$.

(3) $\int_f^\infty f'^{-8/3-n}\mathrm{d}f' = f^{-5/3-n}/(\frac{5}{3}+n)$.

(4) $\Phi_{\rm V} = \frac{2\pi}{K}\cdot\frac{3}{5}f^{-5/3}$, so the ratio is $\frac{5}{3}c_f f^{-n}/(\frac{5}{3}+n) = 5c_f f^{-n}/(5+3n)$, and $5+3n = 16-2\gamma$.

The upper limit. The integrand of $\Delta\Phi$ falls as $f'^{-8/3-n}$, so the part beyond f_c is a fraction $(f/f_c)^{5/3+n}$ of the whole. For the IMRI, with f_c the ISCO frequency of about 4.4 Hz and f the 21.7 mHz at which the four-year window opens, that fraction is about 2×10^{-9} for $\gamma = 7/3$, and smaller still for the other slopes. Integrating to infinity, as here, and stopping at the ISCO, as the script does, give the same numbers to that accuracy.

This is the high-frequency limit quoted by Coogan et al. [89, eq. (16)], and it has the same exponent $-5/3-n$ as $\delta\Psi$. The two coefficients compare in one line: from (T2.94) and (T2.95),

$$\frac{\delta\Psi}{\Delta\Phi} = \frac{-2\pi c_f/[K(n+\frac{8}{3})(n+\frac{5}{3})]}{2\pi c_f/[K(n+\frac{5}{3})]} = -\frac{1}{n+\frac{8}{3}} \qquad (f \gg f_{\rm eq}),$$

and for $n = 0$ (a constant fractional speed-up) this is the factor $\frac{3}{8}$ between the two kinds of phase found in Section T2.5. The sign only records that $\Delta\Phi$ is defined as vacuum minus environment.

Without the expansion. Step (2) needs $\epsilon \ll 1$, i.e. $f \gg f_{\rm eq}$. The integral can also be done exactly. Write the geometric series $1/(1+\epsilon) = \sum_k (-\epsilon)^k$ and integrate term by term:

$$\Phi(f) \overset{(1)}{=} \frac{2\pi}{K}\sum_{k=0}^{\infty}(-c_f)^k \int_f^\infty f'^{-8/3-kn}\mathrm{d}f'$$
$$\overset{(2)}{=} \Phi_{\rm V}(f)\sum_{k=0}^{\infty}\frac{a}{a+k}\big(-c_f f^{-n}\big)^k$$
$$\overset{(3)}{=} \Phi_{\rm V}(f)\;{}_2F_1\big(1, a; 1+a; -c_f f^{-n}\big), \qquad a \equiv \frac{5}{3n}. \tag{T2.96}$$

(1) $\Phi = 2\pi\int_f^\infty f'/\dot f\,\mathrm{d}f'$ with $\dot f = K f'^{11/3}(1 + c_f f'^{-n})$.

(2) $\int_f^\infty f'^{-8/3-kn}\mathrm{d}f' = f^{-5/3-kn}/(\frac{5}{3}+kn)$; factoring out the $k=0$ term, which is $\Phi_{\rm V}$, leaves $\frac{5/3}{5/3+kn} = \frac{a}{a+k}$.

(3) The **Pochhammer symbol** is the rising product $(x)_k = x(x+1)\cdots(x+k-1)$, with $(x)_0 = 1$. In $\frac{(a)_k}{(1+a)_k} = \frac{a(a+1)\cdots(a+k-1)}{(a+1)\cdots(a+k)}$ every factor but the first on top and the last below cancels, leaving $\frac{a}{a+k}$. The **hypergeometric function** is the series ${}_2F_1(\alpha,\beta;\gamma';z) = \sum_k \frac{(\alpha)_k(\beta)_k}{(\gamma')_k\,k!}z^k$; with $\alpha = 1$, $(1)_k = k!$ cancels the $k!$, and with $\beta = a$, $\gamma' = 1+a$ the series is $\sum_k \frac{a}{a+k}z^k$.

The geometric series converges only for $\epsilon < 1$, yet the phase at lower frequencies, where friction dominates, is a perfectly finite integral. The same substitution that turned the series into ${}_2F_1$

shows that nothing is lost. Put $t = (f/f')^n$ in the exact integral, with $z = -c_f f^{-n}$:

$$\frac{K}{2\pi}\Phi(f) \overset{(1)}{=} \int_f^\infty \frac{f'^{-8/3}\,\mathrm{d}f'}{1 + c_f f'^{-n}} \overset{(2)}{=} \frac{f^{-5/3}}{n}\int_0^1 \frac{t^{a-1}\,\mathrm{d}t}{1-zt} \overset{(3)}{=} \frac{3}{5} f^{-5/3}\, a\int_0^1 \frac{t^{a-1}\,\mathrm{d}t}{1-zt}.$$

(1) The definition with $\dot f = K f'^{11/3}(1 + c_f f'^{-n})$.
(2) $f' = f t^{-1/n}$, $\mathrm{d}f' = -(f/n)t^{-1/n-1}\mathrm{d}t$, and $c_f f'^{-n} = c_f f^{-n} t = -zt$; the powers of t combine to $t^{(8/3-1)/n-1} = t^{a-1}$, and f' from f to ∞ is t from 1 to 0.
(3) $1/n = \frac{3}{5}a$.

Expanding $1/(1-zt) = \sum_k (zt)^k$ and using $a\int_0^1 t^{a-1+k}\mathrm{d}t = a/(a+k)$ returns exactly the series of (T2.96), so the last integral *is* ${}_2F_1(1, a; 1+a; z)$ where the series converges. But the integral is finite and smooth for every $z \le 0$, since $1 - zt \ge 1$, so it defines the function at every frequency; this is Euler's integral for the hypergeometric function, the standard way of continuing it beyond $|z| < 1$. Hence (T2.96) holds at all frequencies [89, eq. (12)]. At low frequency, $f \to 0$, $z \to -\infty$ and the integral goes to zero: friction then sweeps the binary through each frequency so fast that it accumulates almost none of the vacuum phase, and $\Delta\Phi \to \Phi_{\mathrm{V}}$.

T2.10.3 The experiment: dephasing for several spike slopes

Their question

Coogan et al. asked whether LISA could tell an intermediate-mass-ratio inspiral wrapped in a dark-matter spike from the same binary in vacuum, and whether the signal would then measure the spike's density and slope [89]. Earlier estimates had used a fixed spike and found enormous dephasings. Their point was that the friction heats the spike it feeds on, so the honest question is what survives once the spike is allowed to respond. We reproduce the static part of their calculation, which is what the equations of this section describe, and use it to see where the static picture stops.

Their setup

The binary is a $10^3\,M_\odot$ black hole with a $1.4\,M_\odot$ companion, the IMRI of Section T2.4. Around the heavy hole sits a spike $\rho = \rho_6(r_6/r)^\gamma$ whose density at the pivot radius $r_6 = 10^{-6}$ pc is fixed to their benchmark value (Section T2.9 derived it from $\rho_{\mathrm{sp}} = 226\,M_\odot\,\mathrm{pc}^{-3}$). The companion feels Chandrasekhar's friction (T2.87), with the fraction ξ of slow particles appropriate for a 7/3 spike and their logarithm $\ln\sqrt{m_1/m_2}$. We keep ρ_6 fixed and vary the slope, to see how the dephasing depends on γ, and we look at the last four years before the ISCO, a nominal LISA mission.

quantity	meaning	value
m_1, m_2	central hole, companion	$10^3\,M_\odot$, $1.4\,M_\odot$
ρ_6	spike density at r_6	$5.45\times10^{15}\,M_\odot\,\mathrm{pc}^{-3}$
r_6	pivot radius	10^{-6} pc
ξ	fraction of slower particles	0.58
$\ln\Lambda$	Coulomb logarithm, $\ln\sqrt{m_1/m_2}$	3.286
γ	spike slope	1.5, 2, 7/3, 2.5
window	last four years before the ISCO	from 21.7 mHz

Their equations in ours

Everything needed has been derived above, and we kept their symbols wherever they had one: ρ_6, r_6, c_f and $f_{\rm eq}$ mean the same here as there, and f is in both cases the gravitational-wave frequency.

Coogan et al.	here	content
eq. (12)	(T2.96)	exact phase to coalescence in a static spike
eq. (15)	$f_{\rm eq} = c_f^{1/n}$, with (T2.93)	where friction and radiation are equal
eq. (16)	(T2.95)	high-frequency power law of the dephasing

The one convention to watch is which phase is quoted: their dephasing is the time-domain $\Delta\Phi$, not the Fourier $\delta\Psi$ of the Fisher matrix, and the two differ by the factor $-1/(n+\frac{8}{3})$ derived above.

What we compute

The script `05_dephasing.py` evaluates these formulas for the four slopes:

Listing T2.7: The friction-to-radiation ratio and the exact dephasing (`code/chT2/05_dephasing.py`)

```
36 def c_f(gam):
37     """eps(f) = P_DF/P_GW = c_f f^{-n} (SI evaluation of the text formula)."""
38     rho_SI = rho6 * MSUN / PC**3 * (r6 * PC) ** gam      # rho = rho_SI r^-gam (SI)
39     pref = 5*np.pi * c**5 * xi * lnL / (8 * G**2.5 * (M*MSUN)**1.5 * (m1*MSUN)**2)
40     n = (11 - 2 * gam) / 3
41     return pref * rho_SI * (G * M * MSUN / np.pi**2) ** ((11 - 2 * gam) / 6), n
42
43 def phi_vac(f):
44     """GW phase left to coalescence, Phi_V = (1/16) (pi G Mc f/c^3)^{-5/3}."""
45     return (np.pi * Mc * TSUN * f) ** (-5 / 3) / 16
46
47 def dphi_closed(f, gam):
48     """Phi_V - Phi_S, static dress: Phi_S = Phi_V 2F1(1,a;1+a;-c_f f^-n), a=5/3n."""
49     cf, n = c_f(gam)
50     b = 5 / (3 * n)
51     return phi_vac(f) * (1 - hyp2f1(1, b, 1 + b, -cf * f ** (-n)))
```

It makes four moves. First, lines 38–39 compute the constants of (T2.93): `rho_SI` is $\rho_6 r_6^\gamma$ in SI units, and `pref` is the prefactor of step (2) of (T2.92). Second, line 41 completes c_f by writing the radius through Kepler's law, the factor $(GM/\pi^2)^{(11-2\gamma)/6}$, so that $\epsilon = c_f f^{-n}$ is a function of the frequency alone. Third, lines 45 and 51 evaluate the vacuum phase $\Phi_{\rm V} = \frac{1}{16}(\pi G\mathcal{M} f/c^3)^{-5/3}$ and the exact dephasing $\Phi_{\rm V}(1-{}_2F_1)$ of (T2.96); the variable `b` on line 50 is the parameter $a = 5/(3n)$, and `dphi_closed` is $\Delta\Phi$. Fourth, the script integrates $2\pi f/\dot f$ directly, with no expansion and no hypergeometric function, and compares: the two agree to 1.343×10^{-13}. That check tests the algebra from ϵ to Φ, the series, the substitution and the code; it says nothing about whether the static-spike physics is right. The dephasing plotted is accumulated between f and the ISCO, as in [86, eq. (49)].

Reading the figure

Figure T2.21 has one idea: the dephasing is a broken power law, with the break at $f_{\rm eq}$. For $\gamma = 7/3$ friction and radiation balance at $f_{\rm eq} = 16.75\,$mHz. Below $f_{\rm eq}$ friction drives the inspiral, almost all of the vacuum phase is lost ($\Delta\Phi \to \Phi_{\rm V}$, as found above), and the local slope of $\Delta\Phi$

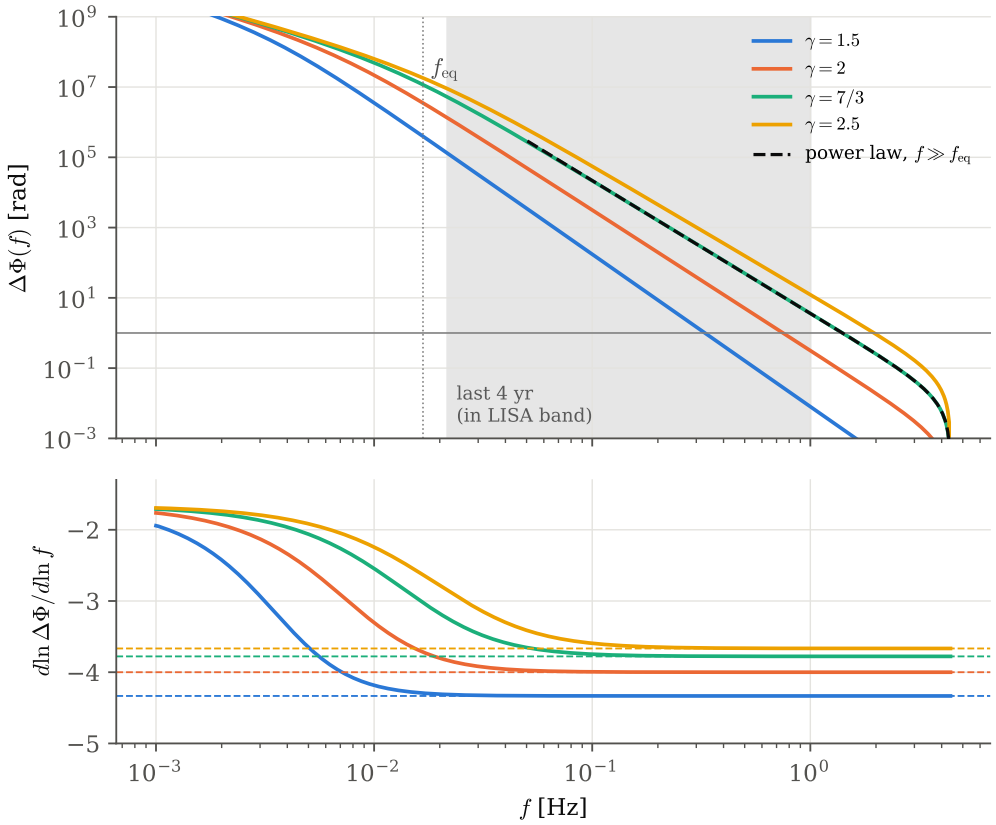


Figure T2.21: Dephasing of the $10^3 + 1.4\,M_\odot$ IMRI by friction in a static spike of slope γ, at fixed density ρ_6 at 10^{-6} pc. Top: phase lost between f and the ISCO; the dashed black line is the power law (T2.95) for $\gamma = 7/3$, drawn where $f > 3f_{\rm eq}$, and the grey band is the last four years. Bottom: the local slope $\mathrm{d}\ln\Delta\Phi/\mathrm{d}\ln f$ of the phase lost before coalescence. It moves from the vacuum value $-5/3$ at low frequency to the predicted $-5/3 - n$ (dashed, one per slope) above $f_{\rm eq}$.

is the vacuum $-5/3$. Above it each curve settles onto the predicted exponent $b = -5/3 - n$ of (T2.95): the measured local slopes are -4.333, -4, -3.778 and -3.667 for $\gamma = 1.5$, 2, 7/3, 2.5, against $-13/3$, -4, $-34/9$ and $-11/3$.

How large

The steeper the spike, the more phase it removes at fixed ρ_6: over the last four years before the ISCO (from 21.7 mHz), $\Delta\Phi = 1.302 \times 10^5$, 1.325×10^6, 5.191×10^6 and 8.937×10^6 rad for the four slopes. To give these a scale: a matched filter at signal-to-noise ratio ρ resolves a phase error of about $1/\rho$ radians, the weighted root-mean-square phase shift that the Fisher matrix can detect (Section 10.9.3), about a tenth of a radian for $\rho = 10$. So even the gentlest of these spikes would be glaringly visible in this idealised, static setting.

Paper and reproduction side by side

quantity ($\gamma = 7/3$)	Coogan et al.	ours (static)
$f_{\rm eq}$	$\approx 15\,$mHz [89, eq. (15)]	16.75 mHz
high-frequency slope of $\Delta\Phi$	$-34/9$ (static, eq. (16))	-3.778
exact vs. power law	eq. (12)	closed form = numerical to 1.343×10^{-13}
$\Delta\Phi$, last four years	much smaller, spike with feedback (Sec. IV)	5.191×10^6 rad

Why they differ

The equality frequency. Ours is 12% above theirs. Since $f_{\rm eq} = c_f^{1/n}$ and $c_f \propto \xi \ln\Lambda\, \rho_6$, $f_{\rm eq} \propto (\xi \ln\Lambda\, \rho_6)^{1/n} = (\xi \ln\Lambda\, \rho_6)^{9/19}$ for $\gamma = 7/3$. A 12% shift in $f_{\rm eq}$ therefore corresponds to a $(1.12)^{19/9} - 1 \approx 26\%$ difference in the product $\xi \ln\Lambda\, \rho_6$. The prescriptions differ by more than that: setting $\xi = 1$ instead of 0.58 alone would move $f_{\rm eq}$ up by $(1/0.58)^{9/19} \approx 1.29$, and the two common choices of $\ln\Lambda$ differ by a factor of two. We have not pinned the remaining 12% on a single choice; it is the size expected from these conventions, and the exponents, which do not depend on them, agree.

Static against responding spike. This is the large difference. The energy friction removes from the orbit goes into the dark matter, and a rough count shows it is far more than the spike can absorb. Take the orbit at $f_{\rm eq}$, of radius $r = (GM/\pi^2 f_{\rm eq}^2)^{1/3} \approx 1.2 \times 10^{-8}\,$pc. The energy friction removes between $f_{\rm eq}$ and the ISCO is the friction share of the orbital energy lost,

$$E_{\rm DF} = \int_{f_{\rm eq}}^{f_c} \frac{\epsilon}{1+\epsilon} \left|\frac{{\rm d}E}{{\rm d}f}\right| {\rm d}f = \frac{\mu}{3}\left(\frac{\pi G M f_{\rm eq}}{c^3}\right)^{2/3} c^2 \int_1^{f_c/f_{\rm eq}} \frac{y^{-1/3}\,{\rm d}y}{1+y^n} \approx 1 \times 10^{-3}\, M_\odot c^2,$$

with $y = f/f_{\rm eq}$, $\epsilon = y^{-n}$, $|{\rm d}E/{\rm d}f|$ from (T2.41), and the last integral ≈ 0.51. The dark matter inside that orbit is $M_{\rm DM}(<r) = 4\pi\rho_6 r_6^\gamma r^{3-\gamma}/(3-\gamma) \approx 5 \times 10^{-3}\, M_\odot$, bound to the hole by roughly $GMM_{\rm DM}/r \approx 2 \times 10^{-5}\, M_\odot c^2$. The friction deposits some fifty times the binding energy of the dark matter it moves through. The spike near the orbit cannot stay as it was: it is heated and partly unbound, its density at a given radius falls during the inspiral, ϵ at fixed f falls with it, and the dephasing is smaller and its break moves. That is the qualitative result Coogan et al. find when they follow this feedback: a much smaller dephasing, still a broken power law, but with a different break frequency and exponents [89, Sec. IV]. Kim et al. mimic the same effect by capping the friction at the binding energy of a shell of halo [86, Sec. IV C].

The chirp itself. Our vacuum chirp is Newtonian and circular. Near the ISCO, where $v \approx 0.4c$, post-Newtonian corrections to $\dot f$ are large (Section T2.6), and relativistic corrections to the spike and to the friction push the other way from feedback, increasing the dephasing [88]. These change the numbers, not the exponents of (T2.94), which survive as the leading-order description of each effect.

What it taught us

The power law is not the whole story even before feedback. LISA's window is not far above $f_{\rm eq}$: at the start of the four-year window the power law misses the exact $\Delta\Phi$ for $\gamma = 7/3$ by 35.35%, and for $\gamma = 2.5$ the window starts *below* $f_{\rm eq} = 24.26\,$mHz, where no expansion in ϵ holds. A ppE forecast is reliable when the observed band lies well above $f_{\rm eq}$, that is, for weak

environments, which are exactly the ones whose detectability is in question; and a static spike strong enough to matter is one that the binary itself would destroy.

Taxonomy: Environmental effects as power laws in the phase

effect	$P_{\rm env}/P_{\rm GW} \propto$	n	$b = -\frac{5}{3} - n$	at $\gamma = \frac{7}{3}$
friction, spike $\rho \propto r^{-\gamma}$	$r^{11/2-\gamma}$	$\frac{11-2\gamma}{3}$	$\frac{2\gamma-16}{3}$	$-\frac{34}{9}$
friction, uniform density	$r^{11/2}$	$\frac{11}{3}$	$-\frac{16}{3}$ (−5.5PN)	—
accretion, collisionless	$r^{9/2-\gamma}$	$\frac{9-2\gamma}{3}$	$\frac{2\gamma-14}{3}$	$-\frac{28}{9}$
accretion, gas (BHL)	as friction			
pull of enclosed mass (no loss)	$r^{3-\gamma}$ (relative)	$\frac{2(3-\gamma)}{3}$	$\frac{2\gamma-11}{3}$	$-\frac{19}{9}$
dipole radiation (Brans–Dicke)	r	$\frac{2}{3}$	$-\frac{7}{3}$	—
vector ultralight DM	resonant: a step near one frequency, not a power law [100]			

The table also marks where the power-law picture stops. The oscillating potential of a vector dark-matter halo (Section T2.9) does nothing on average until the orbital frequency crosses a resonance with the field, and then shifts the phase within a narrow band [100, Sec. V]. A template βu^b cannot represent such a step; it needs its own model. The ppE parametrisation is a statement about smooth, monotonic effects, and its hidden assumption is that the environment changes slowly with radius.

Where each item is used. The waveform (T2.56) with phase (T2.54) and the correction (T2.94), together with the noise (T2.77), are the inputs of the gravitational-wave Fisher forecast in Chapter 10, which computes the error on β for a list of exponents b and checks it with the sampler of Chapter 9. The density profiles and the parametrisation by ρ_6 fix the physical meaning of β there. The likelihood $-\frac{1}{2}(d-h\,|\,d-h)$ with this S_n is the one derived in Chapter 0, and comparing a vacuum template with a dressed one is a Bayesian model comparison of the kind treated in Chapter 5.

Recap

- Energy balance with the quadrupole formula gives the chirp $\dot f = \frac{96}{5}\pi^{8/3}(G\mathcal{M}/c^3)^{5/3}f^{11/3}$; only the chirp mass enters at leading order, and LISA's benchmark inspirals complete millions of cycles in four years.
- The stationary-phase approximation gives $\tilde h = \mathcal{A} f^{-7/6} e^{-i\Psi(f)}$ with $\mathrm{d}^2\Psi/\mathrm{d}f^2 = 2\pi/\dot f$: anything that changes the sweep rate changes the phase.
- LISA's sensitivity $S_n(f)$ combines test-mass acceleration noise (low f), optical metrology noise and the arm response (high f), and the Galactic confusion foreground (around a millihertz); the signal-to-noise ratio is $\int (h_c/h_n)^2 \mathrm{d}\ln f$.
- Slow growth of a black hole compresses a cusp $r^{-\gamma_i}$ into a spike $r^{-(9-2\gamma_i)/(4-\gamma_i)}$, $r^{-7/3}$ for NFW, raising the density at the orbit by some eighteen orders of magnitude; ultralight dark matter falls freely inside the self-gravity radius, $\rho \propto r^{-3/2}$.
- An extra power $P_{\rm env} = c_f f^{-n} P_{\rm GW}$ shifts the phase by βu^b, $b = -5/3 - n$, valid for $f \gg f_{\rm eq}$; resonant effects are not power laws.

T2.11 Type Ia supernovae: candles that need a correction

A white dwarf is the burnt-out core of a star like the Sun, held up by the pressure of its degenerate electrons. If it gains mass from a companion, or merges with a second white dwarf, it can approach the Chandrasekhar limit of about $1.4\,M_\odot$, ignite carbon in its centre and blow itself apart. The explosion makes a few tenths of a solar mass of radioactive nickel-56, and the decay of that nickel powers the light we see for weeks. The fuel and the trigger are nearly the same every time, so the explosions are nearly the same every time. At its peak a type Ia supernova outshines its whole host galaxy, and it can be seen across half the observable universe.

"Nearly the same" is not good enough for cosmology. Raw peak magnitudes scatter by about 0.4 mag [103, p. 33], which is a 40% spread in luminosity and a 20% spread in the distance one would infer. Two observations rescue the method. Supernovae that are brighter at peak also fade more slowly, and supernovae that look redder are fainter. Both properties can be read off the light curve. A correction built from them shrinks the scatter to about 0.12 mag. Such an object is called a **standardisable candle** [its luminosity is not fixed, but it can be predicted from properties we measure on the same object].

The question. Given the light curve of one supernova in a few colour bands and the redshift of its host galaxy, what distance does it imply, how uncertain is that distance, and which statistical tools turn a few hundred such light curves into a measurement of the expansion history?

T2.11.1 What is measured, and the standardisation

A magnitude is a logarithm of a flux F: $m = -2.5\log_{10} F + \text{const}$, so 5 magnitudes is a factor of 100 in flux and a larger magnitude means a fainter object. The **distance modulus** of Chapter 3 links the apparent magnitude m to the absolute magnitude M (the magnitude at 10 pc) through the luminosity distance D_L of Chapter T1, the distance for which the flux is $L/4\pi D_L^2$:

$$\mu \equiv m - M = 5\log_{10}\frac{D_L}{10\,\mathrm{pc}} = 5\log_{10}\frac{D_L}{\mathrm{Mpc}} + 25. \tag{T2.97}$$

A light-curve fitter such as SALT [103, p. 39] compares the observed fluxes in each band, night after night, with a family of template light curves. It returns three numbers per supernova: the peak apparent magnitude m_B in the rest-frame B band, a **stretch** x_1 [how much wider than average the light curve is in time, zero for a typical one], and a **colour** c [how much redder than average the supernova is at peak, in magnitudes of $B - V$]. Figure T2.22 shows the chain from light curve to distance.

Working definition T2.4 (Tripp standardisation)

The observed distance modulus of a type Ia supernova is

$$\mu_{\mathrm{obs}} = m_B - M + \alpha\, x_1 - \beta\, c, \tag{T2.98}$$

with M, α and β the same for every supernova. *Operationally:* fit the light curve to get (m_B, x_1, c); add αx_1 because broad supernovae are brighter than average (a smaller M); subtract βc because red ones are fainter. Typical values are $\alpha \approx 0.15$ and $\beta \approx 3$.

Why do the two corrections work? The amount of nickel-56 sets both the peak luminosity and the width. More nickel means a hotter, more opaque ejecta, so the light takes longer to diffuse out, and the curve is both brighter and broader [103, p. 33]. This is the Phillips relation. Colour

has two sources. Dust in the host galaxy absorbs blue light more than red, so it reddens and dims together. Milky-Way dust would give $\beta \approx R_B = R_V + 1 \approx 4.1$, with $R_V \equiv A_V/E(B-V) = 3.1$. Here A_B and A_V are the dimming by dust in the two bands, and the reddening is their difference, $E(B-V) = A_B - A_V$. The colour correction asks how much dimming in B comes with one magnitude of reddening, $R_B \equiv A_B/E(B-V) = (A_V + E(B-V))/E(B-V) = R_V + 1$. Fits to supernovae return a smaller β, as if R_V were 1.5–2.5 [103, p. 36]. Part of the colour spread is therefore intrinsic to the explosions and not dust at all. That is why β is fitted rather than taken from the physics of dust.

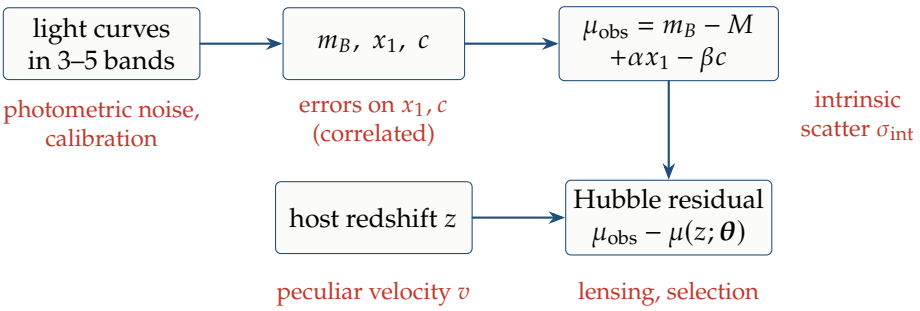


Figure T2.22: From light curves to a distance. Blue boxes are the steps of the analysis; red labels mark where noise enters. The fitted shape and colour turn the peak magnitude into a distance modulus, and the redshift of the host predicts that modulus in a cosmological model. Their difference, the Hubble residual, is the quantity whose sampling distribution every later fit needs.

T2.11.2 Where the noise lives

The datum that enters every fit is the **Hubble residual** $\Delta\mu = \mu_{\rm obs} - \mu(z;\boldsymbol{\theta})$, the difference between the standardised modulus and the one predicted at the host redshift by a model with parameters $\boldsymbol{\theta}$. Its scatter has four parts. The *intrinsic scatter* $\sigma_{\rm int} \approx 0.1$ mag is what the standardisation cannot remove. The *measurement errors* of m_B, x_1 and c enter through (T2.98), so an error σ_c in colour costs $\beta\sigma_c$ in μ. The *peculiar velocity* v of the host [its motion relative to the smooth expansion] shifts the redshift, and therefore the predicted modulus. And at high redshift, *gravitational lensing* by structure along the line of sight magnifies or demagnifies the supernova.

How big is the peculiar-velocity term? At low redshift $D_L \approx cz/H_0$, and (T2.97) turns a fractional error in distance into an additive error in magnitude:

$$\delta\mu \overset{(1)}{=} \frac{5}{\ln 10}\frac{\delta D_L}{D_L} \overset{(2)}{=} \frac{5}{\ln 10}\frac{\delta(cz)}{cz} \overset{(3)}{=} \frac{5}{\ln 10}\frac{v}{cz}, \qquad \sigma_{\mu,\rm pec} = 2.17\,\frac{\sigma_v}{cz}. \tag{T2.99}$$

(1) Differentiate $\mu = 5\log_{10} D_L + \text{const} = (5/\ln 10)\ln D_L + \text{const}$.
(2) $D_L \propto cz$ at low redshift, so the fractional changes are equal.
(3) A line-of-sight velocity v adds v to the measured cz. The number in front is $5/\ln 10 = 5/2.303 = 2.17$.

For $\sigma_v = 250$ km/s this is 0.030 mag at $z = 0.06$, but 0.18 mag at $z = 0.01$. That is why the low-redshift anchor of a supernova Hubble diagram starts at $z \approx 0.02$–0.03, the "sweet spot" where peculiar velocities are small but the distances still do not depend on the dark-energy parameters [103, p. 33].

Example: what one supernova is worth. Solve (T2.97) for D_L and take logarithms:

$$\ln D_L \overset{(1)}{=} \frac{\ln 10}{5}(\mu - 25), \qquad \sigma_{\ln D} \overset{(2)}{=} \frac{\ln 10}{5}\sigma_\mu = \frac{\sigma_\mu}{2\times 1.086} \overset{(3)}{=} \frac{0.12}{2.17} = 0.055. \tag{T2.100}$$

(1) $D_L/\mathrm{Mpc} = 10^{(\mu-25)/5}$ and $\ln 10^x = x \ln 10$.
(2) A constant factor passes through the standard deviation. Weinberg et al. write $\ln 10/5$ as $1/(2 \times 1.086)$: $1.086 = 2.5/\ln 10$ converts magnitudes to natural logarithms, and the 2 converts flux to distance, because flux goes as D_L^{-2}.
(3) The standardised scatter $\sigma_\mu \approx 0.12\,\mathrm{mag}$.

Each well-measured supernova gives its distance to about 6%. Averaging N independent supernovae in a redshift bin divides this by $\sqrt{N}$, which is Weinberg et al.'s eq. (51) [103, p. 34]. How many supernovae does a 1% distance in one bin need? Set the error of the mean equal to the target and solve for N:

$$\frac{0.055}{\sqrt{N}} = 0.01 \implies N = \left(\frac{0.055}{0.01}\right)^2 = 5.5^2 \approx 30, \qquad \frac{0.055}{\sqrt{N}} = 0.005 \implies N = 11^2 \approx 120,$$

as long as nothing systematic is left. The Pantheon+ compilation has 1550 spectroscopically confirmed supernovae from $z = 0.001$ to 2.26 [26]. Its statistical errors are already comparable to its systematic ones, which come from flux calibration across telescopes, the treatment of dust and colour, and a possible drift of the population with redshift, such as the 0.03 mag per decade of host stellar mass found in the residuals [103, pp. 39–41].

T2.11.3 Fitting α and β: least squares with noisy regressors

Write $y_i \equiv m_{B,i} - \mu(z_i)$ for a low-redshift sample with known cosmology. Then (T2.98) says $y_i = M - \alpha x_{1,i} + \beta c_i + \epsilon_i$. This is a linear model, and least squares (Chapter 3) seems to give M, α, β at once. There is a catch. The regressors x_1 and c are themselves measured with noise. Ordinary least squares assumes the regressors are exact. What does it do when they are not?

Take the colour term alone: $y = \beta c + \epsilon$, with the true colours spread with variance V_c and measured as $\hat{c} = c + e$, where e is independent noise of variance σ_c^2. For a large sample the least-squares slope tends to

$$\hat{\beta} \overset{(1)}{=} \frac{\mathrm{Cov}(y, \hat{c})}{\mathrm{Var}(\hat{c})} \overset{(2)}{=} \frac{\mathrm{Cov}(\beta c + \epsilon,\ c + e)}{\mathrm{Var}(c) + \mathrm{Var}(e)} \overset{(3)}{=} \frac{\beta V_c}{V_c + \sigma_c^2} \equiv \beta \lambda_c, \qquad \lambda_c = \frac{V_c}{V_c + \sigma_c^2}. \tag{T2.101}$$

(1) The least-squares slope of y on $\hat{c}$ is the sample covariance over the sample variance, and both converge to their expectations (the law of large numbers).
(2) Substitute the model and the measurement. c and e are independent, so their variances add.
(3) ϵ and e are independent of everything else, so only $\mathrm{Cov}(\beta c, c) = \beta V_c$ survives.

The fitted slope is pulled towards zero by the factor $\lambda_c < 1$. Statisticians call this **regression dilution**, and λ_c the reliability ratio. The scatter left after the correction follows from the same quantities:

$$\mathrm{Var}(y - \hat{\beta}\hat{c}) \overset{(1)}{=} \mathrm{Var}\left(\beta(1-\lambda_c)c - \beta\lambda_c e + \epsilon\right) \overset{(2)}{=} \beta^2\left[(1-\lambda_c)^2 V_c + \lambda_c^2 \sigma_c^2\right] + \sigma_\epsilon^2 \overset{(3)}{=} \beta^2 \lambda_c\, \sigma_c^2 + \sigma_\epsilon^2. \tag{T2.102}$$

(1) Substitute $y = \beta c + \epsilon$, $\hat{c} = c + e$ and $\hat{\beta} = \beta\lambda_c$.
(2) The three terms are independent, so their variances add.
(3) $1 - \lambda_c = \sigma_c^2/(V_c + \sigma_c^2)$, so $(1-\lambda_c)^2 V_c + \lambda_c^2 \sigma_c^2 = \dfrac{\sigma_c^4 V_c + V_c^2 \sigma_c^2}{(V_c + \sigma_c^2)^2} = \lambda_c \sigma_c^2$.

Here σ_ϵ^2 collects every other source: σ_{int}^2, the error on m_B and the peculiar velocities of (T2.99).

Watch out: A fitted slope is not the physical slope

With noisy regressors, least squares underestimates $|\beta|$ by the factor λ_c of (T2.101), however many supernovae are used. A diluted β is then applied to supernovae at high redshift, whose colours are noisier, so the bias does not cancel between the two ends of the Hubble diagram. Modern analyses treat the true x_1 and c of each supernova as unknown latent parameters with a population distribution, a hierarchical model of the kind built in Chapter 5.

The experiment. `06_sn_standardise.py` draws 400 supernovae at $0.02 < z < 0.1$ with $\alpha = 0.15$, $\beta = 3$, an intrinsic scatter of 0.1 mag, true colours of spread 0.1 measured with errors 0.03, and peculiar velocities of 250 km/s:

Listing T2.8: A mock low-redshift supernova sample and the least-squares standardisation (`code/chT2/06_sn_standardise.py`)

```
42 def mock(rng):
43     z = rng.uniform(0.02, 0.10, N)
44     x1 = rng.normal(0, SIG_X1, N)
45     c = rng.normal(0, SIG_C, N)
46     Mtrue = M0 - ALPHA * x1 + BETA * c + rng.normal(0, SIG_INT, N)
47     zobs = z + SIG_V / C_KMS * rng.normal(0, 1, N) * (1 + z)    # peculiar velocity
48     mB = Mtrue + mu_of_z(z) + rng.normal(0, ERR_M, N)
49     return zobs, mB, x1 + rng.normal(0, ERR_X1, N), c + rng.normal(0, ERR_C, N)
50
51
52 def fit(zobs, mB, x1, c):
53     """Least squares for y = m_B - mu(z) = M - alpha x1 + beta c."""
54     y = mB - mu_of_z(zobs)
55     A = np.column_stack([np.ones_like(y), -x1, c])
56     th, *_ = np.linalg.lstsq(A, y, rcond=None)
57     return th, y, y - A @ th
```

Line 46 is (T2.98) read backwards: the true absolute magnitude of each supernova. Line 47 moves the redshift by the peculiar velocity, and line 49 adds measurement noise to the regressors, which is what causes the dilution. The fit in lines 54–56 is ordinary least squares on the design matrix $[1, -x_1, c]$.

Figure T2.23 shows the result. Before any correction, the residual $m_B - \mu(z) - M$ carries the whole spread of the stretch and colour terms as well as the noise. Write V_x for the spread (variance) of the true stretches, as V_c is for the colours, and σ_m, $\sigma_{\rm pec}$ for the magnitude error and the peculiar-velocity term of (T2.99). All the terms are independent, so their variances add:

$$\sigma^2_{\rm raw} = \sigma^2_{\rm int} + \alpha^2 V_x + \beta^2 V_c + \sigma^2_m + \sigma^2_{\rm pec}.$$

The raw residuals scatter by 0.344 mag, against 0.353 from this sum. After standardisation they scatter by 0.142 mag, against the prediction 0.140 of (T2.102). Over 2000 mock samples the fitted β averages 2.754 ± 0.071, not 3: it agrees with the diluted value $\beta\lambda_c = 2.752$ ($\lambda_c = 0.917$). The stretch has its own reliability ratio, built as in (T2.101) from its spread and its measurement error σ_x, $\lambda_x = V_x/(V_x + \sigma_x^2)$. It is measured more precisely relative to its spread ($\lambda_x = 0.978$), and $\hat\alpha$ averages 0.1467 ± 0.0002, again $\alpha\lambda_x = 0.1467$ within its Monte Carlo error.

Where the statistics comes back in. The datum is the model's prediction plus one realisation of the noise, $\mu_{\rm obs} = \mu(z;\boldsymbol{\theta}) + \Delta\mu$ with $\Delta\mu \sim \mathcal{N}(0, \sigma^2)$. For a trial $\boldsymbol{\theta}$ the likelihood asks how plausible the particular noise realisation is that would turn that model's prediction into the magnitudes actually observed; a $\boldsymbol{\theta}$ that needs a large, improbable $\Delta\mu$ in every supernova

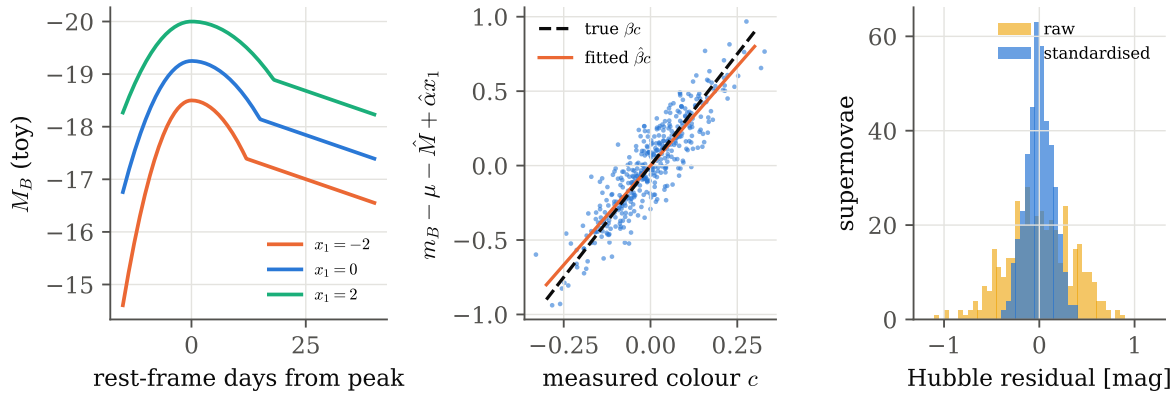


Figure T2.23: Standardising type Ia supernovae. Left: toy light curves; the broad one ($x_1 = 2$) is brighter, the narrow one fainter (the trend is exaggerated for visibility). Middle: after the stretch correction, the residual magnitude grows linearly with the measured colour; the fitted slope (orange) is shallower than the true one (dashed) because the colours are noisy. Right: Hubble residuals before (yellow) and after (blue) the correction.

is disfavoured. The residual $\Delta\mu$ is Gaussian in magnitudes, so the likelihood of Chapter 3 with the Hubble law, and the H_0 estimator built there, apply directly. Fitting α, β, M and the cosmology together is a linear-in-nuisance problem, solved by the least squares of Chapter 3 or marginalised in Chapter 5. Correlated systematics (a calibration offset shared by all supernovae from one telescope) enter as a full covariance matrix $\mathbf{C}$, whose use in the likelihood and in the Fisher matrix is the subject of Chapter 10. The hierarchical treatment of latent colours is sampled with the MCMC of Chapter 9.

T2.12 The distance ladder: why H_0 needs an anchor

The supernovae of Section T2.11 measure *relative* distances. Their brightnesses might tell us that one is 1.8 times farther away than another, while leaving both distances in megaparsecs undetermined. We encountered this ambiguity in the likelihood calculation of Chapter 3. In the Hubble law with supernova magnitudes, the absolute magnitude M and $5\log_{10} H_0$ enter only in the combination $M - 5\log_{10} H_0$, so the Fisher matrix of the two is singular. Brighter candles in a faster-expanding universe look exactly like fainter candles in a slower one. To measure H_0 someone has to tell us M, and the only way to do that is to find supernovae in galaxies whose distances are known by other means. Those distances in turn need calibrating, and the chain ends only at geometry. This chain is the **distance ladder**.

Where we have met H_0 before. In Section 3.17 we fitted mock surveys of 100 supernovae made with $H_0 = 70$ and an absolute magnitude $M = -19.25$ that we simply declared known, and the maximum-likelihood estimator, the weighted mean of $5\log_{10} H_0$, found H_0 to about 0.5 km/s/Mpc. Section 3.17.4 then freed M, saw the Fisher matrix of $(M, 5\log_{10} H_0)$ turn singular, and closed the gap by handing the fit one more number, $M_{\rm cal} \pm 0.03$ mag, from "the calibrators"; the calibration error then dominated, and the scatter of $\hat{H}_0$ grew to 1.1 km/s/Mpc. Here the parameter, the magnitude likelihood and the degeneracy are the same, and so is the estimator, now written as generalised least squares. What changes is where M comes from: instead of arriving as one number with an assumed error, it is produced by fitting the rungs below, so its error is assembled from parallaxes, Cepheids and calibrator supernovae, and we can ask which rung pays for most of it. The data are still mock, made with $H_0 = 73$ to sit near the published value, which is met in Section T2.12.3. The two error bars come out almost equal, 1.1 km/s/Mpc in both, because the ladder below delivers M to 0.03 mag, the precision the

worked problem assumed.

The question. Which measurements make up the ladder, what does each one measure and with what noise, and how do the errors of the rungs combine into the error on H_0?

T2.12.1 The rungs

Rung 1, geometry. The **parallax** p of a star is the half-angle through which it appears to move as the Earth goes round the Sun. Its distance is $d = 1\,\mathrm{AU}/\tan p$, that is $d = 1\,\mathrm{pc} \times (1''/p)$ for small angles [104, Sec. 8]. The Gaia satellite measures parallaxes to some 20 μas, so a Cepheid at 2 kpc ($p = 0.5$ mas) has its distance to 4%. Two other geometric anchors reach farther: detached eclipsing binaries in the Large Magellanic Cloud (the stars' sizes from the eclipse timing, their angular sizes from their surface brightness) and the water masers orbiting the black hole in NGC 4258 (a Keplerian disc whose velocity and proper motion give its physical and angular size). Both give distances to about 1–1.5% [24].

Rung 2, a stellar standard candle. **Cepheids** are pulsating giant stars whose brightness varies with periods of days to months. Brighter Cepheids pulsate more slowly. The **Leavitt law** [the period–luminosity relation of Cepheids, a straight line in $\log P$] says

$$m_W = \mu + M_W + b_W\,(\log_{10} P - 1), \tag{T2.103}$$

where m_W is a reddening-free combination of magnitudes in two bands (a "Wesenheit" magnitude), M_W the absolute magnitude of a 10-day Cepheid, and $b_W \approx -3.3$ the slope. The scatter about the line is about 0.25 mag per star, but a galaxy contains dozens of Cepheids, so its distance modulus μ is found to a few hundredths of a magnitude. The alternative second rung is the **tip of the red giant branch** (TRGB): the brightest red giants are those about to ignite helium in their cores, and the ignition happens at a fixed core mass, so the tip of the giant branch is a standard candle with $M_I \approx -4.05$. Both second rungs are reviewed in [105, Sec. IV.A].

Rung 3, supernovae in calibrator galaxies. A few dozen galaxies within ~ 40 Mpc have both Cepheids (or a measured TRGB) and a well-observed type Ia supernova. In each of them the Cepheids give μ, and the supernova gives m_B, so $M_B = m_B - \mu$. SH0ES uses 42 supernovae in 37 such hosts [24].

Rung 4, the Hubble flow. Supernovae at $0.023 < z < 0.15$ are far enough that peculiar velocities matter little ((T2.99)) and near enough that the distance depends on the cosmology only through the deceleration parameter q_0 [$q_0 = -\ddot{a}a/\dot{a}^2$ today, about -0.55 in ΛCDM]. What fixes that dependence? Expand the comoving distance $D_C = c\int_0^z \mathrm{d}z'/H(z')$ to second order with $H(z) \approx H_0[1 + (1 + q_0)z]$, and use $D_L = (1 + z)D_C$ for a flat universe [104, eqs. (15), (21)]:

$$\begin{aligned} D_L &\overset{(1)}{=} (1+z)\,\frac{c}{H_0}\int_0^z \left[1 - (1+q_0)z'\right]\mathrm{d}z' \\ &\overset{(2)}{=} (1+z)\,\frac{c}{H_0}\left[z - \frac{1+q_0}{2}z^2\right] \overset{(3)}{=} \frac{cz}{H_0}\left[1 + \frac{1-q_0}{2}z\right] + O(z^3). \end{aligned} \tag{T2.104}$$

(1) How fast does H change with redshift? Both H and z change with time, so divide their rates of change. The definition of q gives $\dot{H} = \ddot{a}/a - \dot{a}^2/a^2 = -H^2(1+q)$, and $1 + z = 1/a$ gives $\dot{z} = -\dot{a}/a^2 = -(1+z)H$. Then

$$\frac{\mathrm{d}H}{\mathrm{d}z} = \frac{\dot{H}}{\dot{z}} = \frac{-H^2(1+q)}{-(1+z)H} = \frac{H(1+q)}{1+z} \xrightarrow{z=0} (1+q_0)H_0,$$

so to first order $H(z) \approx H_0[1+(1+q_0)z]$, and $1/(1+\epsilon) \approx 1-\epsilon$ gives $1/H(z) \approx H_0^{-1}[1-(1+q_0)z]$.

(2) Integrate term by term.
(3) Multiply out: $(1+z)\big(z - \frac{1+q_0}{2}z^2\big) = z + z^2 - \frac{1+q_0}{2}z^2 + O(z^3)$.

What does a Hubble-flow supernova then tell us? Its apparent magnitude is its absolute magnitude plus its distance modulus (T2.97), and (T2.104) gives the distance. The logarithm of a product is a sum of logarithms, which splits the known part from the unknown one:

$$\begin{aligned} m_{B,i} &\overset{(1)}{=} M_B + 5\log_{10}\frac{D_L(z_i)}{\mathrm{Mpc}} + 25 \\ &\overset{(2)}{=} M_B + 5\log_{10}\left\{cz_i\left[1 + \frac{1-q_0}{2}z_i\right]\right\} + 25 - 5\log_{10}H_0 \overset{(3)}{=} M_B + x_i - a, \end{aligned} \tag{T2.105}$$

with $x_i \equiv 5\log_{10}\{cz_i[1+\frac{1}{2}(1-q_0)z_i]\} + 25$ and $a \equiv 5\log_{10}H_0$ (c in km/s and H_0 in km/s/Mpc, so that D_L is in Mpc).

(1) $m = M + \mu$, with μ from (T2.97).
(2) Substitute (T2.104) and use $\log(XY/H_0) = \log X + \log Y - \log H_0$.
(3) Name the pieces: x_i is computed from the measured redshift alone (with q_0 from the cosmology, a small correction at these redshifts); a is the one unknown that carries H_0.

Here the degeneracy of Chapter 3 is visible: only $M_B - a$ appears.

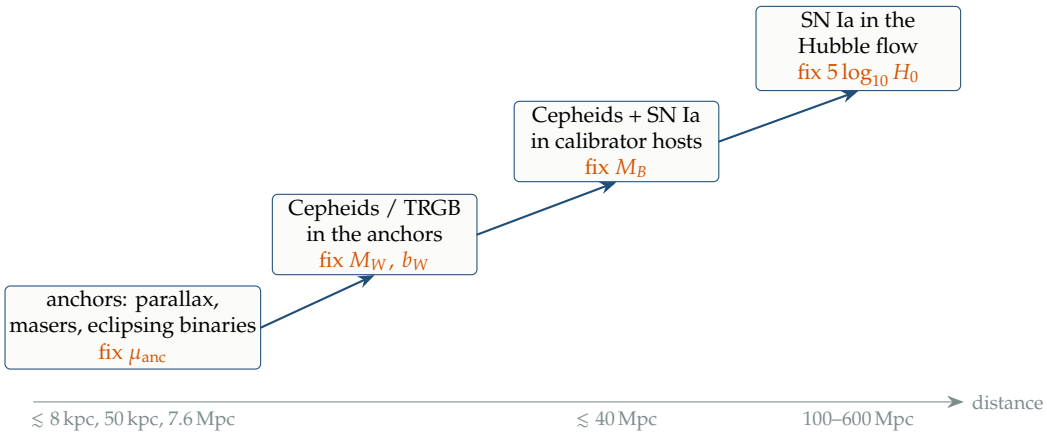


Figure T2.24: The distance ladder. Each rung measures distances to the objects of the next, and each pair of neighbouring rungs shares a parameter (orange): the anchor distances calibrate the Cepheid law, the Cepheids give the distances of the calibrator hosts and hence the supernova absolute magnitude M_B, and M_B turns Hubble-flow magnitudes into H_0. The grey scale gives the typical distances of each rung.

T2.12.2 One linear model for the whole ladder

Each measurement on every rung is a sum of unknowns plus noise. A geometric measurement says $\hat{\mu}_{\mathrm{anc}} = \mu_{\mathrm{anc}} + \text{noise}$. A Cepheid says (T2.103). A calibrator supernova says $m_B = \mu_{\mathrm{cal}} + M_B + \text{noise}$, and a Hubble-flow supernova says $m_B - x = M_B - a + \text{noise}$ by (T2.105). None of them multiplies two unknowns together, so the whole ladder is one linear model.

The smallest ladder first. Take two anchors, one calibrator host and one Hubble-flow supernova, with one Cepheid in each galaxy, and treat the Cepheid slope b_W as known, so that the term $b_W(\log_{10}P - 1)$ can be moved to the data side. The unknowns are the two anchor moduli, the calibrator modulus, the two absolute magnitudes and a, collected as $\boldsymbol{\theta} = (\mu_1, \mu_2, \mu_{\mathrm{cal}}, M_W, M_B, a)$. Write each measurement as one row: its coefficients on the

unknowns go into a row of the matrix A, and the measured number into $\boldsymbol{y}$,

$$\underbrace{\begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \\ m_{W,1} - b_W(\log_{10} P_1 - 1) \\ m_{W,2} - b_W(\log_{10} P_2 - 1) \\ m_{W,\mathrm{cal}} - b_W(\log_{10} P_{\mathrm{cal}} - 1) \\ m_{B,\mathrm{cal}} \\ m_{B,\mathrm{HF}} - x \end{pmatrix}}_{\boldsymbol{y}} = \underbrace{\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix}}_{A} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_{\mathrm{cal}} \\ M_W \\ M_B \\ a \end{pmatrix} + \boldsymbol{n}.$$

Read down a column to see how the rungs are tied. The column of M_W has a 1 in every Cepheid row, in the anchors and in the calibrator alike, so the anchors' geometry reaches the calibrator through it. The column of M_B has a 1 in both supernova rows, so the calibrator passes its information to the Hubble flow. The column of μ_{cal} joins the calibrator's Cepheid to its supernova. Only the last row involves a.

The real ladder is this pattern repeated: N_{anc} anchors and N_{cal} calibrator hosts, each with its own modulus column, dozens to hundreds of Cepheid rows per galaxy, one supernova row per calibrator, a row for each Hubble-flow supernova, and a seventh kind of column for b_W, whose entry in a Cepheid row is that star's $\log_{10} P - 1$. Stack all of them as $\boldsymbol{y} = A\boldsymbol{\theta} + \boldsymbol{n}$ with noise covariance $\mathbf{C}$. The generalised least-squares solution of Chapter 3 is

$$\hat{\boldsymbol{\theta}} = (A^{\mathsf{T}}\mathbf{C}^{-1}A)^{-1}A^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{y}, \qquad \mathrm{Cov}(\hat{\boldsymbol{\theta}}) = (A^{\mathsf{T}}\mathbf{C}^{-1}A)^{-1} = \mathbf{F}^{-1}, \tag{T2.106}$$

and the matrix $A^{\mathsf{T}}\mathbf{C}^{-1}A$ is exactly the Fisher matrix $\mathbf{F}$ of this linear Gaussian model (Chapter 10). The ladder is "one big fit": the information flows up the rungs through the shared parameters, and so do the errors.

How the errors add up. It helps to see the answer rung by rung before letting the matrix do it. The ingredients, one at a time: there are N_{anc} anchors, each with a geometric error σ_{geo} (in magnitudes of distance modulus) and n_A Cepheids; there are N_{cal} calibrator hosts, each with n_C Cepheids and one supernova; a single Cepheid scatters about the Leavitt law by σ_{ceph}, a single calibrator supernova by σ_{SN}; and there are N_{HF} Hubble-flow supernovae, each of scatter σ_{HF}. Take the Cepheid slope b_W as known (it is fixed by thousands of Cepheids, so its error is small) and follow the estimate up the ladder, one rung at a time.

One anchor. Average its n_A Cepheids, after removing the known slope term, and subtract the geometric modulus. By (T2.103) what is left is M_W:

$$\hat{M}_W^{(j)} = \overline{\left[m_W - b_W(\log_{10} P - 1)\right]}_j - \hat{\mu}_{\mathrm{anc},j}, \qquad \mathrm{Var}\,\hat{M}_W^{(j)} = \frac{\sigma_{\mathrm{ceph}}^2}{n_A} + \sigma_{\mathrm{geo}}^2,$$

because the mean of n_A independent Cepheids has variance $\sigma_{\mathrm{ceph}}^2/n_A$, and the geometric error is independent of the Cepheids. *Then all anchors.* The N_{anc} anchors are independent of one another, so their average $\hat{M}_W$ has $1/N_{\mathrm{anc}}$ of this variance. That is line 1 of (T2.107) below.

One calibrator. Its Cepheids give its modulus as $\hat{\mu}_{\mathrm{cal},i} = \overline{\left[m_W - b_W(\log_{10} P - 1)\right]}_i - \hat{M}_W$, and its supernova then gives

$$\hat{M}_B^{(i)} = m_{B,i} - \hat{\mu}_{\mathrm{cal},i} = \underbrace{m_{B,i} - \overline{\left[m_W - b_W(\log_{10} P - 1)\right]}_i}_{\text{private to host } i} + \underbrace{\hat{M}_W}_{\text{common to all hosts}}.$$

The private part has variance $\sigma_{\rm SN}^2+\sigma_{\rm ceph}^2/n_C$ and is independent from host to host. The common part is the *same* number in every host: if $\hat{M}_W$ came out 0.02 mag too bright, every calibrator inherits the same 0.02 mag. *Average the $N_{\rm cal}$ hosts.* The private errors average down, to $1/N_{\rm cal}$ of their variance. The common error does not: the average of $N_{\rm cal}$ copies of the same offset is that offset. So

$$\operatorname{Var}\hat{M}_B=\operatorname{Var}\left(\overline{\text{private}}\right)+\operatorname{Var}(\hat{M}_W)=\frac{1}{N_{\rm cal}}\left(\sigma_{\rm SN}^2+\frac{\sigma_{\rm ceph}^2}{n_C}\right)+\operatorname{Var}(\hat{M}_W),$$

the two parts adding because the hosts' own errors are independent of the anchors'. That is line 2.

The Hubble flow, the same way. By (T2.105) each Hubble-flow supernova gives $\hat{a}^{(k)}=\hat{M}_B-(m_{B,k}-x_k)$. The term in brackets is private, with variance $\sigma_{\rm HF}^2$; $\hat{M}_B$ is common to all of them. Averaging $N_{\rm HF}$ supernovae shrinks only the private part. Collecting the three rungs,

$$\begin{aligned}
\operatorname{Var}(\hat{M}_W)&\overset{(1)}{=}\frac{1}{N_{\rm anc}}\left(\sigma_{\rm geo}^2+\frac{\sigma_{\rm ceph}^2}{n_A}\right),\\
\operatorname{Var}(\hat{M}_B)&\overset{(2)}{=}\operatorname{Var}(\hat{M}_W)+\frac{1}{N_{\rm cal}}\left(\sigma_{\rm SN}^2+\frac{\sigma_{\rm ceph}^2}{n_C}\right),\\
\operatorname{Var}(\hat{a})&\overset{(3)}{=}\operatorname{Var}(\hat{M}_B)+\frac{\sigma_{\rm HF}^2}{N_{\rm HF}},\qquad \frac{\sigma_{H_0}}{H_0}\overset{(4)}{=}\frac{\ln 10}{5}\sqrt{\operatorname{Var}(\hat{a})}.
\end{aligned}\tag{T2.107}$$

(1) One anchor gives $\sigma_{\rm geo}^2+\sigma_{\rm ceph}^2/n_A$; $N_{\rm anc}$ independent anchors divide it by $N_{\rm anc}$.

(2) The calibrators' private errors average down; the error of $\hat{M}_W$ is common to all of them and passes through whole.

(3) The same split one rung higher: the error of $\hat{M}_B$ is common to all Hubble-flow supernovae.

(4) $H_0=10^{a/5}=e^{(\ln 10/5)a}$, so $\delta H_0/H_0=\delta\ln H_0=(\ln 10/5)\,\delta a$.

Every rung adds its own variance, and the variance of a lower rung is never diluted by more objects on a higher one. A million Hubble-flow supernovae cannot fix a bad anchor.

Example: the error budget of a SH0ES-sized ladder. Take $N_{\rm anc}=3$ anchors with $\sigma_{\rm geo}=0.03$ mag and $n_A=100$ Cepheids each, $N_{\rm cal}=40$ calibrators with $n_C=30$ Cepheids each, $\sigma_{\rm ceph}=0.25$, $\sigma_{\rm SN}=0.13$, and $N_{\rm HF}=280$ Hubble-flow supernovae with $\sigma_{\rm HF}=0.14$. Substituting line 1 into line 2 and line 2 into line 3 of (T2.107) writes $\operatorname{Var}(\hat{a})$ as five separate terms, one per source of noise:

$$\begin{aligned}
\operatorname{Var}(\hat{a})&=\underbrace{\frac{\sigma_{\rm geo}^2}{N_{\rm anc}}}_{\text{anchor geometry}}+\underbrace{\frac{\sigma_{\rm ceph}^2}{N_{\rm anc}n_A}}_{\text{anchor Cepheids}}+\underbrace{\frac{\sigma_{\rm SN}^2}{N_{\rm cal}}}_{\text{calibrator SNe}}+\underbrace{\frac{\sigma_{\rm ceph}^2}{N_{\rm cal}n_C}}_{\text{calibrator Cepheids}}+\underbrace{\frac{\sigma_{\rm HF}^2}{N_{\rm HF}}}_{\text{Hubble flow}}\\
&=\underbrace{\frac{0.03^2}{3}}_{3.0\times10^{-4}}+\underbrace{\frac{0.25^2}{300}}_{2.1\times10^{-4}}+\underbrace{\frac{0.13^2}{40}}_{4.2\times10^{-4}}+\underbrace{\frac{0.25^2}{1200}}_{0.5\times10^{-4}}+\underbrace{\frac{0.14^2}{280}}_{0.7\times10^{-4}}=1.05\times10^{-3},
\end{aligned}$$

where $300=N_{\rm anc}n_A=3\times100$ and $1200=N_{\rm cal}n_C=40\times30$. So $\sigma_a=0.032$ mag and $\sigma_{H_0}/H_0=0.4605\times0.032=1.49\%$. The calibrator supernovae and the anchors dominate; the Hubble flow, with its 280 supernovae, contributes least. Doubling the number of calibrators is worth far more than doubling the Hubble-flow sample.

The experiment. `07_distance_ladder.py` builds the full matrix A for this ladder (47 parameters, one row per measurement) and solves (T2.106):

Listing T2.9: The ladder as one linear model, solved by generalised least squares (`code/chT2/07_distance_ladder.py`)

```
44      # rung 1: geometric distance moduli of the anchors
45      for i in range(Na):
46          add([(i, 1)], mu_anc[i] + rng.normal(0, cfg["sig_geo"]), cfg["sig_geo"])
47      # rung 2: Cepheids, m = mu + M_W + b_W (log P - 1)
48      for i, (mu, n) in enumerate([(m, cfg["Nceph_anc"]) for m in mu_anc] +
49                                  [(m, cfg["Nceph_cal"]) for m in mu_cal]):
50          lp = rng.uniform(0.6, 1.8, n) - 1.0
51          m = mu + truth["MW"] + truth["bW"] * lp + rng.normal(0, cfg["sig_ceph"], n)
52          for k in range(n):
53              add([(i, 1), (iMW, 1), (ibW, lp[k])], m[k], cfg["sig_ceph"])
54      # rung 3a: one supernova in each calibrator host, m = mu + M_B
55      for i in range(Nc):
56          add([(Na + i, 1), (iMB, 1)], mu_cal[i] + truth["MB"] + rng.normal(0,
        cfg["sig_sn"]), cfg["sig_sn"])
57      # rung 3b: Hubble-flow supernovae, m = M_B + 5 log10(cz [1 + (1-q0) z/2]) + 25 - a
58      z = rng.uniform(0.023, 0.15, cfg["Nhf"])
59      x = 5 * np.log10(C_KMS * z * (1 + 0.5 * (1 - truth["q0"]) * z)) + 25
60      a_true = 5 * np.log10(truth["H0"])
61      m = truth["MB"] + x - a_true + rng.normal(0, cfg["sig_hf"], cfg["Nhf"])
62      for k in range(cfg["Nhf"]):
63          add([(iMB, 1), (ia, -1)], m[k] - x[k], cfg["sig_hf"])
64      return np.array(y), np.array(rows), np.array(sig), ia
65
66
67  def gls(y, A, sig):
68      W = 1 / sig**2
69      F = A.T @ (A * W[:, None])                    # Fisher matrix A^T C^-1 A
70      cov = np.linalg.inv(F)
71      theta = cov @ (A.T @ (W * y))
72      return theta, cov
```

Each call to `add` writes one row of A: its nonzero entries say which parameters that measurement involves. Lines 53, 56 and 63 are the three kinds of shared-parameter rows that tie the rungs together. One mock ladder gives $H_0 = 73.90 \pm 1.10$ (true value 73), a 1.49% measurement, the same as the rung-by-rung estimate. Over 400 independent mock ladders the estimates average 72.99 with standard deviation 1.10, while the quoted errors average 1.09, and the 1σ interval covers the truth in 70% of them: the inverse Fisher matrix is the true covariance here, as it must be for a linear Gaussian model. Figure T2.25 shows the budget as the number of calibrators grows: the error falls from 3.05% with 5 calibrators to 1.22% with 160, and then flattens on the anchor floor of 0.80%.

T2.12.3 The real ladder and the Hubble tension

Where we have met H_0 before. Every H_0 so far was made up by us: the 70 of the mock supernovae of Section 3.17, where M was known, and the 73 of the mock ladder just built, recovered as 73.90 ± 1.10. In both the truth was known, and the only question was whether the estimator found it and whether its error bar covered it. Here the numbers are published measurements, the truth is not known, and the question becomes whether two of them can be measurements of the same constant. The ladder side is the same construction as above, the Hubble-flow fit of Section 3.17.4 with M supplied by the rungs, and its published error, 1.04, is close to our mock 1.10 because the mock was built to be SH0ES-sized. The other side is calibrated by something else entirely: no candle, but the sound horizon at recombination, whose angle on the sky fixes

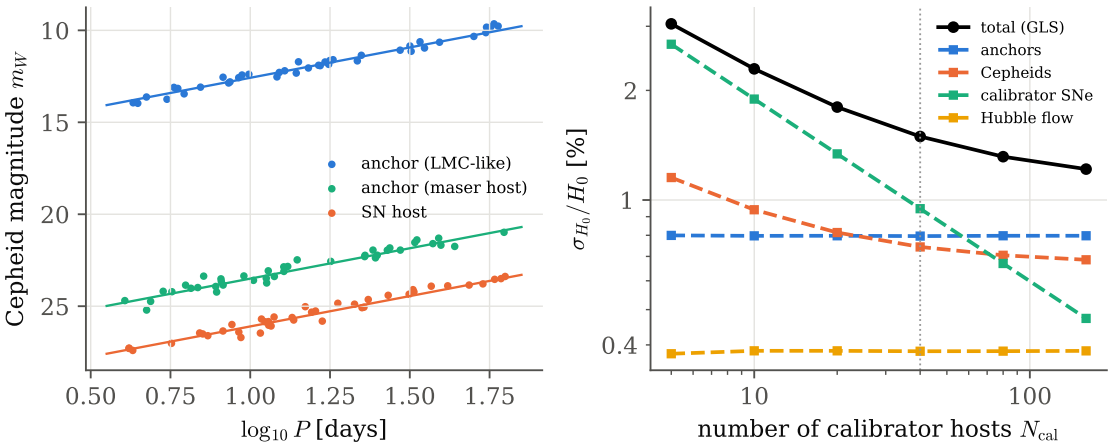


Figure T2.25: Left: mock Cepheids in two anchors and one supernova host follow parallel Leavitt laws; their vertical offsets are the differences of distance modulus. Right: the fractional error on H_0 from the full fit (black) and the contribution of each rung (computed by making that rung noiseless and subtracting in quadrature), against the number of calibrator hosts. The dotted line marks the default ladder.

H_0 as a parameter derived from the CMB fit (Section 10.6.3). The degeneracy of Section 3.17.4 reappears as the way to read the disagreement: since supernovae fix only $M - 5\log_{10} H_0$, a gap in H_0 is a gap in the luminosity of the candle. Section 9.23.1 later adds the ladder value to a CMB chain by reweighting, and Section 11.20 sets all the routes side by side.

Two kinds of measurement give H_0, and they answer different questions. The SH0ES team asks: *how fast are galaxies around us receding now, per megaparsec of distance, with distances from geometry alone?* Their ladder is the one built above, and the only physics it assumes is in the rungs themselves (parallax, the Leavitt law, standardised supernovae); no cosmological model enters except the small q_0 term of (T2.104). The Planck team asks a different question: *which ΛCDM universe produces the pattern of hot and cold spots we see in the CMB, emitted at $z \approx 1100$?* H_0 is one of the outputs of that fit. It is not measured locally but predicted forward, by running the fitted model from $z \approx 1100$ to today. So the two numbers are not two measurements of the same quantity by two instruments; one is a local measurement, the other a prediction of ΛCDM. If they disagree, either a rung of the ladder is wrong or the model that carries the CMB to today is.

method	H_0 [km/s/Mpc]	what it assumes
SH0ES, Cepheid ladder	73.04 ± 1.04	local geometry and the physics of each rung
Planck, CMB	67.36 ± 0.54	flat ΛCDM from $z \approx 1100$ to today
CCHP TRGB ladder	69.8 ± 1.7	local geometry, red giants in place of Cepheids

The values are from [24], [25, Table 2] (TT,TE,EE+lowE+lensing) and [105, Sec. IV.A], which quotes the TRGB value of the Carnegie–Chicago Hubble Program (CCHP) of Freedman et al. The TRGB ladder lands between the other two.

How many standard deviations apart? The two measurements are independent, so the variance of their difference is the sum of their variances (Chapter 4), and

$$\frac{73.04 - 67.36}{\sqrt{1.04^2 + 0.54^2}} = \frac{5.68}{1.17} = 4.8,$$

the "about 5σ" of [105, Sec. IV.A].

The same gap in magnitudes. By (T2.105) the Hubble-flow supernovae fix only $M_B - a = M_B - 5\log_{10} H_0$. Hold that combination fixed and raise H_0 by a factor r: a grows by $5\log_{10} r$, so M_B must grow by the same amount. A larger M_B means intrinsically fainter supernovae, which for the same apparent magnitude must be nearer, and nearer galaxies with the same redshifts mean a faster expansion. For the two values,

$$\Delta M_B = 5\log_{10}\frac{73.04}{67.36} = 5\times 0.035 = 0.18\ \text{mag}:$$

the whole tension is equivalent to a disagreement of 0.18 mag about how bright type Ia supernovae are, with the ladder's supernovae the fainter ones. That is why it is cleanest to carry the ladder into a cosmological fit not as a value of H_0 but as a Gaussian prior on M_B [105, Sec. IV.E, eq. (1)]: the ladder calibrates M_B, and H_0 then follows from the supernovae together with whatever cosmological model is being tested.

Example: can a local void explain it? Suppose we live inside a sphere whose mean density is below the cosmic mean, with δ the mean density contrast inside the sphere ($\delta < 0$ for a void). Why would the local expansion be faster? The denser surroundings pull the matter at the edge of the sphere outwards, and the sphere has less mass inside to hold it back, so matter streams out of the void. On top of the Hubble flow there is an outflow, and someone inside who measures recession velocities sees them grow faster with distance than the global H_0: the local rate is $H_0 + \delta H$ with $\delta H > 0$. How much faster? Linear continuity, $\dot\delta = -\nabla\cdot\boldsymbol{v}$ for the peculiar velocity $\boldsymbol{v}$ (derived in Section T2.14, (T2.118)), and the growth rate $\dot\delta = fH\delta$ with $f \approx \Omega_m^{0.55} \approx 0.53$ give

$$\frac{\delta H_0}{H_0} \overset{(1)}{=} \frac{\nabla\cdot\boldsymbol{v}}{3H_0} \overset{(2)}{=} -\frac{\dot\delta}{3H_0} \overset{(3)}{=} -\frac{f}{3}\,\delta, \tag{T2.108}$$

(1) The extra outflow grows in proportion to distance, $\boldsymbol{v} = \delta H\,\boldsymbol{r}$, which is what a change δH of the local Hubble rate means. Its divergence is $\nabla\cdot\boldsymbol{v} = \partial_x v_x + \partial_y v_y + \partial_z v_z = 3\,\delta H$.

(2) Continuity: matter flowing out of the sphere lowers its density, $\dot\delta = -\nabla\cdot\boldsymbol{v}$.

(3) Linear growth, $\delta \propto D(t)$, so $\dot\delta = (\mathrm{d}\ln D/\mathrm{d}\ln a)(\dot a/a)\,\delta = fH\delta$, with $H = H_0$ today.

which is Abdalla et al.'s eq. (2) [105, Sec. IV.F]. The sign is as expected: $\delta < 0$ gives $\delta H_0 > 0$. How empty would the void have to be? The gap between the ladder and the CMB is $(73.04 - 67.36)/67.36 = 8.4\%$, about 8%, so

$$0.08 = \frac{0.53}{3}\,|\delta| \quad\Longrightarrow\quad |\delta| = \frac{3\times 0.08}{0.53} = 0.45,$$

a region almost half empty, over the whole volume probed by the Hubble-flow supernovae, out to $z = 0.15$ or some 600 Mpc. Density fluctuations on such scales are at the percent level, and theory and simulations put the expected scatter of the local H_0 at 0.5–1% [105, Sec. IV.F], an order of magnitude too small.

Where the statistics comes back in. The ladder is generalised least squares (Chapter 3), and its covariance is the inverse Fisher matrix (Chapter 10). Testing whether two measurements of H_0 agree is a hypothesis test on their difference (Chapter 4). Using the ladder as a prior on M_B is Bayesian combination of independent data (Chapter 5). Peculiar velocities and the local Hubble flow are taken up again as data in their own right in Section 11.16, and the local cosmic variance of H_0, the careful version of the void estimate above, in Section 11.17.

T2.13 Selection effects: what a flux limit does to a sample

Every survey has a faintest object it can detect. Close to that limit the survey does not see a fair sample of what is there: it sees the objects that happened to be brighter than average, because their fainter siblings at the same distance were lost. Treat the survivors as typical, and their distances come out too small. At high redshift, where most of the cosmological information sits, the supernovae look brighter than the model predicts, and the fit moves the cosmological parameters to make them so.

The plain claim is that "the mean absolute magnitude of type Ia supernovae is M_0". The everyday question is: *if I average the absolute magnitudes of the supernovae in my catalogue, do I get M_0?* The question hides a sampling rule. "The mean" of the luminosity function is an average over supernovae drawn uniformly *per unit volume*. A catalogue is drawn by a different rule: a supernova enters it if its apparent magnitude is below the limit. Change the sampling rule and the distribution changes, exactly as the distribution of an oscillator's position changes when we sample it uniformly in position rather than in time. The physics of the explosions is the same in both cases; what differs is how the objects were selected.

Selection is conditioning, and conditioning has a picture (Figure 1.11, Section 5.2). Draw every supernova that went off in the survey volume as a point in a rectangle: that rectangle is the full sample space, and the luminosity function describes how the points are spread over it. Detection draws a line through the rectangle: on one side are the supernovae brighter than the limit, $m < m_{\lim}$, on the other the ones we never see. Once we know a supernova is in the catalogue, everything on the far side of the line is discarded. The surviving region becomes the new sample space, and every probability is measured again as a fraction of its area. If the survivors are a fraction $P(\text{detected})$ of the whole, every probability of the full population is divided by that fraction. The mean of the survivors, and the likelihood of what they show us, are both computed in the shrunken space. Nothing about an individual explosion has changed; only the region against which it is measured has.

The question. For candles with a Gaussian spread of absolute magnitudes, (1) what is the mean absolute magnitude of the ones a flux-limited survey detects, (2) how biased are the Hubble residuals near the limit, and (3) what likelihood removes the bias?

T2.13.1 The classical Malmquist bias

The inputs are three.

1. *Physical:* the candles are spread uniformly through a static Euclidean space, with number density n.
2. *Physical:* their absolute magnitudes follow a Gaussian **luminosity function** [the distribution of absolute magnitudes among the candles] $\varphi(M) = \mathcal{N}(M_0, \sigma^2)$, the same everywhere and independent of position.
3. *Selection:* we look at the objects whose apparent magnitude lies in a narrow interval $[m, m + \mathrm{d}m]$.

What we observed is an object of apparent magnitude m. *What could have produced it?* Any absolute magnitude M, placed at the distance modulus $\mu = m - M$ that makes it appear at m. *How likely is each candidate?* It is the probability of drawing that M from the luminosity function, times the number of places at that distance where the object could be. The number of objects in

a shell of radius r and thickness $\mathrm{d}r$ is $n\,4\pi r^2\mathrm{d}r$, and r is tied to μ by $r = 10^{\mu/5+1}$ pc. So

$$\mathrm{d}N(M,\mu) \overset{(1)}{=} n\,\varphi(M)\,\mathrm{d}M\,4\pi r^2\,\frac{\mathrm{d}r}{\mathrm{d}\mu}\,\mathrm{d}\mu \overset{(2)}{=} n\,\varphi(M)\,\mathrm{d}M\,4\pi\frac{\ln 10}{5}\,r^3\,\mathrm{d}\mu \;\propto\; \varphi(M)\,10^{0.6\mu}\,\mathrm{d}M\,\mathrm{d}\mu,$$

$$p(M\mid m) \overset{(3)}{=} \frac{\varphi(M)\,10^{0.6(m-M)}}{\int \varphi(M')\,10^{0.6(m-M')}\mathrm{d}M'} \overset{(4)}{=} \mathcal{N}\big(M_0 - 0.6\ln 10\,\sigma^2,\ \sigma^2\big). \tag{T2.109}$$

(1) Independence of luminosity and position (inputs 1 and 2): the joint density is the product. Change variable from r to μ.
(2) $r \propto 10^{\mu/5}$, so $\mathrm{d}r/\mathrm{d}\mu = (\ln 10/5)\,r$, and $r^3 \propto 10^{3\mu/5}$.
(3) Fix m (input 3): $\mu = m - M$, and $\mathrm{d}\mu = \mathrm{d}m$ does not depend on M. Normalise over M.
(4) $10^{-0.6M} = e^{-0.6\ln 10\,M}$. Multiply it into the Gaussian and complete the square: $-\frac{(M-M_0)^2}{2\sigma^2} - kM = -\frac{(M-M_0+k\sigma^2)^2}{2\sigma^2}$ + const with $k = 0.6\ln 10$.

Key idea: Malmquist bias

In a flux-limited sample of candles spread uniformly in Euclidean space, the detected objects are on average brighter than the luminosity function by

$$\langle M\rangle_{\mathrm{det}} - M_0 = -0.6\ln 10\;\sigma^2 = -1.382\,\sigma^2,$$

at every apparent magnitude. The reason is volume: there is more space at large distances, so an object at a given m is more likely to be a bright one far away than a faint one nearby.

The shift does not depend on m or on where the limit is, so the whole flux-limited sample shares it. The size matters. For raw supernovae, $\sigma = 0.4$ mag, the shift is

$$\langle M\rangle_{\mathrm{det}} - M_0 = -1.382\times 0.4^2 = -0.22\ \mathrm{mag}.$$

Which way does that move a distance? We infer the distance modulus as $\mu_{\mathrm{inf}} = m - M_0$. The detected supernovae really have, on average, $M = M_0 - 0.22$, so their true modulus is $\mu = m - M_0 + 0.22$, and the inferred one is too small by 0.22: a candle that is brighter than we assume looks closer than it is. Since $D_L \propto 10^{\mu/5}$ by (T2.97),

$$\frac{D_{\mathrm{inf}}}{D_{\mathrm{true}}} = 10^{-0.22/5} = 10^{-0.044} = 0.90,$$

so distances come out 10% too small. After standardisation, $\sigma = 0.12$ mag, the shift is $1.382\times 0.12^2 = 0.020$ mag and $10^{-0.020/5} = 0.99$, a 1% distance error, which is still as large as the statistical error of a modern survey. The key idea depends on its inputs. In an expanding universe the volume does not grow as r^3; if the spread σ itself depends on distance, or the luminosity function has a non-Gaussian tail, the number changes. The mechanism, a luminosity function reweighted by the volume where each luminosity can be seen, does not.

T2.13.2 Near the limit: a truncated Gaussian

Now the version that matters for a supernova Hubble diagram. At redshift z, the model predicts a mean apparent magnitude $\bar m(z;\boldsymbol\theta) = M_B + \mu(z;\boldsymbol\theta)$, and the observed magnitudes scatter about it, $m = \bar m + \epsilon$ with $\epsilon \sim \mathcal{N}(0,\sigma^2)$. The survey keeps a supernova only if $m < m_{\mathrm{lim}}$. *What is the mean residual of the supernovae it keeps?* With $t \equiv (m_{\mathrm{lim}} - \bar m)/\sigma$ the distance of the limit from the mean in units of σ, and ϕ, Φ the standard normal density and distribution function,

$$\mathbb{E}[\epsilon\mid \epsilon < t\sigma] \overset{(1)}{=} \frac{\int_{-\infty}^{t\sigma} \epsilon\,\frac{1}{\sigma}\phi(\epsilon/\sigma)\,\mathrm{d}\epsilon}{\Phi(t)} \overset{(2)}{=} \frac{-\sigma\big[\phi(\epsilon/\sigma)\big]_{-\infty}^{t\sigma}}{\Phi(t)} \overset{(3)}{=} -\sigma\,\frac{\phi(t)}{\Phi(t)}. \tag{T2.110}$$

(1) Conditional expectation: integrate over the kept region and divide by its probability $\mathrm{P}(\epsilon < t\sigma) = \Phi(t)$. The division is the renormalisation of the shrunken sample space: the kept region has total probability $\Phi(t)$, and dividing by it makes the probabilities inside it add up to one again.
(2) $\frac{\mathrm{d}}{\mathrm{d}\epsilon}\phi(\epsilon/\sigma) = -\frac{\epsilon}{\sigma^2}\phi(\epsilon/\sigma)$, so $\epsilon\,\phi(\epsilon/\sigma)/\sigma$ is $-\sigma$ times the derivative of $\phi(\epsilon/\sigma)$.
(3) ϕ vanishes at $-\infty$.

Two cases give the result a size. Far from the limit ($t \gg 1$), $\Phi(t) \to 1$ (almost nothing is cut) and $\phi(t) \to 0$ (the Gaussian has no weight there), so the bias vanishes. Where half the supernovae are lost ($t = 0$), $\phi(0) = 1/\sqrt{2\pi}$ and $\Phi(0) = 1/2$, so the bias is $-\sigma \cdot 2/\sqrt{2\pi} = -\sigma\sqrt{2/\pi} = -0.80\,\sigma$. Beyond that ($t < 0$) it grows without bound. Figure T2.26 shows the mechanism at three distances.

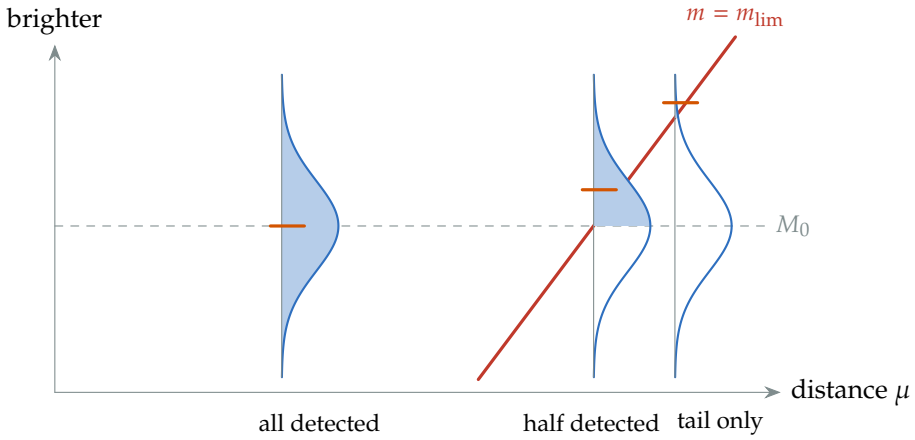


Figure T2.26: The luminosity function of the candles (blue bell, brighter upwards) at three distances, with the flux limit (red). Only the shaded part, brighter than the limit, is detected. The orange tick marks the mean brightness of the detected objects: at the true M_0 nearby, 0.8σ brighter where the limit cuts the bell in half, and about 2.7σ brighter where only the bright tail survives.

The fix: write the likelihood of what was selected. The likelihood asks how probable the observed magnitudes are under each model. For a catalogue, the honest version of that question is: how probable are they *among the supernovae that could have entered the catalogue*? That is the shrunken sample space again, now for a density instead of a mean:

$$p(m \mid z, \boldsymbol{\theta}, \text{detected}) \overset{(1)}{=} \frac{p(m, \text{detected} \mid z, \boldsymbol{\theta})}{\mathrm{P}(\text{detected} \mid z, \boldsymbol{\theta})} \overset{(2)}{=} \frac{\mathcal{N}\big(m; \bar{m}(z;\boldsymbol{\theta}), \sigma^2\big)}{\Phi\big(t(z;\boldsymbol{\theta})\big)} \qquad (m < m_{\mathrm{lim}}),$$

$$-\ln\mathcal{L} \overset{(3)}{=} \sum_i \left[\frac{(m_i - \bar{m}_i)^2}{2\sigma^2} + \ln\sigma + \tfrac{1}{2}\ln 2\pi + \ln\Phi(t_i)\right]$$

$$\overset{(4)}{=} \sum_i \left[\frac{(m_i - \bar{m}_i)^2}{2\sigma^2} + \ln\Phi(t_i)\right] + \text{const.} \tag{T2.111}$$

(1) The definition of conditional probability: the joint density divided by the probability of the condition.

(2) For $m < m_{\lim}$ "detected" is automatically true, so the joint density is just the Gaussian. The probability of detection is $\mathrm{P}(m < m_{\lim}) = \Phi(t)$, the same $\Phi(t)$ that renormalised the mean in (T2.110).

(3) The supernovae are independent, so the likelihood is a product and $-\ln \mathcal{L}$ a sum. Write out $-\ln$ of the Gaussian density $\frac{1}{\sqrt{2\pi}\sigma} e^{-(m-\bar{m})^2/2\sigma^2}$, and note that $-\ln(1/\Phi) = +\ln \Phi$.

(4) σ is known and the same for every model, so $\ln \sigma + \frac{1}{2}\ln 2\pi$ is a constant that does not move the fit.

The first term is the naive $\chi^2/2$. The second is new, and it depends on $\boldsymbol{\theta}$ through $t_i = (m_{\lim} - \bar{m}_i)/\sigma$. Since $\Phi \le 1$, $\ln \Phi(t) \le 0$: the term *lowers* $-\ln \mathcal{L}$, most strongly for a model whose mean $\bar{m}$ lies near or beyond the limit, where $\Phi(t)$ is small. *What is that credit for?* Take one supernova detected at $m = m_{\lim} - 0.5\sigma$ and two models. Model A puts the mean exactly there, $\bar{m} = m_{\lim} - 0.5\sigma$, so $t = 0.5$. Model B puts the mean fainter than the limit, $\bar{m} = m_{\lim} + \sigma$, so $t = -1$ and the supernova sits 1.5σ out in the bright tail.

model	$\chi^2/2$	$\ln \Phi(t)$	$-\ln \mathcal{L}$ (no constant)
A, $t = 0.5$	0	$\ln 0.69 = -0.37$	-0.37
B, $t = -1$	$1.5^2/2 = 1.13$	$\ln 0.16 = -1.84$	-0.72

The naive fit prefers A by a factor $e^{1.13} \approx 3$, because under B this supernova is a 1.5σ excursion. But under B, *every* supernova we can see is an excursion into the bright tail: the selection guarantees it. Being in the tail is not evidence against B; it is what B predicts for the supernovae that got into the catalogue. The $\ln \Phi$ term returns exactly that credit, and with it B is, if anything, slightly preferred (by $e^{0.35} \approx 1.4$), because under B the detected supernovae pile up just inside the limit, which is where this one is. With many supernovae the balance is set by how they are spread inside the limit, which is what the full sum in (T2.111) weighs. (The number of detected supernovae is not used here. A likelihood that also used the count would add a term that does penalise a model for predicting too few detections.) The result is the likelihood of Chapter 3, written for the data *as they were collected*: a density for what could have been observed, given the rule that decided what got into the sample. Real surveys have no sharp limit. The detection probability $\eta(m)$ rises smoothly from 0 to 1 and depends on the observing conditions, so it is measured by injecting fake supernovae into the real images, and $\Phi(t)$ in (T2.111) becomes $\int \eta(m)\mathcal{N}(m; \bar{m}, \sigma^2)\,\mathrm{d}m$. Pantheon+ corrects each supernova by a bias computed from large simulations of the whole survey [26].

The experiment. `08_malmquist.py` does both calculations.

The Euclidean test. It places 4×10^5 candles uniformly in a ball and keeps those brighter than a limit. The mean M of the detected ones is shifted by -0.0138, -0.1280 and -0.348 mag for $\sigma = 0.1$, 0.3 and 0.5, against $-1.382\sigma^2 = -0.0138$, -0.1243 and -0.345 from the key idea.

The survey. Then it simulates supernova surveys in an expanding universe, with two samples:

sample	redshifts	selection
low redshift	$0.02 < z < 0.1$	complete: every supernova is kept
deep	$0.1 < z < 1.2$	kept only if $m < m_{\lim} = 24$
both	flat ΛCDM with $\Omega_{m0} = 0.3$; scatter $\sigma = 0.15$ mag about $\bar{m}(z)$	

Each fit has two free numbers, Ω_{m0} and an overall magnitude offset (the combination $M_B - 5\log_{10} H_0$ of Section T2.12), which the complete low-redshift sample pins down. The two fits differ by one line:

Listing T2.10: The naive and the selection-aware negative log-likelihood (code/chT2/08_malmquist.py)

```
68 def nll(p, z, m, lim, corrected):
69     om, off = p
70     if not 0.02 < om < 0.98:
71         return 1e10
72     mod = mu_model(z, om) + off
73     t = (m - mod) / SIG
74     out = 0.5 * np.sum(t**2)
75     if corrected:
76         fin = np.isfinite(lim)
77         out += np.sum(norm.logcdf((lim[fin] - mod[fin]) / SIG))
78     return out
```

Line 74 is the $\chi^2/2$ of a standard fit. Line 77 adds $\sum \ln \Phi(t_i)$, the second term of (T2.111), for the supernovae of the deep sample.

The result. Figure T2.27 shows it. Half the supernovae are lost at $z \approx 0.73$, and in the last redshift bin that still contains supernovae ($z \approx 0.75$) the mean residual is -0.140 mag. Over 200 mock surveys the naive χ^2 returns $\Omega_{m0} = 0.353 \pm 0.019$, too high. The chain that leads there has three links. The selected high-redshift supernovae are brighter than the model mean, so they look nearer than they are. The offset is fixed by the complete low-redshift sample, so the only way the naive fit can bring the model to them is to shorten the model's distances at high redshift. In flat ΛCDM at fixed H_0, a shorter $D_L = (1+z)\,c\int_0^z \mathrm{d}z'/H(z')$ at high redshift needs a larger $H(z')$ in the past, so the expansion must have slowed down more since then: more deceleration, which needs more matter. The selection-aware likelihood returns 0.301 ± 0.018, unbiased, with the same scatter. The naive bias is almost three times its own error bar, and it would not shrink with more supernovae.

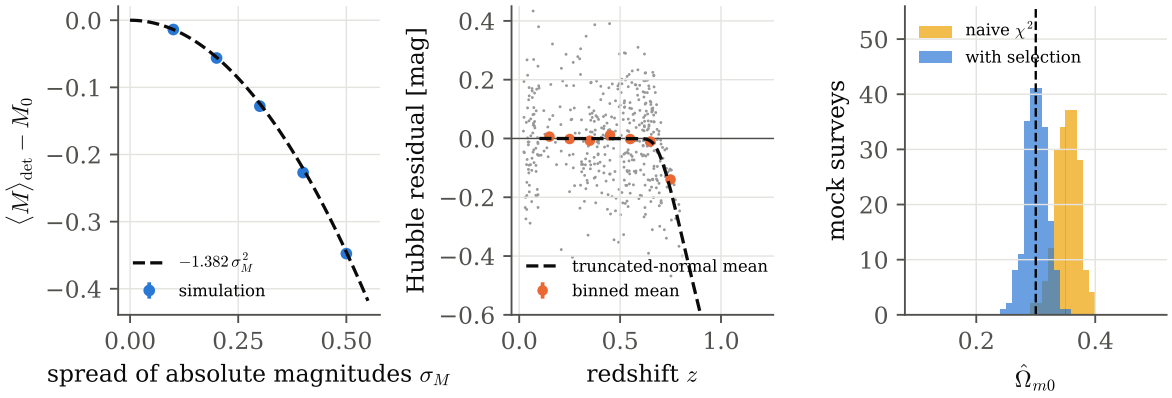


Figure T2.27: Left: the classical Malmquist shift measured in simulation against $-1.382\sigma^2$. Middle: Hubble residuals of a magnitude-limited survey; the binned means (orange) follow the truncated-normal mean $-\sigma\phi(t)/\Phi(t)$ (dashed) and plunge as the limit cuts into the population. Right: Ω_{m0} from many mock surveys; the naive fit is biased, the likelihood with the selection term is not.

A relative: Eddington bias. A related effect acts on counts rather than on luminosities. If the number of objects rises steeply towards faint fluxes, measurement noise scatters more objects up across a threshold than down, so the objects just above the threshold are, on average, fainter than they were measured. Both biases have the same cure: model the selection, and write the likelihood of the data that were actually kept.

Where the statistics comes back in. The conditional density (T2.109) is Bayes' theorem of Chapter 5, with the luminosity function as prior and the volume factor as likelihood. The truncated likelihood is the likelihood of Chapter 3 written for the data actually collected. Selection functions that are known only from simulations are a Monte Carlo problem (Chapter 6), and when they are folded into a hierarchical supernova model the posterior is explored by MCMC (Chapter 9).

T2.14 Galaxies as biased tracers in redshift space

A redshift survey records, for each galaxy, two angles on the sky and a redshift. The Dark Energy Spectroscopic Instrument (DESI) did this for more than six million galaxies and quasars in its first year: bright galaxies at $0.1 < z < 0.4$, luminous red galaxies at $0.4 < z < 1.1$, star-forming emission-line galaxies at $0.8 < z < 1.6$, quasars at $0.8 < z < 2.1$, and the absorption of quasar light by intergalactic hydrogen (the Lyman-α forest) out to $z \approx 4$ [106, Sec. 2.2, Table 1]. A fiducial cosmology turns each redshift into a distance and the survey into a three-dimensional map of points. The counts in cells of that map define the galaxy overdensity $\delta_g(\boldsymbol{x}) = n(\boldsymbol{x})/\bar{n} - 1$, and its two-point statistics, the correlation function $\xi_g(r)$ and the power spectrum $P_g(k)$ of Chapter 7, are what the statistics chapters work with.

The map is not the matter field. Three things stand between them. Galaxies form only where matter has collapsed into haloes, so they are *biased*: their density contrast is larger than that of the matter. Galaxies are discrete points, so their map carries *shot noise*. And the redshift contains the galaxy's peculiar velocity as well as its distance, so the map is *distorted* along the line of sight.

The question. How is the galaxy power spectrum related to the matter power spectrum on large scales, and what is the noise of a measured P_g?

T2.14.1 Why galaxies cluster more strongly than matter

The inputs are two. *Physical:* a galaxy forms where the local density, smoothed on the scale of a halo, exceeds a threshold. *Mathematical:* the initial density field is Gaussian. Write the local density at $\boldsymbol{x}$ as $u(\boldsymbol{x})$, with variance σ^2, and say a galaxy forms where $u > \nu\sigma$; the number ν measures how rare the galaxies are. The local density is a sum of two parts, $u = u_{\text{long}} + u_{\text{short}}$: a long-wavelength part, shared with the surroundings over tens of megaparsecs, and a short-wavelength part that varies from halo to halo and decides locally whether the threshold is crossed. Only the first can link a galaxy to matter far away.

What we know is that a galaxy formed at $\boldsymbol{x}$. *What does that tell us about the matter at* $\boldsymbol{y}$*, a distance* r *away?* The galaxy is a sign that $u(\boldsymbol{x})$ was high. Because $u(\boldsymbol{x})$ and $\delta(\boldsymbol{y})$ are correlated, the matter at $\boldsymbol{y}$ is denser than average too. The cross-correlation of galaxies and matter is exactly that expected density:

$$\xi_{gm}(r) \overset{(1)}{=} \frac{\mathbb{E}[I(\boldsymbol{x})\,\delta(\boldsymbol{y})]}{\mathbb{E}[I]} \overset{(2)}{=} \mathbb{E}\left[\delta(\boldsymbol{y}) \,|\, u(\boldsymbol{x}) > \nu\sigma\right] \overset{(3)}{=} \mathbb{E}\left[\frac{\xi(r)}{\sigma^2}\, u \,\middle|\, u > \nu\sigma\right] \overset{(4)}{=} \frac{\xi(r)}{\sigma}\,\frac{\phi(\nu)}{1-\Phi(\nu)} \equiv b_L\,\xi(r). \tag{T2.112}$$

(1) $I(\boldsymbol{x})$ is 1 where a galaxy forms and 0 elsewhere, so $\delta_g = I/\mathbb{E}[I] - 1$; the -1 drops out because $\mathbb{E}[\delta] = 0$.

(2) Dividing by $\mathbb{E}[I] = \mathrm{P}(u > \nu\sigma)$ turns the average into a conditional expectation.

(3) For two jointly Gaussian variables the conditional mean is linear: $\mathbb{E}[\delta(\boldsymbol{y}) \,|\, u] = \frac{\mathrm{Cov}(\delta(\boldsymbol{y}),u(\boldsymbol{x}))}{\mathrm{Var}\,u}\,u$ (derived in Eq. (7.7) and Section 2.12.4). Which part of $u(\boldsymbol{x})$ is correlated with matter a

distance r away? Split $u = u_{\text{long}} + u_{\text{short}}$ as above. The short-wavelength part varies over the size of a halo and has no memory of what happens tens of megaparsecs away, so $\operatorname{Cov}(\delta(\boldsymbol{y}), u_{\text{short}}(\boldsymbol{x})) = 0$. The long-wavelength part is the matter field itself, smoothed over a region much smaller than r, so $\operatorname{Cov}(\delta(\boldsymbol{y}), u_{\text{long}}(\boldsymbol{x})) = \xi(r)$, the matter correlation. Hence $\operatorname{Cov}(\delta(\boldsymbol{y}), u(\boldsymbol{x})) = \xi(r)$, while $\operatorname{Var} u = \sigma^2$ counts both parts.

(4) The mean of a Gaussian above $\nu\sigma$ is $\sigma\phi(\nu)/[1-\Phi(\nu)]$, the upper-tail version of (T2.110).

For rare galaxies, $\nu \gg 1$, the ratio $\phi(\nu)/[1-\Phi(\nu)] \to \nu$, and $b_L \to \nu/\sigma$: the rarest objects are the most strongly clustered. This is Kaiser's explanation of why rich galaxy clusters cluster more strongly than galaxies.

From galaxy–matter to galaxy–galaxy. Equation (T2.112) conditions on one galaxy and looks at the matter. The galaxy map has a galaxy at both ends of each pair, so both points are now conditioned. Write $x_1 = u(\boldsymbol{x})/\sigma$ and $x_2 = u(\boldsymbol{y})/\sigma$ for the two local densities in units of σ. They are jointly Gaussian with unit variances and a small correlation $\rho = \xi(r)/\sigma^2$ (the long-wavelength parts carry it, as in item 3). Then

$$1+\xi_{gg}(r) \overset{(1)}{=} \frac{\mathbb{E}[I(\boldsymbol{x})I(\boldsymbol{y})]}{\mathbb{E}[I]^2} \overset{(2)}{=} \frac{1}{\mathbb{E}[I]^2}\int_\nu^\infty\!\!\int_\nu^\infty \phi(x_1)\phi(x_2)\big(1+\rho\, x_1x_2\big)\,\mathrm{d}x_1\mathrm{d}x_2$$
$$\overset{(3)}{=} 1+\rho\Big[\frac{\int_\nu^\infty x\,\phi(x)\,\mathrm{d}x}{1-\Phi(\nu)}\Big]^2 \overset{(4)}{=} 1+\frac{\xi(r)}{\sigma^2}\Big[\frac{\phi(\nu)}{1-\Phi(\nu)}\Big]^2 = 1+b_L^2\,\xi(r). \tag{T2.113}$$

(1) $\delta_g = I/\mathbb{E}[I]-1$, so $\mathbb{E}[(1+\delta_g(\boldsymbol{x}))(1+\delta_g(\boldsymbol{y}))] = 1+\xi_{gg}$.

(2) The bivariate Gaussian density with correlation ρ is $\frac{1}{2\pi\sqrt{1-\rho^2}}\exp\big[-\frac{x_1^2-2\rho x_1x_2+x_2^2}{2(1-\rho^2)}\big]$. To first order in ρ the prefactor and the $1-\rho^2$ are unchanged, and the exponent gains $+\rho x_1x_2$, so the density is $\phi(x_1)\phi(x_2)e^{\rho x_1x_2} \approx \phi(x_1)\phi(x_2)(1+\rho x_1x_2)$.

(3) The 1 integrates to $[1-\Phi(\nu)]^2 = \mathbb{E}[I]^2$. The ρx_1x_2 term factorises into two identical one-dimensional integrals.

(4) $\int_\nu^\infty x\phi(x)\mathrm{d}x = [-\phi(x)]_\nu^\infty = \phi(\nu)$, as in (T2.110). Compare with (T2.112): $b_L = \phi(\nu)/\{\sigma[1-\Phi(\nu)]\}$.

Each conditioned point contributes one factor b_L: a galaxy at $\boldsymbol{x}$ raises the expected density near $\boldsymbol{y}$ by $b_L\xi$ (that is (T2.112)), and the threshold at $\boldsymbol{y}$ turns that raised density into b_L times as many extra galaxies. So $\xi_{gg} = b_L^2\,\xi$, and since the power spectrum is the Fourier transform of the correlation function, $P_{gg} = b_L^2 P$.

This bias is "Lagrangian": it describes where the galaxies formed. Afterwards they move with the matter, and the motion adds the matter's own contrast. *Label each bit of matter by where it started.* Call the starting (Lagrangian) position $\boldsymbol{q}$. Write $\delta_L(\boldsymbol{q})$ for the linear matter contrast there, grown to the time we observe with the linear growth factor, so that $b_L\delta_L$ is the galaxies' contrast of (T2.112) at their birth place. At the start the matter is uniform to first order, while the galaxies already carry their contrast $\delta_g^L = b_L\delta_L$. *Move everything.* Each bit of matter, and any galaxy riding in it, moves to its present (Eulerian) position $\boldsymbol{x} = \boldsymbol{q} + \boldsymbol{\psi}(\boldsymbol{q})$, with $\boldsymbol{\psi}$ the displacement. A small volume d^3q becomes $\mathrm{d}^3x = J\,\mathrm{d}^3q$, with $J = |\det(\partial x_i/\partial q_j)|$ the Jacobian of the map. *Count.* No galaxy and no matter is created or lost on the way, so the number in the small volume is the same before and after:

$$\bar n_g\big(1+\delta_g(\boldsymbol{x})\big)\,\mathrm{d}^3x \overset{(1)}{=} \bar n_g\big(1+b_L\delta_L(\boldsymbol{q})\big)\,\mathrm{d}^3q \quad\Longrightarrow\quad 1+\delta_g \overset{(2)}{=} \frac{1+b_L\delta_L}{J}, \qquad 1+\delta \overset{(3)}{=} \frac{1}{J},$$
$$1+\delta_g \overset{(4)}{=} \big(1+b_L\delta_L\big)\big(1+\delta\big) \overset{(5)}{=} 1+b_L\delta+\delta+O(\delta^2). \tag{T2.114}$$

(1) Number conservation for the galaxies; $\bar n_g$, the mean density, is the same before and after because the total volume and number do not change.

(2) Divide by $\bar{n}_g\,\mathrm{d}^3x = \bar{n}_g J\,\mathrm{d}^3q$.
(3) The same count for the matter, which started uniform.
(4) Eliminate J between the two.
(5) Multiply out and keep first-order terms. To first order the evolved contrast and the linear one are the same field, $\delta = \delta_L$, and $\boldsymbol{x}$ and $\boldsymbol{q}$ differ only by the small displacement, so $\delta_L(\boldsymbol{q}) = \delta(\boldsymbol{x})$ to this order.

So, to first order,

$$\delta_g = (1 + b_L)\,\delta \equiv b\,\delta, \qquad P_g(k) = b^2 P(k), \tag{T2.115}$$

the "continuity" derivation that Cooray and Sheth mention for the large-scale bias of haloes [107, Sec. 3.3].

The halo-model version of the same idea gives each halo mass its bias, quoted in a different variable. A halo of mass m forms where the linear density smoothed on the scale of that mass, of rms $\sigma(m)$, exceeds the spherical-collapse threshold $\delta_{\rm sc} \approx 1.69$. In our language the threshold in units of σ is $\nu = \delta_{\rm sc}/\sigma(m)$. Cooray and Sheth use instead its square, which we write $\nu_h \equiv \delta_{\rm sc}^2/\sigma^2(m) = \nu^2$ to keep the two apart, and quote $b_1(m) = 1 + (\nu_h - 1)/\delta_{\rm sc}$ for spherical collapse, with an improved form fitted to simulations [107, eqs. (53), (68)–(70)]. *How does this compare with ours?* The "1+" is the step from Lagrangian to Eulerian just derived. For the Lagrangian part, our high-threshold limit gives

$$b_L \to \frac{\nu}{\sigma} = \frac{\delta_{\rm sc}}{\sigma^2} = \frac{\nu_h}{\delta_{\rm sc}},$$

so the halo formula differs from ours only by the -1. It comes from what is being counted. Our b_L belongs to the fraction of space above a threshold, a Gaussian tail. The halo model counts haloes in a narrow mass range, which is the rate at which that tail changes with mass, and for spherical collapse this abundance goes as $\nu_h^{1/2} e^{-\nu_h/2}$ instead of a tail probability. A long-wavelength contrast $\delta_{\rm long}$ lowers the barrier to $\delta_{\rm sc} - \delta_{\rm long}$, so $\nu_h \to (\delta_{\rm sc} - \delta_{\rm long})^2/\sigma^2$ and $\mathrm{d}\nu_h/\mathrm{d}\delta_{\rm long} = -2\nu_h/\delta_{\rm sc}$. The fractional change of the abundance is then

$$\frac{\mathrm{d}}{\mathrm{d}\delta_{\rm long}}\Big[\tfrac{1}{2}\ln\nu_h - \tfrac{1}{2}\nu_h\Big] = \Big(\frac{1}{2\nu_h} - \frac{1}{2}\Big)\Big(-\frac{2\nu_h}{\delta_{\rm sc}}\Big) = \frac{\nu_h - 1}{\delta_{\rm sc}},$$

the Lagrangian bias of the halo formula. The ν_h is our threshold effect, and the -1 comes from the prefactor $\nu_h^{1/2}$. For rare haloes, $\nu_h \gg 1$, the two agree. A galaxy population then has the occupation-weighted average bias $b_{\rm gal} = \int \mathrm{d}m\, n(m)\, b_1(m) \langle N_{\rm gal} \,|\, m\rangle/\bar{n}_{\rm gal}$, and $P_{\rm gal} \approx b_{\rm gal}^2 P_{\rm lin}$ on large scales [107, eqs. (130)–(131)]. Since $b_1(m)$ grows with ν_h, and so with mass, luminous red galaxies, which live in massive haloes, are more strongly biased than emission-line galaxies, which live in smaller ones.

Key idea: Linear bias

On scales much larger than a halo, the galaxy density contrast is proportional to the matter density contrast, $\delta_g = b\,\delta$, so $P_g(k) = b^2 P(k)$ and $\xi_g(r) = b^2 \xi(r)$. The shape of the clustering is the matter's; only its amplitude belongs to the galaxies. The bias is a nuisance parameter in every fit of galaxy clustering.

T2.14.2 Shot noise

Galaxies are a discrete sample of a continuous field: given the field, the number of galaxies in a small cell is Poisson-distributed with mean $\bar{n}(1 + \delta_g)\,\mathrm{d}V$ (Section 2.5.1). What does that do to the measured power? Put $N = \bar{n}V$ galaxies at positions $\boldsymbol{x}_i$ in a box of volume V. The density is

a sum of spikes, $n(\boldsymbol{x}) = \sum_i \delta_D(\boldsymbol{x} - \boldsymbol{x}_i)$, and its Fourier transform at any $\boldsymbol{k} \neq 0$ is a sum of phases, so with the convention $\mathbb{E}[\delta_{\boldsymbol{k}} \delta^*_{\boldsymbol{k}}] = VP(k)$ of Section 11.1.3,

$$\hat{P}_g(k) = \frac{|\delta_{g,\boldsymbol{k}}|^2}{V} \overset{(1)}{=} \frac{1}{V\bar{n}^2} \sum_i \sum_j e^{-i\boldsymbol{k}\cdot(\boldsymbol{x}_i - \boldsymbol{x}_j)} \overset{(2)}{=} \frac{1}{V\bar{n}^2}\Big[N + \sum_{i \neq j} e^{-i\boldsymbol{k}\cdot(\boldsymbol{x}_i - \boldsymbol{x}_j)}\Big],$$
$$\langle \hat{P}_g(k) \rangle \overset{(3)}{=} \frac{N}{V\bar{n}^2} + b^2 P(k) \overset{(4)}{=} b^2 P(k) + \frac{1}{\bar{n}}. \tag{T2.116}$$

(1) $\delta_g = n/\bar{n} - 1$, and the -1 only touches $\boldsymbol{k} = 0$, so $\delta_{g,\boldsymbol{k}} = \bar{n}^{-1} \sum_i e^{-i\boldsymbol{k}\cdot\boldsymbol{x}_i}$; the modulus squared is a double sum.
(2) Split off the $i = j$ terms: each is $e^0 = 1$, and there are N of them.
(3) The pairs of different galaxies are the clustering: averaged over realisations, their phases add up to $V\bar{n}^2\, b^2 P(k)$, the power of the field the galaxies trace.
(4) $N = \bar{n}V$.

The self-pairs, each galaxy paired with itself, add a constant $1/\bar{n}$ at every k: a white-noise floor, the **shot noise** (derived in full in Section 11.1.3).

How noisy is the measured $\hat{P}_g$? The Gaussian-field result of Chapter 7 answers it in two steps. *Count the modes.* In a box of volume V the allowed wavevectors sit on a grid, one per $(2\pi)^3/V$ of $\boldsymbol{k}$-space. A shell of radius k and width Δk has volume $4\pi k^2 \Delta k$, so it holds $N_k = 4\pi k^2 \Delta k\, V/(2\pi)^3 = Vk^2\Delta k/(2\pi^2)$ grid points. The field is real, so $\delta_{-\boldsymbol{k}} = \delta^*_{\boldsymbol{k}}$ and only $N_k/2$ of them are independent. *The error of one shell.* One independent mode gives $|\delta_{\boldsymbol{k}}|^2/V$, whose standard deviation equals its mean (it is $\frac{1}{2}\chi^2_2$ times the mean, (7.26)), and the mean is now the total power $b^2P + 1/\bar{n}$. Averaging $N_k/2$ independent modes divides the variance by $N_k/2$ ((7.27)), so $\sigma_P = \sqrt{2/N_k}\,(b^2P + 1/\bar{n})$. Divide by the signal b^2P:

$$\frac{\sigma_P}{b^2P} = \sqrt{\frac{2}{N_k}}\Big(1 + \frac{1}{\bar{n}b^2P}\Big). \tag{T2.117}$$

Two terms limit a survey. Cosmic variance, the $\sqrt{2/N_k}$, falls only with volume. Shot noise is negligible once $\bar{n}b^2P \gtrsim 1$. A survey designer therefore covers as much volume as possible with just enough galaxies to reach $\bar{n}P \approx 1$ on the scales of interest, around $k = 0.1\,h\,\mathrm{Mpc}^{-1}$ [103, Sec. 4.4.5]. DESI's "effective volume" $V_{\mathrm{eff}} = \int [\bar{n}P/(1 + \bar{n}P)]^2 \mathrm{d}V$ is this trade-off in one number [106, Table 1].

If several galaxies share one halo, their pairs add a second constant on large scales, the one-halo term $\int \mathrm{d}m\, n(m) \langle N(N-1) \,|\, m\rangle / \bar{n}^2$ [107, eq. (128)]. It behaves like extra shot noise.

T2.14.3 Redshift-space distortions

The redshift of a galaxy measures its distance plus its peculiar velocity along the line of sight. In comoving coordinates the map position is $\boldsymbol{s} = \boldsymbol{x} + u_z \hat{\boldsymbol{z}}$, with $\hat{\boldsymbol{z}}$ the line of sight and $u_z = v_z/(aH)$ the velocity expressed as a distance. Galaxies falling towards an overdensity from in front are pushed back, those behind are pulled forward, and the structure looks squashed along the line of sight (Figure T2.28).

How strong is the squashing? The inputs are: linear theory ($|\delta| \ll 1$); a distant observer, so the line of sight is the same direction $\hat{\boldsymbol{z}}$ for every galaxy; and velocities set by the growth of structure. That last input needs a short derivation, because the velocity is what moves the galaxies in the map.

The velocities from the growth of structure. Matter that flows out of a region lowers its density. In comoving coordinates, with $\boldsymbol{v}$ the peculiar velocity, mass conservation to first order in δ and

$\boldsymbol{v}$ reads

$$\frac{\partial\delta}{\partial t} \overset{(1)}{=} -\frac{1}{a}\,\nabla\cdot\boldsymbol{v}, \qquad \frac{\partial\delta}{\partial t} \overset{(2)}{=} f H\,\delta \quad\Longrightarrow\quad \nabla\cdot\boldsymbol{v} \overset{(3)}{=} -a f H\,\delta, \qquad \nabla\cdot\boldsymbol{u} \overset{(4)}{=} -f\,\delta. \tag{T2.118}$$

(1) The continuity equation: the outflow rate through the surface of a small comoving volume, per unit volume, is $\nabla\cdot\boldsymbol{v}$ in physical units, and a comoving gradient is a times a physical one, hence the $1/a$.
(2) Linear growth: $\delta(\boldsymbol{x},t) = \delta(\boldsymbol{x},t_0)D(t)$, with D the linear growth factor, so $\partial\delta/\partial t = (\dot D/D)\delta = (\mathrm{d}\ln D/\mathrm{d}\ln a)(\dot a/a)\delta = fH\delta$, which defines the growth rate $f = \mathrm{d}\ln D/\mathrm{d}\ln a \approx \Omega_m(z)^{0.55}$.
(3) Equate the two and multiply by $-a$.
(4) Divide by aH: $\boldsymbol{u} = \boldsymbol{v}/(aH)$ is the velocity expressed as a comoving distance.

In physical coordinates, with $\boldsymbol{r} = a\boldsymbol{x}$, the first relation is $\dot\delta = -\nabla_{\boldsymbol{r}}\cdot\boldsymbol{v}$, the form used for the local void in Section T2.12. The divergence fixes only part of $\boldsymbol{u}$. In linear theory any rotation of the flow decays as the universe expands, so the flow is a gradient, $\boldsymbol{u} = \nabla\Phi_u$, and in Fourier space $\boldsymbol{u}_{\boldsymbol{k}} = i\boldsymbol{k}\,\Phi_{u,\boldsymbol{k}}$ points along $\boldsymbol{k}$. Its divergence is $i\boldsymbol{k}\cdot\boldsymbol{u}_{\boldsymbol{k}} = -k^2\Phi_{u,\boldsymbol{k}}$, and setting this equal to $-f\,\delta_{\boldsymbol{k}}$ gives $\Phi_{u,\boldsymbol{k}} = f\,\delta_{\boldsymbol{k}}/k^2$, so

$$\boldsymbol{u}_{\boldsymbol{k}} = \frac{i\boldsymbol{k}\,f\,\delta_{\boldsymbol{k}}}{k^2}.$$

Now the map. Galaxies are conserved by the mapping from real to redshift space, $(1+\delta_s)\,\mathrm{d}^3s = (1+\delta_g)\,\mathrm{d}^3x$. So

$$\delta_s \overset{(1)}{=} \delta_g - \frac{\partial u_z}{\partial z} \xrightarrow{\text{Fourier}} \delta_{s,\boldsymbol{k}} \overset{(2)}{=} b\,\delta_{\boldsymbol{k}} - ik_z\,\frac{ik_z f}{k^2}\,\delta_{\boldsymbol{k}} \overset{(3)}{=} \left(b + f\mu^2\right)\delta_{\boldsymbol{k}},$$
$$P_s(k,\mu) \overset{(4)}{=} b^2\left(1+\beta\mu^2\right)^2 P(k), \qquad \beta \equiv \frac{f}{b}, \qquad \mu \equiv \frac{k_z}{k}. \tag{T2.119}$$

(1) $\mathrm{d}^3s = (1+\partial u_z/\partial z)\,\mathrm{d}^3x$, because only the z coordinate is moved; expand to first order.
(2) $\delta_g = b\delta$, and $\partial/\partial z \to ik_z$.
(3) $-i\cdot i = 1$ and $k_z^2/k^2 = \mu^2$, the squared cosine of the angle between $\boldsymbol{k}$ and the line of sight.
(4) Square the modulus and average over realisations.

This is the **Kaiser formula** [108]. Waves across the line of sight ($\mu = 0$) are untouched; waves along it are enhanced by $(1+\beta)^2$. Averaging over all directions gives the **monopole**. Directions are spread uniformly in μ between -1 and 1, and the factor is even in μ, so the average is $\int_0^1 \mathrm{d}\mu$. Expand the square and integrate term by term, with $\int_0^1 \mu^2\mathrm{d}\mu = \frac{1}{3}$ and $\int_0^1 \mu^4\mathrm{d}\mu = \frac{1}{5}$:

$$\int_0^1 \left(1+\beta\mu^2\right)^2\mathrm{d}\mu = \int_0^1 \left(1+2\beta\mu^2+\beta^2\mu^4\right)\mathrm{d}\mu = 1 + \frac{2}{3}\beta + \frac{1}{5}\beta^2.$$

Why this separates growth from bias. The amplitude of P is usually written through σ_8 [the rms matter contrast in spheres of $8\,h^{-1}$Mpc], so that $P(k)/\sigma_8^2$ is a fixed shape. Write (T2.119) in that form, using $b(1+\beta\mu^2) = b + f\mu^2$:

$$P_s(k,\mu) = \left(b\sigma_8 + f\sigma_8\,\mu^2\right)^2\,\frac{P(k)}{\sigma_8^2}.$$

The part of the amplitude that does not depend on direction is $b\sigma_8$; the part that grows as μ^2 is $f\sigma_8$. Fitting how the power changes with μ gives the two separately: the angular dependence of the galaxy power spectrum measures the growth rate of structure, independently of the unknown bias. On small scales the random velocities inside haloes do the opposite and stretch groups into "fingers" pointing at the observer; that regime needs models beyond linear theory.

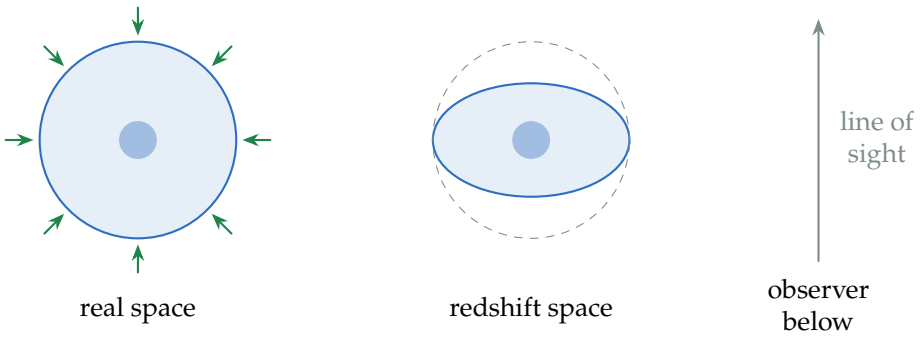


Figure T2.28: The Kaiser effect. Left: a spherical shell of galaxies around an overdensity, falling inwards (green arrows). Right: the same shell placed at the distances inferred from redshifts. Galaxies on the near side move away from us and appear farther; those on the far side move towards us and appear nearer. The shell is flattened along the line of sight; across it nothing changes (the dashed circle is the true shell).

T2.14.4 The experiment: thresholds, bias and squashing in a box

`09_tracers_rsd.py` tests (T2.112), (T2.115) and (T2.119) on a simulated universe. The ingredients:

ingredient	value	what it stands for
box side, grid	$1000\,h^{-1}$Mpc, 128^3 cells	a survey volume, periodic
matter field	Gaussian, linear CAMB $P(k)$ at $z=1$	the initial density, grown linearly
smoothing scale R	$10\,h^{-1}$Mpc, $\sigma_R = 0.229$	the size of a halo: $u_{\rm long}$
small-scale part	independent, rms 1.5	unresolved physics inside haloes: $u_{\rm short}$
total rms σ	1.516	$\sigma^2 = \sigma_R^2 + 1.5^2$
threshold	cells with $u > \nu\sigma$	where galaxies form
growth rate	$f = 0.876$ at $z=1$	sets the velocities

One sentence per ingredient. The box cannot resolve haloes, so the local density that decides where galaxies form is built by hand from the two parts of u: the field smoothed on the scale of a halo is $u_{\rm long}$, the part that correlates with distant matter; the independent small-scale part is $u_{\rm short}$, which by construction correlates with nothing. Cells above the threshold host galaxies. Then every particle and every galaxy is moved by its Zel'dovich displacement $\boldsymbol{\psi}_{\boldsymbol{k}} = i\boldsymbol{k}\delta_{\boldsymbol{k}}/k^2$, the first-order displacement of (T2.114) (it has $-\nabla\cdot\boldsymbol{\psi} = \delta$, so $1/J = 1+\delta$ to first order), and in redshift space by an extra $f\psi_z\hat{\boldsymbol{z}}$, which is the u_z of (T2.118).

Measuring b without cosmic variance. The box has only about a thousand independent modes at $k < 0.05\,h\,\mathrm{Mpc}^{-1}$, and in each of them the auto spectrum $|\delta_{g,\boldsymbol{k}}|^2$ scatters as much as its mean ((7.26)). But the box also holds the matter field that made it, and for each mode $\delta_{g,\boldsymbol{k}} = b\,\delta_{\boldsymbol{k}} + \text{noise}$. Then

$$\frac{\mathrm{Re}\big(\delta_{g,\boldsymbol{k}}\delta^*_{\boldsymbol{k}}\big)}{|\delta_{\boldsymbol{k}}|^2} = b + \frac{\mathrm{Re}(\text{noise}\cdot\delta^*_{\boldsymbol{k}})}{|\delta_{\boldsymbol{k}}|^2},$$

and the random amplitude $|\delta_{\boldsymbol{k}}|^2$ of the particular mode cancels in the first term. The ratio is b in every mode, whatever the realisation drew. The script averages it over $k < 0.05\,h\,\mathrm{Mpc}^{-1}$.

Figure T2.29 shows what came out. Before the move, the measured bias follows b_L of (T2.112): 1.06 ± 0.03 against 1.01 at $\nu = 1$, and 1.71 ± 0.09 against 1.57 at $\nu = 2$. The high-ν limit ν/σ (1.32 at $\nu = 2$) is still poor at these thresholds. After the move the bias is 2.62 ± 0.08 at $\nu = 2$, against $1 + b_L = 2.57$: the matter's own clustering has been added, as (T2.115) says.

A galaxy sample with $\nu = 1.5$ and $\bar{n} = 3\times10^{-4}\,h^3\mathrm{Mpc}^{-3}$ (299091 galaxies) has $b = 2.24 \pm 0.06$ from the cross spectrum (2.28 predicted). Its auto spectrum shows the noise terms of (T2.116).

Subtracting only the Poisson term $1/\bar{n} = 3343\,h^{-3}\mathrm{Mpc}^3$ gives an apparent bias of 2.42, too high. *What is missing?* Galaxies are placed in the selected cells as a Poisson number N with mean $\lambda = 2.13$ per cell, so a cell often holds two or more, and its galaxies pair with one another. Those pairs are at zero separation, like the self-pairs of (T2.116), so they too add a constant to the power. Counting them is one line. A Poisson number has $\mathbb{E}\,N = \mathrm{Var}\,N = \lambda$, so $\mathbb{E}[N(N-1)] = \mathbb{E}\,N^2 - \mathbb{E}\,N = (\lambda + \lambda^2) - \lambda = \lambda^2$. There are $N_{\mathrm{sel}} = \bar{n}V/\lambda$ selected cells, so the number of within-cell pairs is $N_{\mathrm{sel}}\lambda^2 = \bar{n}V\lambda$, and in the same normalisation as the $N/(V\bar{n}^2)$ of (T2.116),

$$\text{one-halo constant} = \frac{N_{\mathrm{sel}}\;\mathbb{E}[N(N-1)]}{V\bar{n}^2} = \frac{\bar{n}V\lambda}{V\bar{n}^2} = \frac{\lambda}{\bar{n}} = 2.13 \times 3343 \approx 7114\,h^{-3}\mathrm{Mpc}^3,$$

the cell version of the one-halo term above. Subtracting it too gives 2.19, in agreement with the cross spectrum. In redshift space the ratio $P_s(k,\mu)/P_r(k)$ rises with μ as $(1+\beta\mu^2)^2$ with $\beta = f/b = 0.391$; the monopole ratio is 1.314 against $1 + \frac{2}{3}\beta + \frac{1}{5}\beta^2 = 1.292$. The highest-$\mu$ bin, which holds the fewest modes, scatters most. For the matter ($b = 1$) one box is not enough to read that bin, so the script repeats the matter part on fresh boxes. Averaged over 11 boxes, the enhancement at $\mu \approx 0.95$ is 3.08 ± 0.01 against 3.21: the Zel'dovich motions have begun to smear the largest line-of-sight displacements, the first sign of the non-linear regime.

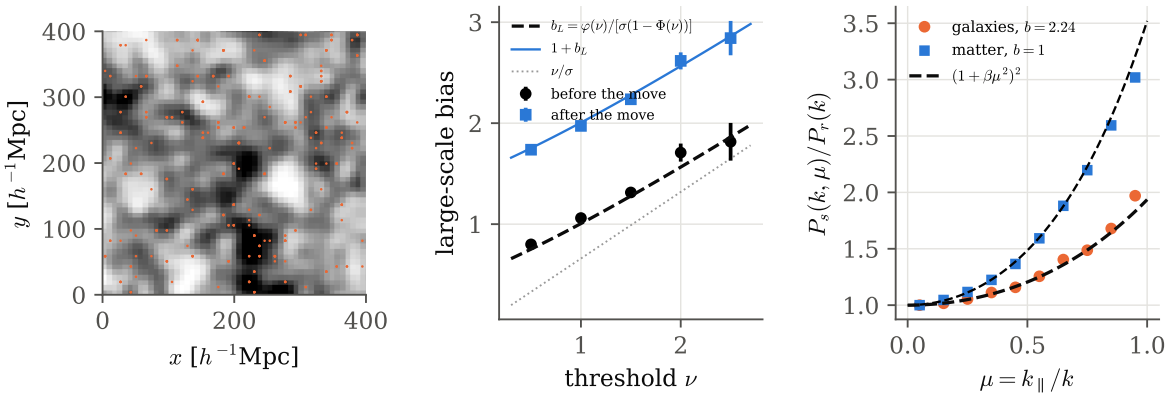


Figure T2.29: Left: a slice of the smoothed density field with the galaxies of one thin slab (orange); the galaxies sit in the dense regions. Middle: the large-scale bias against the threshold ν; before the move it follows b_L (dashed), after it $1 + b_L$ (blue line); the dotted line is the high-threshold limit ν/σ. Right: redshift-space power divided by real-space power, against μ; galaxies (orange) and matter (blue) follow the Kaiser factor with $\beta = f/b$ and $\beta = f$ respectively.

Nerfed, honestly. The box is a single Gaussian realisation of $(1\,h^{-1}\mathrm{Gpc})^3$ with $7.8\,h^{-1}\mathrm{Mpc}$ cells, run in seconds, and the haloes are replaced by a threshold on a field with a white-noise small-scale part. Research analyses use N-body simulations with 10^{10} particles, find haloes in them, populate the haloes with galaxies by an occupation model fitted to the small-scale clustering, and build hundreds to thousands of approximate mock catalogues to estimate the covariance of $\hat{P}_g$.

Where the statistics comes back in. The galaxy power spectrum and correlation function, their estimators and their Gaussian covariance, are the random-field statistics of Chapter 7. Shot noise is the Poisson process of Chapter 2. Bias, the shot-noise amplitude and $f\sigma_8$ are the parameters of a clustering likelihood whose forecast is a Fisher calculation (Chapter 10) and whose posterior is sampled with MCMC (Chapter 9). Galaxy clustering, redshift-space distortions and peculiar velocities as cosmological data are the subject of Chapter 11.

T2.15 Baryon acoustic oscillations: a standard ruler

Before recombination, photons and baryons were held together by Thomson scattering into a single hot fluid. Every small overdensity in that fluid launched a spherical sound wave, which ran outwards until the baryons were released from the photons and then stopped. How fast did it run? *The pressure comes from the photons, the inertia from both.* Radiation has pressure $p = \rho_\gamma c^2/3$; the baryons add mass but, being cold, no pressure. A sound wave squeezes photons and baryons together, so both number densities rise by the same fraction. For photons $\rho_\gamma \propto T^4$ and $n_\gamma \propto T^3$, so $\delta\rho_\gamma/\rho_\gamma = \frac{4}{3}\,\delta n/n$; for baryons $\delta\rho_b/\rho_b = \delta n/n$. Hence $\delta\rho_b = \frac{3}{4}(\rho_b/\rho_\gamma)\,\delta\rho_\gamma$, and

$$c_s^2 = \frac{\delta p}{\delta\rho} = \frac{c^2\,\delta\rho_\gamma/3}{\delta\rho_\gamma + \delta\rho_b} = \frac{c^2}{3(1+R)}, \qquad R \equiv \frac{3\rho_b}{4\rho_\gamma}.$$

Without baryons this is $c/\sqrt{3}$; the baryons' inertia slows the wave a little. *How far did it get?* In a time $\mathrm{d}t$ the wave covers a physical distance $c_s\mathrm{d}t$, which is a comoving distance $c_s\mathrm{d}t/a$. With $a = 1/(1+z)$ and $\mathrm{d}z/\mathrm{d}t = -(1+z)H$, this is $c_s\mathrm{d}t/a = -c_s\,\mathrm{d}z/H$. Summing from the beginning ($z = \infty$) to the moment the baryons stop feeling the photons' push, the **drag epoch** $z_d \approx 1060$ (slightly after recombination, because photons outnumber baryons by about a billion to one and keep dragging them a little longer), each initial peak was left with a thin shell of extra baryons at the comoving distance sound had travelled,

$$r_d = \int_{z_d}^{\infty} \frac{c_s(z)}{H(z)}\,\mathrm{d}z \approx 147\ \mathrm{Mpc} \tag{T2.120}$$

[106, eqs. (2.3)–(2.4)]. Dark matter fell into both the central peak and the shell, and galaxies later formed in both. So each overdensity is surrounded by a shell at r_d, and pairs of galaxies separated by r_d, one in a peak and one in its shell, are slightly more common than chance. Counting pairs by separation, that excess shows up as a bump in the correlation function at $r \approx r_d$, the feature computed in Chapter T1. The integral (T2.120) depends only on the physical matter and baryon densities $\omega_m = \Omega_m h^2$ and $\omega_b = \Omega_b h^2$ (through H and R) and the number of relativistic species $N_{\rm eff}$ (through H). A fit to its numerical value, quoted by DESI, is $r_d \approx 147.05\,(\omega_m/0.1432)^{-0.23}(N_{\rm eff}/3.04)^{-0.1}(\omega_b/0.02236)^{-0.13}$ Mpc [106, eq. (2.5)]; the exponents are fitted numbers, not derived ones. The CMB measures it to 0.2%: $r_d = 147.09 \pm 0.26$ Mpc [25, Table 2]. A feature of known length, seen at many redshifts, is a **standard ruler**.

The question. What exactly does the position of the bump measure, how is it extracted from a noisy correlation function, how precisely, and what did DESI find?

T2.15.1 What the ruler measures

The survey measures angles and redshifts, not lengths. A pair of galaxies at redshift z separated by r_d *across* the line of sight subtends an angle

$$\Delta\theta = \frac{r_d}{D_M(z)}, \qquad D_M(z) = \int_0^z \frac{c\,\mathrm{d}z'}{H(z')}\ \ \text{(flat)},$$

where D_M is the transverse comoving distance [104, eq. (16)]. A pair separated by r_d *along* the line of sight differs in redshift by Δz, and since the comoving distance changes by $\mathrm{d}\chi = c\,\mathrm{d}z/H(z)$,

$$\Delta z = \frac{r_d}{D_H(z)}, \qquad D_H(z) \equiv \frac{c}{H(z)}.$$

So the bump measures D_M/r_d from its angular size and D_H/r_d from its depth in redshift [106, Sec. 2.1]. Figure T2.30 draws both.

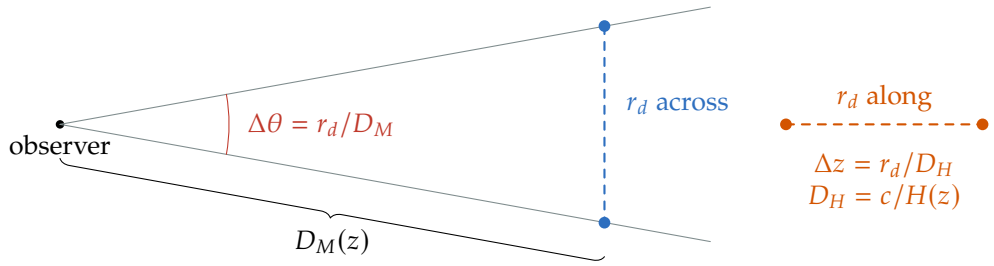


Figure T2.30: What a standard ruler measures. Left: two galaxies separated by r_d across the line of sight appear an angle $\Delta\theta = r_d/D_M$ apart. Right: two galaxies separated by r_d along the line of sight differ in redshift by $\Delta z = r_d/D_H$. The first gives the distance, the second the expansion rate, both in units of r_d.

The map is drawn with an assumed cosmology. To place galaxies in space at all, the survey assumes a fiducial cosmology, with distances $D_M^{\rm fid}$ and $D_H^{\rm fid}$ and a sound horizon $r_d^{\rm fid}$. *If the true distances differ from the assumed ones, where in the fiducial map does the true ring of radius r_d end up, and how do we read the true distances off its position?*

Across the line of sight first. The survey turns a measured angle into a length with the assumed distance, and the angle is set by the true one:

$$r_\perp^{\rm app} \overset{(1)}{=} D_M^{\rm fid}\,\Delta\theta \overset{(2)}{=} r_d\,\frac{D_M^{\rm fid}}{D_M} \overset{(3)}{=} \frac{r_d^{\rm fid}}{\alpha_\perp}, \qquad \alpha_\perp \equiv \frac{D_M/r_d}{D_M^{\rm fid}/r_d^{\rm fid}}. \tag{T2.121}$$

(1) The map converts angle to transverse length with the fiducial distance.
(2) The true pair at separation r_d subtends $\Delta\theta = r_d/D_M$, with the true D_M.
(3) Name the ratio of the apparent ring to the ring the template expects, $r_\perp^{\rm app}/r_d^{\rm fid} = 1/\alpha_\perp$.

Along the line of sight, the same way. A redshift difference becomes a length through the fiducial D_H, and the true one sets $\Delta z = r_d/D_H$: $r_\parallel^{\rm app} = D_H^{\rm fid}\Delta z = r_d D_H^{\rm fid}/D_H = r_d^{\rm fid}/\alpha_\parallel$. Together,

$$\alpha_\perp = \frac{D_M/r_d}{D_M^{\rm fid}/r_d^{\rm fid}}, \qquad \alpha_\parallel = \frac{D_H/r_d}{D_H^{\rm fid}/r_d^{\rm fid}}. \tag{T2.122}$$

So a sphere of radius r_d in the true universe appears in the fiducial map as an ellipsoid with semi-axes $r_d^{\rm fid}/\alpha_\perp$ across the line of sight and $r_d^{\rm fid}/\alpha_\parallel$ along it. Read the direction carefully: $\alpha > 1$ means the true distance (in units of r_d) is larger than assumed, the ring looks *smaller* than the template, and the feature appears at smaller r.

From the ring to the template fit. The template correlation function $\xi_{\rm fid}(r)$ of the fiducial cosmology has its bump at $r = r_d^{\rm fid}$. The data have it at $r^{\rm app} = r_d^{\rm fid}/\alpha$. The dilated template $\xi_{\rm fid}(\alpha r)$ has its bump where $\alpha r = r_d^{\rm fid}$, that is at $r = r_d^{\rm fid}/\alpha$, exactly where the data put it. So fitting the data with $\xi_{\rm fid}(\alpha r)$ and finding the best α measures the ratio of true to fiducial distance in (T2.122). DESI reports the same information in the form D_M/r_d and D_H/r_d [106, Sec. 2.1], which is our $\alpha_\perp$ and $\alpha_\parallel$ multiplied by the fiducial $D_M^{\rm fid}/r_d^{\rm fid}$ and $D_H^{\rm fid}/r_d^{\rm fid}$.

Why D_V, and why the factor z. When the data are too noisy to separate the two directions, one fits the angle-averaged correlation function with a single dilation α. The single scale that preserves the volume of the ellipsoid, which is what an angle average responds to at first order, is the geometric mean of its three axes:

$$\alpha_{\rm iso} \overset{(1)}{=} \left(\alpha_\perp^2\,\alpha_\parallel\right)^{1/3} \overset{(2)}{=} \frac{\left(D_M^2 D_H\right)^{1/3}/r_d}{\left(D_M^{\rm fid\,2} D_H^{\rm fid}\right)^{1/3}/r_d^{\rm fid}} \overset{(3)}{=} \frac{D_V/r_d}{D_V^{\rm fid}/r_d^{\rm fid}}, \qquad D_V \equiv \left(z\,D_M^2\,D_H\right)^{1/3}. \tag{T2.123}$$

(1) Two transverse axes and one radial axis: the volume scales as $\alpha_\perp^2 \alpha_\parallel$.
(2) Substitute (T2.122).
(3) Multiply numerator and denominator by $z^{1/3}$, which cancels.

The factor z is a convention, and it is there so that D_V is a distance with a familiar limit: at low redshift $D_M \approx cz/H_0$ and $D_H \approx c/H_0$, so $D_V \approx \left(z \cdot (cz/H_0)^2 \cdot c/H_0\right)^{1/3} = cz/H_0$, the Hubble-law distance. This is Eisenstein et al.'s eq. (2), $D_V = [D_M^2\, cz/H(z)]^{1/3}$ [42, Sec. 4], and DESI's eq. (2.6) [106].

Watch out: Which "angular diameter distance"?

Three conventions meet here. Hogg's $D_A = D_M/(1+z)$ is the physical angular diameter distance [104, eq. (18)], and so is the D_A of Chapter T1. Weinberg et al. write D_A for the *comoving* angular diameter distance, which is D_M [103, p. 46]. Eisenstein et al. and DESI write D_M. A factor $(1+z)$ in a BAO formula is usually this.

T2.15.2 Measuring the dilation, and how well

A correlation function is measured in bins of separation r_i. How noisy is each bin, and how are neighbouring bins tied together? For a Gaussian field the answer follows from the power spectrum. The correlation function is the Fourier transform of the power, so the estimate in bin i is a weighted sum of the power measured in thin k-shells, $\hat{\xi}(r_i) = \sum_{\text{shells}} \frac{k^2\Delta k}{2\pi^2} \hat{P}(k)\, \bar{j}_0(kr_i)$, where $\bar{j}_0$ is the spherical Bessel function $\sin x/x$ averaged over the width of the bin (Section 7.3.2). Different shells are independent, and each has $\mathrm{Var}\,\hat{P} = 2P_{\text{tot}}^2/N_k$ with $P_{\text{tot}} = b^2P + 1/\bar{n}$ and $N_k = Vk^2\Delta k/(2\pi^2)$ modes ((T2.117)). So

$$\mathrm{Cov}\left[\hat{\xi}(r_i), \hat{\xi}(r_j)\right] \overset{(1)}{=} \sum_{\text{shells}} \left(\frac{k^2\Delta k}{2\pi^2}\right)^2 \bar{j}_0(kr_i)\, \bar{j}_0(kr_j)\, \frac{2P_{\text{tot}}^2}{Vk^2\Delta k/(2\pi^2)}$$
$$\overset{(2)}{=} \frac{2}{V} \int \frac{k^2 \mathrm{d}k}{2\pi^2} \left[b^2P(k) + \frac{1}{\bar{n}}\right]^2 \bar{j}_0(kr_i)\, \bar{j}_0(kr_j). \tag{T2.124}$$

(1) The covariance of two weighted sums of independent terms is the sum of weight times weight times variance.
(2) Cancel one factor $k^2\Delta k/(2\pi^2)$ and let the shells become thin.

Each piece now has a meaning. The $1/V$ is the averaging over independent modes: a bigger survey has more of them. The square is there because a variance of a power is the square of the power (one mode's $|\delta_k|^2$ scatters as much as its mean). The $1/\bar{n}$ inside it is the shot noise of (T2.116). And since every bin draws on the same shells, neighbouring bins share their errors.

The model is the fiducial template, dilated, plus a smooth curve that absorbs everything the BAO feature does not determine [106, p. 9]:

$$\xi_{\text{model}}(r) = B^2\, \xi_{\text{fid}}(\alpha r) + a_0 + \frac{a_1}{r} + \frac{a_2}{r^2}. \tag{T2.125}$$

Each term has one job. α is the dilation of (T2.122), the only number we want. B^2 absorbs the amplitude: the galaxy bias b^2 of Section T2.14 relative to the template's, and the redshift-space boost. a_0, a_1/r and a_2/r^2 absorb the broadband shape, whatever smooth errors the template has from non-linear growth or a bias that changes slowly with scale. None of them can make a narrow bump, so the bump's position, and with it α, is left to the data.

Solving for the nuisance terms. For a fixed trial α the model is a fixed matrix X (its four columns are $\xi_{\text{fid}}(\alpha r_i)$, 1, $1/r_i$, $1/r_i^2$) times the vector $\boldsymbol{\beta} = (B^2, a_0, a_1, a_2)$. So $\chi^2(\boldsymbol{\beta}) = (\hat{\boldsymbol{\xi}} - X\boldsymbol{\beta})^{\mathsf{T}} \mathbf{C}^{-1} (\hat{\boldsymbol{\xi}} - X\boldsymbol{\beta})$

is quadratic in $\boldsymbol{\beta}$, and a quadratic has one minimum, found exactly by generalised least squares (Chapter 3), $\hat{\boldsymbol{\beta}} = (X^{\mathsf{T}}\mathbf{C}^{-1}X)^{-1}X^{\mathsf{T}}\mathbf{C}^{-1}\hat{\boldsymbol{\xi}}$. Marginalising $\boldsymbol{\beta}$ with flat priors instead (Chapter 5) integrates $e^{-\chi^2/2}$ over a Gaussian in $\boldsymbol{\beta}$, which gives $e^{-\chi^2_{\min}/2}$ times $\det(X^{\mathsf{T}}\mathbf{C}^{-1}X)^{-1/2}$, the same structure as the amplitude integral of (T2.128); that factor changes only slowly with α, so the two routes agree closely. What is left is a profile $\chi^2(\alpha)$.

Is the bump really there? The **BAO detection significance** compares the profile with the best fit of a template with the wiggles removed: $\Delta\chi^2 = \chi^2_{\text{no wiggle}} - \chi^2_{\text{BAO}}$. The null hypothesis is the smooth, wiggle-free correlation function, and the test asks: if there were no bump, would a $\Delta\chi^2$ this large, or larger, be something we would reasonably expect from noise? Think of the BAO model as the no-wiggle model plus the wiggles with a free amplitude; the null sets that one amplitude to zero. Then Wilks' theorem (Section 4.7, (4.30)) says that under the null $\Delta\chi^2$ is a χ^2 with one degree of freedom, which is the square of one unit Gaussian Z. So

$$\mathrm{P}\left(\Delta\chi^2 > n^2 \,|\, \text{no bump}\right) = \mathrm{P}(|Z| > n) = 2\big[1 - \Phi(n)\big],$$

the probability of a Gaussian fluctuation beyond n standard deviations. A measured $\Delta\chi^2$ therefore corresponds to a detection at $n = \sqrt{\Delta\chi^2}$ standard deviations. (The amplitude being one extra parameter is an approximation: α too is free only in the BAO model, which makes the real tail a little heavier.)

Example: the first detection. *Their question.* In 2005 the acoustic bump had been predicted for decades but never seen in galaxies. Eisenstein et al. asked: is there a bump near $100\,h^{-1}$Mpc in the correlation function of galaxies, and if so, what distance does its position give?

Their setup. They used luminous red galaxies (LRGs) from SDSS: 46,748 of them at $0.16 < z < 0.47$ over 3816 square degrees. LRGs are rare and bright, so they can be seen far away and fill a large volume, which beats down the cosmic variance of (T2.117); and they live in massive haloes, so their bias is high and the bump in $\xi_g = b^2\xi$ stands out against the shot noise. The covariance of $\hat{\xi}$ came from 1278 mock catalogues, which play the role of our (T2.124) with the survey's real shape and selection built in.

Their key equation. With galaxies at a range of redshifts and a correlation function averaged over directions, the bump measures the single dilation α_{iso} of (T2.123), that is D_V/r_d at the effective redshift $z = 0.35$. Their eq. (2), $D_V = [D_M^2\, cz/H(z)]^{1/3}$, is our $D_V = (zD_M^2 D_H)^{1/3}$ with $D_H = c/H$.

Their result, through our recipe. The best BAO model had $\chi^2 = 16.1$ for 17 degrees of freedom; the best model without the acoustic peak had 27.8. The significance recipe above gives

$$\Delta\chi^2 = 27.8 - 16.1 = 11.7, \qquad n = \sqrt{11.7} = 3.4,$$

a 3.4σ detection. The peak position gave $D_V(0.35) = 1370 \pm 64\,$Mpc, so $64/1370 = 4.7\%$ [42, Sec. 4, Fig. 6]. Our mock survey below, with a somewhat larger volume and a denser sample, finds a median detection of 3.8σ and a scatter of 0.0272 in α, about 3%: the same order, a little better, as it should be for more volume and more galaxies.

The experiment. `10_bao_ruler.py` needs two templates and a survey.

Why the bump is smeared. Between the drag epoch and today each galaxy has moved from where it formed, by some $8\,h^{-1}$Mpc in each direction, pulled by the large-scale flows. Treat that displacement $\boldsymbol{\psi}$ as a random Gaussian vector with variance Σ^2_{nl} per direction. A wave $e^{i\boldsymbol{k}\cdot\boldsymbol{x}}$ sampled at the displaced positions is multiplied, on average, by $\mathbb{E}[e^{i\boldsymbol{k}\cdot\boldsymbol{\psi}}] = e^{-k^2\Sigma^2_{\text{nl}}/2}$, the characteristic function of a Gaussian. The acoustic wiggles live at $k \gtrsim 0.05\,h\,\text{Mpc}^{-1}$, where this

factor bites, so they fade; in $\xi(r)$ the narrow bump is smoothed into a broader one. The value $\Sigma_{\rm nl} = 8\,h^{-1}$Mpc is the one before reconstruction (the step that real surveys use to move galaxies back towards where they started).

What "no wiggle" means. The comparison template is a power spectrum with the same broadband shape as the real one but without the oscillations. Eisenstein and Hu give a smooth transfer function for that [109, eqs. (26), (28)–(31)]. Calling the CAMB linear spectrum $P_{\rm lin}$ and the smooth one $P_{\rm nw}$, the difference $P_{\rm lin} - P_{\rm nw}$ is the BAO signal alone, and the template damps only that part: $P_{\rm w} = P_{\rm nw} + (P_{\rm lin} - P_{\rm nw})\, e^{-k^2\Sigma_{\rm nl}^2/2}$.

The survey, about one DESI redshift bin:

quantity	value	enters through
volume V	$1\,(h^{-1}\text{Gpc})^3$	the $1/V$ of (T2.124)
number density $\bar{n}$	$3\times10^{-4}\,h^3\text{Mpc}^{-3}$	the shot noise $1/\bar{n}$
bias b	2	the signal b^2P
separation bins	20, from 50 to 150 h^{-1}Mpc	the data vector $\hat{\boldsymbol{\xi}}$
mock surveys	10000	draws from $\mathcal{N}(\xi_{\rm true}, \mathbf{C})$

It draws the mock $\hat{\xi}$ vectors with $\mathbf{C}$ from (T2.124), and fits each one:

Listing T2.11: The profile $\chi^2(\alpha)$ with the linear nuisance parameters solved at each α (`code/chT2/10_bao_ruler.py`)

```
80 alphas = np.linspace(0.8, 1.2, 401)
81 # design matrices: [xi_fid(alpha r), 1, 1/r, 1/r^2]; the linear parameters are solved
       exactly at each alpha
82 basis = lambda a, wig: np.column_stack([template(rc, a, wig), np.ones_like(rc), 1 / rc, 1 /
       rc**2])
83 PROJ = {}
84 for wig in (True, False):
85     PROJ[wig] = []
86     for a in alphas:
87         X = basis(a, wig)
88         F = X.T @ CINV @ X
89         PROJ[wig].append((X, np.linalg.solve(F, X.T @ CINV)))
90
91
92 def chi2_curve(d, wig=True):
93     out = np.empty(alphas.size)
94     for i, (X, P) in enumerate(PROJ[wig]):
95         res = d - X @ (P @ d)
96         out[i] = res @ CINV @ res
97     return out
98
99
100 def fit_alpha(d):
101     cw, cn = chi2_curve(d, True), chi2_curve(d, False)
102     i = np.argmin(cw)
103     inside = alphas[cw <= cw[i] + 1]
104     return alphas[i], 0.5 * (inside.max() - inside.min()), cw.min(), cn.min(), cw, cn
```

Lines 83–89 compute, for every trial α, the design matrix X and the matrix that turns data into the best nuisance values,

$$P_\alpha = (X^{\mathsf T}\mathbf{C}^{-1}X)^{-1}X^{\mathsf T}\mathbf{C}^{-1}, \qquad \hat{\boldsymbol{\beta}} = P_\alpha\hat{\boldsymbol{\xi}}.$$

The residual is then $\hat{\boldsymbol{\xi}} - X\hat{\boldsymbol{\beta}} = (1 - XP_\alpha)\,\hat{\boldsymbol{\xi}}$, and

$$\chi^2(\alpha) = \hat{\boldsymbol{\xi}}^{\mathsf T}(1 - XP_\alpha)^{\mathsf T}\,\mathbf{C}^{-1}\,(1 - XP_\alpha)\,\hat{\boldsymbol{\xi}}.$$

The data enter only at the two ends; everything in between depends on α through X and on nothing else. So these projection matrices are computed once per trial α and reused for all 10000 mocks, and each fit is a matrix product (line 95). Line 103 reads the $\Delta\chi^2 = 1$ interval off the profile.

Figure T2.31 shows one mock: $\hat{\alpha} = 1.030 \pm 0.026$, $\chi^2_{\rm min} = 11.8$ for 15 degrees of freedom, and a no-wiggle penalty $\Delta\chi^2 = 13.7$. Over all mocks the median detection is 3.8σ (16–84% range 2.7–4.8σ), comparable to the first SDSS detection. The estimates $\hat{\alpha}$ average 0.9997: the ruler is unbiased. Their scatter, 0.0272, is larger than both the Fisher error 0.0207 (the width predicted from the curvature of the profile $\chi^2(\alpha)$ of noise-free data at its minimum, Chapter 10) and the average $\Delta\chi^2 = 1$ half-width 0.0218, and the $\Delta\chi^2 = 1$ interval covers the truth in 66% of the mocks (±0.5%, the binomial error) instead of 68%.

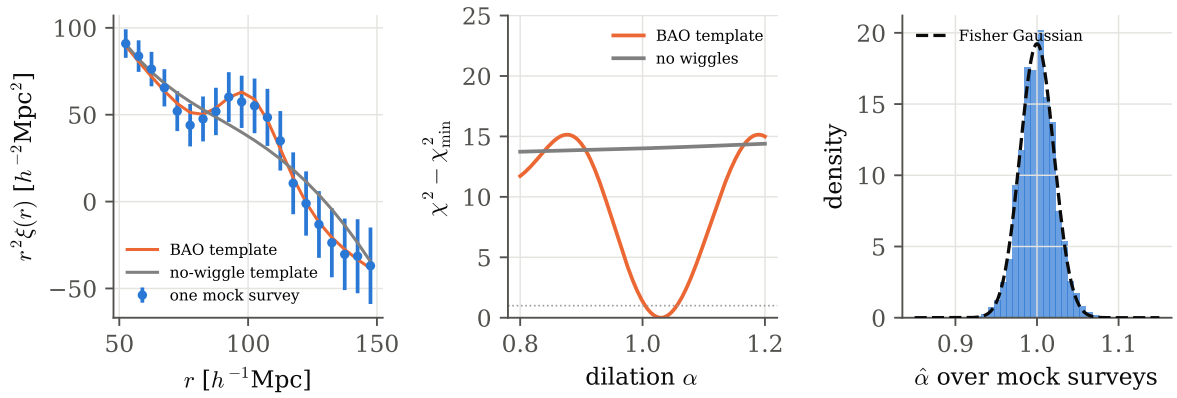


Figure T2.31: Fitting the BAO dilation. Left: one mock correlation function (times r^2, with the diagonal errors) and the best fits of the BAO template and of the no-wiggle template. Middle: the profile $\chi^2(\alpha)$ for both; the BAO template has a clear minimum near $\alpha = 1$, the no-wiggle one is flat and higher. Right: the distribution of $\hat{\alpha}$ over the mock surveys and the Gaussian of the Fisher error; the tails are heavier than the Gaussian.

Why the heavier tails? In a mock where noise happens to weaken the bump, $\chi^2(\alpha)$ is shallow and can have a second minimum where a noise wiggle lines up with the template. The likelihood is then far from Gaussian in α, and the curvature at the minimum underestimates the real spread. At a 4σ detection this happens often; at the 10–20σ of a DESI bin it does not, and DESI finds its posteriors "well approximated by Gaussian distributions" [106, Sec. 2.3.2].

How this is used in real research: Verification versus research

Here the mocks *verify* the analysis: we know $\alpha = 1$, and we check that the estimator is unbiased and that its error bar covers the truth at the stated rate. In research the same machinery is the *instrument*. DESI estimates the covariance of $\hat{\xi}$ from large sets of mock catalogues and analytic models, because the real survey has a mask, varying completeness and non-linear clustering that (T2.124) ignores; it calibrates the reconstruction and the theoretical systematics (below 0.1–0.2%) on simulations; and it blinds the data, shifting the BAO position in the catalogues until the pipeline is frozen [106, Sec. 2.3.1]. A full analysis uses thousands of mocks; ours uses 10000 Gaussian draws from an analytic covariance and no reconstruction.

T2.15.3 What DESI's distances say

Their question. DESI measured the ruler at seven redshifts in its first year. The question we reproduce is: *do the BAO distances alone, with no CMB and no supernovae, fix the matter density* Ω_m

and the expansion scale $H_0 r_d$ in flat ΛCDM, and do they agree with the ΛCDM that fits the CMB? It matters because the ruler is calibrated by early-universe physics and read at $z < 2.5$, a test of the model across most of cosmic time.

Their data. Each bin gives D_M/r_d and D_H/r_d with their correlation coefficient where the signal is strong, and D_V/r_d alone for the bright galaxies (BGS) and the quasars (QSO), where it is weak [106, Table 1]:

tracer	$z_{\rm eff}$	D_M/r_d	D_H/r_d	D_V/r_d	correlation
BGS	0.295	–	–	7.93 ± 0.15	–
LRG	0.510	13.62 ± 0.25	20.98 ± 0.61	–	−0.445
LRG	0.706	16.85 ± 0.32	20.08 ± 0.60	–	−0.420
LRG+ELG	0.930	21.71 ± 0.28	17.88 ± 0.35	–	−0.389
ELG	1.317	27.79 ± 0.69	13.82 ± 0.42	–	−0.444
QSO	1.491	–	–	26.07 ± 0.67	–
Lyα	2.330	39.71 ± 0.94	8.52 ± 0.17	–	−0.477

Five bins with two numbers and two bins with one make $5\times2+2 = 12$ numbers. The correlation is negative because the data pin down the overall size of the ring better than its shape: a fit that stretches the ring across the line of sight compensates by shrinking it along the line of sight.

The model, derived. In flat ΛCDM at these redshifts (radiation is 0.1% of H^2 even at $z = 2.3$) the predictions are

$$\frac{D_H}{r_d}(z) = \frac{c}{H_0 r_d}\frac{1}{E(z)}, \qquad \frac{D_M}{r_d}(z) = \frac{c}{H_0 r_d}\int_0^z \frac{\mathrm{d}z'}{E(z')}, \qquad E(z) = \sqrt{\Omega_m(1+z)^3 + 1 - \Omega_m}. \tag{T2.126}$$

Only two numbers appear: Ω_m, which sets how the distances change with redshift, and the product $H_0 r_d$, which sets their overall scale. BAO alone therefore cannot separate H_0 from r_d.

Where we have met H_0 before. The mock supernovae of Section 3.17, made with $H_0 = 70$, fixed only $M - 5\log_{10} H_0$, and H_0 came out only once M was supplied from outside. The ladder of Section T2.12 supplied it from parallaxes and Cepheids, and the real ladder gave 73.04 ± 1.04, 4.8σ above the CMB's 67.36 ± 0.54 (Section T2.12.3). The same degeneracy is back here with a ruler in place of the candle: the distances fix only the product $H_0 r_d$, as the magnitudes fixed only $M - 5\log_{10} H_0$, and the ruler's length plays the role of the candle's luminosity. Three things differ. The data are real DESI distances at $0.3 < z < 2.3$, far beyond the $z < 0.15$ of the Hubble-flow supernovae, so Ω_m must be fitted alongside the scale rather than entering as the small q_0 correction. The calculation is a Bayesian posterior on a grid rather than a least-squares fit. And the calibration comes from the early Universe: r_d is set by the sound waves before recombination, the same physics that fixes the CMB value of H_0 (Section 10.6.3), not by geometry nearby. That is why the H_0 found below lands near the CMB value and not near the ladder's.

What we compute. Each bin is a measurement $\boldsymbol{d}_b$ (two numbers or one) with Gaussian noise, and the model $\boldsymbol{m}_b(\Omega_m, h r_d)$ is (T2.126) evaluated at $z_{\rm eff}$ (with $D_V = (z D_M^2 D_H)^{1/3}$ of (T2.123) for the one-number bins). The bins are independent, so the χ^2 is a sum over bins,

$$\chi^2(\Omega_m, h r_d) = \sum_{b=1}^{7} \big(\boldsymbol{d}_b - \boldsymbol{m}_b\big)^{\mathsf{T}} C_b^{-1} \big(\boldsymbol{d}_b - \boldsymbol{m}_b\big), \qquad C_b = \begin{pmatrix} \sigma_M^2 & \rho\,\sigma_M\sigma_H \\ \rho\,\sigma_M\sigma_H & \sigma_H^2 \end{pmatrix}, \tag{T2.127}$$

where σ_M and σ_H are the quoted errors of D_M/r_d and D_H/r_d and ρ is their correlation; for a D_V bin C_b is just σ_V^2. The posterior with flat priors on a grid of (Ω_m, hr_d) is Bayes' theorem (Chapter 5) with a constant prior:

$$p(\Omega_m, hr_d \mid \boldsymbol{d}) = \frac{e^{-\chi^2(\Omega_m, hr_d)/2}}{\sum_{\text{grid}} e^{-\chi^2/2}},$$

and the means, standard deviations and correlation are sums over the grid weighted by it.

DESI and ours, side by side. Figure T2.32 shows the distances and the posterior; the numbers are

	DESI [106, eqs. (4.1), (4.3)]	ours
Ω_m	0.295 ± 0.015	0.296 ± 0.015
hr_d [Mpc]	101.8 ± 1.3	101.8 ± 1.3
H_0 [km/s/Mpc], with the CMB r_d	69.29 ± 0.87	69.2 ± 0.9

with correlation -0.92 between our Ω_m and hr_d. The best fit has $\chi^2 = 12.8$ for $12 - 2$ degrees of freedom, an ordinary value.

Why they differ, and by how much. They barely do: the largest difference, 0.001 in Ω_m, is a fifteenth of the error bar. Three things could cause differences of that size. Our grid has steps of 0.001 in Ω_m and 0.1 Mpc in hr_d, where DESI samples its posterior with MCMC. Our Gaussian is built from the errors and correlations as quoted in the table, rounded to two or three digits. And we neglect radiation, which changes H^2 by at most 0.1% at the highest redshift, the Lyα bin, so D_H by at most 0.05%, far below its 2% error. For H_0, DESI also propagates the 0.2% error of r_d, which we leave out; it is small next to the 1.3% from hr_d.

Is Planck's ΛCDM consistent with it? The Planck best fit ($hr_d = 99.1$ Mpc) has $\chi^2 = 21.1$ against our minimum 12.8, so $\Delta\chi^2 = 21.1 - 12.8 = 8.3$. Under the hypothesis that the Planck point is the truth, $\Delta\chi^2$ between it and the best fit is a χ^2 with two degrees of freedom, one per fitted parameter (Wilks' theorem, Section 4.7). Its tail is $\mathrm{P}(\chi^2_2 > v) = e^{-v/2}$ ((10.61)), so $p = e^{-8.3/2} = 0.016$. Converted to the two-sided Gaussian language of the detection test, $2[1 - \Phi(n)] = 0.016$ gives $n = 2.4\sigma$. That treats the Planck point as exact. DESI, including the errors of the CMB values, which spread the Planck point into its own small ellipse, quantifies the difference as a 1.9σ discrepancy (1.6σ without CMB lensing) [106, eq. (4.2)]: a mild tension, not a contradiction.

The lesson. Twelve published numbers and a two-parameter grid are enough to recover DESI's headline constraint; what the grid cannot do on its own is separate H_0 from r_d, which is the subject of the next example.

Example: the inverse distance ladder. Take $r_d = 147.09$ Mpc from the CMB. Then $h = (hr_d)/r_d$, and the posterior above gives $H_0 = 69.2 \pm 0.9$ km/s/Mpc; DESI, propagating the small error of r_d too, finds 69.29 ± 0.87, and 68.53 ± 0.80 with r_d computed from a big-bang-nucleosynthesis value of ω_b instead of the CMB [106, eqs. (4.3)–(4.4)]. This is a distance ladder run backwards, anchored at $z \approx 1060$ by the physics of sound waves rather than at $z \approx 0$ by parallaxes. It lands near the CMB value and well below the 73.04 ± 1.04 of the Cepheid ladder of Section T2.12: the Hubble tension appears again when a ruler replaces a candle.

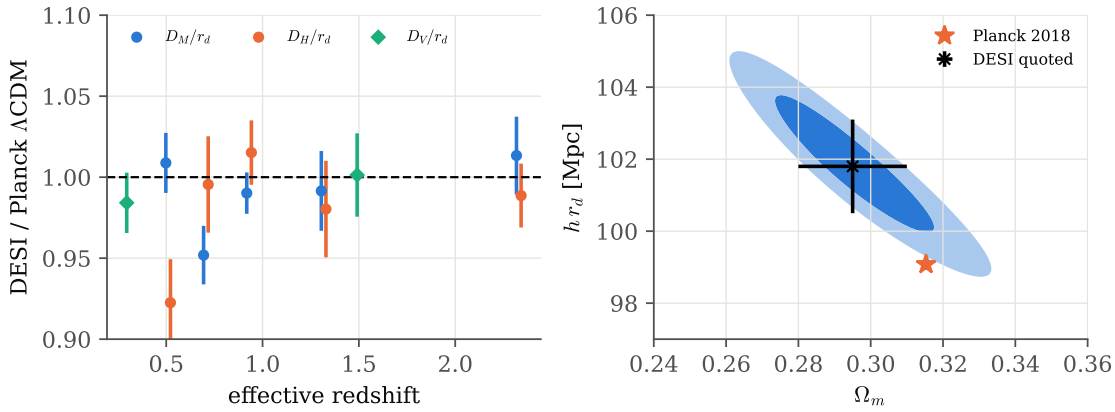


Figure T2.32: Left: DESI DR1 BAO distances divided by the Planck ΛCDM prediction. Right: the 68% and 95% regions in (Ω_m, hr_d) from the twelve DESI numbers, computed on a grid, with DESI's quoted result (cross) and the Planck best fit (star). The long, tilted contour is the degeneracy of (T2.126): a higher Ω_m can be traded for a lower hr_d.

Where the statistics comes back in. The estimator and covariance of $\hat{\xi}$ and $\hat{P}$ are Chapter 7. The template fit is least squares with linear nuisances (Chapter 3), the detection is a likelihood-ratio test (Chapter 4), and the posterior on (Ω_m, hr_d) is the Bayesian grid of Chapter 5, replaced by MCMC (Chapter 9) when more parameters are free. The forecast precision of a BAO survey is a Fisher calculation (Chapter 10), and galaxy clustering as cosmological data is Chapter 11.

T2.16 Photometric redshifts and their error distributions

A spectrum gives a redshift to a part in ten thousand, but it costs a fibre and many minutes of a large telescope for each galaxy. Imaging surveys such as the ones used for weak lensing (Chapter T1) detect a billion galaxies, far too many and mostly too faint for spectra. What they have is the flux of each galaxy in five or six broad filters. That is enough for a rough redshift, because galaxy spectra have sharp features. The strongest in the optical is the break near 4000 Å, where old stars and metal absorption make the light drop by a factor of two or three to the blue. At high redshift the Lyman series of intergalactic hydrogen removes the light blueward of 1216 Å. As a galaxy is moved to higher redshift its spectrum slides to the red, and the break passes from one filter into the next (Figure T2.33). The ratio of fluxes in neighbouring filters, a **colour**, therefore encodes the redshift. An estimate made this way is a **photometric redshift** (photo-z).

The question. How is a redshift computed from a few fluxes, what is the distribution of its error, and how does that error propagate into the redshift distribution $n(z)$ of a sample, which is what cosmology actually uses?

T2.16.1 Template fitting as a likelihood

What we observed is a vector of fluxes F_b in bands $b = 1, \dots, 6$, each with a Gaussian error σ_b. *What could have produced it?* A galaxy of some spectral type T, at some redshift z, with some overall brightness A, predicting fluxes $A\,T_b(z)$, where $T_b(z)$ is the template spectrum redshifted to z and integrated through filter b. *How likely are the observed fluxes under each* (z, T)*?* The brightness is not of interest, so we integrate it out with a flat prior. The χ^2 is quadratic in A,

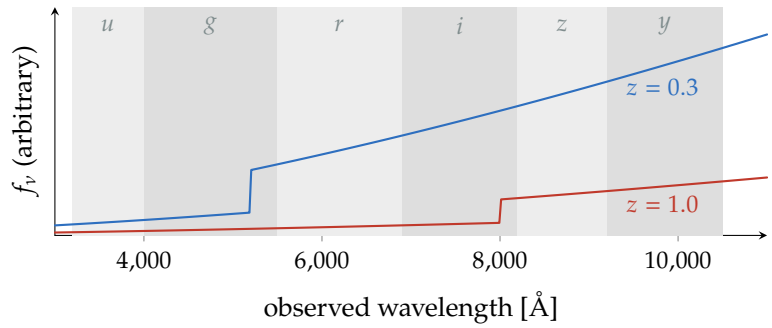


Figure T2.33: The same red-galaxy spectrum at two redshifts, seen through six broad filters (grey bands). At $z = 0.3$ the 4000 Å break lands at 5200 Å, inside the g band; at $z = 1.0$ it has moved to 8000 Å, inside the i band. Which filter holds the break, and how deep the drop is, is what a photometric redshift reads.

$\chi^2(A) = \chi^2_{\min} + (A - \hat{A})^2 S$ with $S = \sum_b T_b^2/\sigma_b^2$, so

$$\mathcal{L}(z,T) \overset{(1)}{=} \int \mathrm{d}A \, \exp\Big[-\frac{1}{2}\sum_b \frac{(F_b - AT_b)^2}{\sigma_b^2}\Big] \overset{(2)}{=} e^{-\chi^2_{\min}/2} \int \mathrm{d}A \, e^{-S(A-\hat{A})^2/2} \overset{(3)}{=} \sqrt{\frac{2\pi}{S}}\, e^{-\chi^2_{\min}(z,T)/2},$$

$$\chi^2_{\min} \overset{(4)}{=} \sum_b \frac{F_b^2}{\sigma_b^2} - \frac{\left(\sum_b F_b T_b/\sigma_b^2\right)^2}{S}, \qquad p(z \mid \boldsymbol{F}) \propto n(z) \sum_T \mathcal{L}(z,T). \tag{T2.128}$$

(1) Independent Gaussian errors in each band; marginalise the nuisance amplitude (Chapter 5).
(2) Expand the square in A and complete it; $\hat{A} = \sum_b F_b T_b \sigma_b^{-2}/S$ is the least-squares amplitude.
(3) The Gaussian integral $\int e^{-Sx^2/2}\mathrm{d}x = \sqrt{2\pi/S}$.
(4) Substitute $\hat{A}$ into χ^2.

The last expression is Bayes' theorem with the redshift distribution $n(z)$ of the population as the prior and the type summed over. The sum gives every template the same weight: it is a flat prior over the types, even though the population the experiment below draws from is 35% red and 65% blue. A type prior would multiply each $\mathcal{L}(z,T)$ by the fraction of galaxies of type T. The photo-z is then usually the peak of $p(z \mid \boldsymbol{F})$, but the whole posterior is the honest answer.

T2.16.2 The error distribution

The distribution of the error $z_{\mathrm{phot}} - z$ has a roughly Gaussian core whose width grows with redshift as $\sigma_z \propto 1 + z$. This scaling follows from the way broad filters locate a spectral feature. A feature at rest wavelength λ_{rest} appears at $\lambda_{\mathrm{obs}} = (1+z)\lambda_{\mathrm{rest}}$, so $\delta(1+z)/(1+z) = \delta\lambda_{\mathrm{obs}}/\lambda_{\mathrm{obs}}$. Since a filter's width is roughly proportional to its central wavelength, the fractional uncertainty $\delta\lambda_{\mathrm{obs}}/\lambda_{\mathrm{obs}}$ is approximately constant, and then $\sigma_z = \delta z$ is a constant times $1 + z$. Good surveys reach $\sigma_z/(1+z) \approx 0.02$–$0.05$. Outside the core, the distribution also contains **catastrophic outliers** with $|z_{\mathrm{phot}} - z| \sim 1$. The main cause is a degeneracy: a sharp drop observed at a wavelength λ_{obs} may be the 4000 Å break of a nearby galaxy or the Lyman-α edge of a distant one, $(1 + z_{\mathrm{low}})\,4000$ Å $= (1 + z_{\mathrm{high}})\,1216$ Å. For $z_{\mathrm{low}} = 0.16$,

$$1 + z_{\mathrm{high}} = 1.16 \times \frac{4000}{1216} = 3.82, \qquad z_{\mathrm{high}} = 2.8,$$

so a galaxy at $z_{\mathrm{high}} = 2.8$ can be mistaken for one at $z_{\mathrm{low}} = 0.16$ [103, p. 96]. Filters on both sides of both features, from u to the near infrared, break the degeneracy.

Weinberg et al. describe the error by the joint distribution $P(z_p, z)$ of photo-z and true redshift, and summarise it by the conditional bias and scatter $\delta z(z_p) = z_p - \langle z \rangle_{z_p}$ and $\sigma_z(z_p)$ [103, p. 94, eq. (100)]. They then make a point that is pure Bayes: $P(z \mid z_p) = P(z)\,P(z_p \mid z)/P(z_p)$ [103, eq. (101)], so a photo-z that is unbiased in the sense $\langle z_p \rangle_z = z$ can still have $\delta z(z_p) \neq 0$.

A picture makes the reason plain (Section 5.3). Think of each true redshift z as an urn, and of the photo-z as a ball drawn from it: the urn at z gives out balls spread around z, with $P(z_p \mid z)$ saying how readily it gives the ball labelled z_p. The population $n(z)$ says how many urns of each kind there are. We are handed one ball, labelled z_p, and asked which urn it came from. Two things decide: how readily each kind of urn gives this ball, and how many urns of that kind there are. Where urns are plentiful, even a modest chance per urn adds up to many balls labelled z_p, so the answer is pulled towards the crowded side. The ball tells us about its urn; it does not change the urns. This is the shrinking sample space once more: knowing the label z_p discards every draw with a different label, and what survives is weighted by the number of urns that produced it (Section 5.2). It is the same structure as the Malmquist calculation of Section T2.13, where the urns were absolute magnitudes and their number was set by the volume. How large is the difference? Take a Gaussian error, $P(z_p \mid z) = \mathcal{N}(z_p; z, \sigma^2)$, and a population whose number rises locally as $P(z) \propto e^{sz}$. Then

$$P(z \mid z_p) \overset{(1)}{=\!\propto} e^{sz} \exp\left[-\frac{(z - z_p)^2}{2\sigma^2}\right] \overset{(2)}{=\!\propto} \exp\left[-\frac{(z - z_p - s\sigma^2)^2}{2\sigma^2}\right], \qquad \langle z \rangle_{z_p} \overset{(3)}{=} z_p + s\sigma^2. \quad \text{(T2.129)}$$

(1) Bayes' theorem; $P(z_p)$ does not depend on z.
(2) Complete the square, exactly as in the Malmquist calculation (T2.109).
(3) The mean of the Gaussian.

Where $n(z)$ rises, galaxies observed at z_p are more likely to have scattered *up* from below, since there were more of them below; where it falls, the opposite. It is the Malmquist bias of Section T2.13 with redshift in place of magnitude: the population we draw from reweights the measurement.

Why the mean of $n(z)$ is the number that matters. A weak-lensing analysis bins galaxies by photo-z and needs the true $n(z)$ of each bin. The lensing signal depends mostly on its mean redshift. A lens bends the light of a source more the farther the source lies behind it (the lensing efficiency grows with the source distance, Chapter T1). So a bin whose mean is estimated too high predicts too much lensing for a given clustering amplitude; to match the observed shear, the fit lowers the amplitude, and S_8 comes out low [S_8 is the clustering amplitude that weak lensing constrains best]. Stage-III surveys need the mean redshift of each bin to about 0.01, and Euclid and LSST need it 5–10 times better [105, Sec. V.B.2]. Outliers matter in proportion to their number. Suppose a fraction f of a high-redshift bin are really foreground galaxies, in front of the lenses and so not lensed by them. The mean shear of the bin is then $(1 - f)$ times the true one, and the shear power spectrum, which is quadratic in the shear, is $(1 - f)^2 \approx 1 - 2f$ times. For $f = 1\%$ that is 1% and 2% [103, p. 94]. Photometric redshifts can also be used for BAO, but only the transverse measurement survives, and only if $\sigma_z/(1 + z) \lesssim 4\%$ [103, pp. 62–63]. The line-of-sight measurement is lost much earlier. A redshift error equivalent to a velocity σ_v, that is $\sigma_z = (1 + z)\sigma_v/c$, smears positions along the line of sight by $c\sigma_z/H(z) = (1 + z)\sigma_v/H(z)$. For $\sigma_v = 1000\,\mathrm{km/s}$ at $z = 1$, with $H(1) \approx 120\,\mathrm{km/s/Mpc}$, that is $2 \times 1000/120 \approx 17\,\mathrm{Mpc}$, comparable to or larger than the width of the acoustic bump, so its radial position, and with it $H(z)$, is erased [103, pp. 62–63]. The distribution $P(z \mid z_p)$ is calibrated with spectroscopic subsamples (10^5 spectra for Stage IV) or with the angular cross-correlation of photometric galaxies with a spectroscopic survey. Why does the cross-correlation measure $P(z \mid z_p)$? Both samples trace the same matter, $\delta_{\mathrm{spec}} = b_{\mathrm{spec}}\delta$ and $\delta_{\mathrm{phot}} = b_{\mathrm{phot}}\delta$ (Section T2.14). Galaxies at different true redshifts are too far apart to be correlated, so the photometric galaxies selected by

z_p correlate with a thin spectroscopic slice at z only through the fraction of them that really sit at z. Their cross-correlation is therefore $b_{\rm spec}\, b_{\rm phot}\, P(z \mid z_p)$ times the matter correlation at z [103, p. 97]: the bias of Section T2.14 returns as a nuisance.

The experiment. `11_photoz.py` builds a mock imaging survey:

ingredient	choice	why
galaxies	20000, from $n(z) \propto z^2 e^{-(z/0.55)^{1.5}}$	a typical deep-imaging redshift distribution
types	35% red, 65% blue	two templates, strong and weak 4000 Å break
spectra	power law, 4000 Å break, Lyman absorption	the features a photo-z reads
mismatch	each galaxy's slope and break differ a little	templates are never exactly right
filters	six top-hats, u to y (Figure T2.33)	broad bands
depths	5σ limits 24.0–25.4 mag	realistic noise
redshift grid	800 values, $0 < z < 4$	where (T2.128) is evaluated

It computes the six fluxes, adds noise, and evaluates (T2.128) on the grid:

Listing T2.12: The template likelihood, with the amplitude and the type marginalised, and the two posteriors (`code/chT2/11_photoz.py`)

```
65 def likelihood(f):
66     """L(z) for each galaxy: Gaussian flux errors, amplitude marginalised (flat prior),
       summed over types."""
67     w = 1 / sig**2
68     chis, norms = [], []
69     for T in TEMPL.values():
70         Tw = T * w                                          # (nz, 6)
71         a = f @ Tw.T                                        # sum f T / sigma^2   (ngal,
       nz)
72         b = (T * Tw).sum(1)                                 # sum T^2 / sigma^2   (nz,)
73         chis.append((f**2 * w).sum(1)[:, None] - a**2 / b[None, :])   # chi^2 at the best
       amplitude
74         norms.append(1 / np.sqrt(b))                        # width of the amplitude
       integral
75     cmin = np.minimum(*chis).min(1, keepdims=True)          # common offset, for numerical
       safety
76     return sum(np.exp(-0.5 * (c - cmin)) * n[None, :] for c, n in zip(chis, norms))
77
78
79 post_flat = np.zeros((NGAL, ZG.size))
80 for s in range(0, NGAL, 2000):
81     L = likelihood(fobs[s:s + 2000])
82     post_flat[s:s + 2000] = L / L.sum(1, keepdims=True)
83 post_prior = post_flat * nz[None, :]
84 post_prior /= post_prior.sum(1, keepdims=True)
85 zp_flat = ZG[post_flat.argmax(1)]
86 zp_prior = ZG[post_prior.argmax(1)]
```

Line 73 is $\chi^2_{\rm min}$ of (T2.128), line 74 the width $\sqrt{1/S}$ of the amplitude integral, and line 76 the sum over types. Line 83 multiplies by the prior $n(z)$.

Two summary numbers describe the errors, both measured on $\Delta z \equiv (z_p - z)/(1+z)$, the error in the units in which it is roughly constant. A galaxy is an **outlier** if $|\Delta z| > 0.15$, that is $|z_p - z| > 0.15(1+z)$: an error far too large for the Gaussian core, typically a degeneracy. The width of the core is measured on the remaining galaxies by the **normalised median absolute**

deviation, NMAD = $1.48 \times \text{median}\,|\Delta z - \text{median}(\Delta z)|$. The median is used instead of the mean so that a few large errors cannot inflate it, and the factor 1.48 makes it equal to σ for a Gaussian (the median of $|x|$ for a unit Gaussian is $0.674 = 1/1.48$).

With a flat prior the core width is NMAD = 0.018, and 9.8% of the galaxies are outliers.

With the $n(z)$ prior, the cause first. A degenerate galaxy has a likelihood with two peaks, one at low and one at high redshift. The prior multiplies each peak by the number of galaxies at that redshift, and $n(z)$ is far larger at $z \lesssim 1$ than at $z \approx 3$; so the peak on the poorly populated side is down-weighted, and when it was a spurious peak the error is removed. Then the size: the core barely changes (0.019), but the outliers drop from 9.8% to 4.3%. The price is paid by the rare galaxies that really sit on the poorly populated side: the high-redshift galaxy of Figure T2.34 ($z = 2.00$) has a bimodal likelihood whose higher peak is right, and the prior moves it to the wrong, low-redshift peak, because the prior is doing exactly what it should for the typical galaxy with that colour, and this one is not typical.

The mean of a bin. For the bin $0.8 < z_p < 1.2$ (6199 galaxies) the true mean redshift is 0.965 and the mean photo-z is 0.975, an error of 0.01, at the Stage-III requirement. Stacking the posteriors instead of the point estimates gives 0.975, no better. Why no better? A posterior accounts for the photometric noise, but every posterior here is computed with the same two templates, and every real galaxy differs from its template in the same kind of way. That template mismatch shifts all the posteriors together, and averaging many of them cannot remove an error they share. The asymmetry of (T2.129) is visible where $n(z)$ is steep: at $z = 0.5$, on the rising side, $\langle z_p \rangle_z = 0.531$ but $\langle z \rangle_{z_p} = 0.500$; at $z = 1.5$, on the falling side, 1.449 against 1.407.

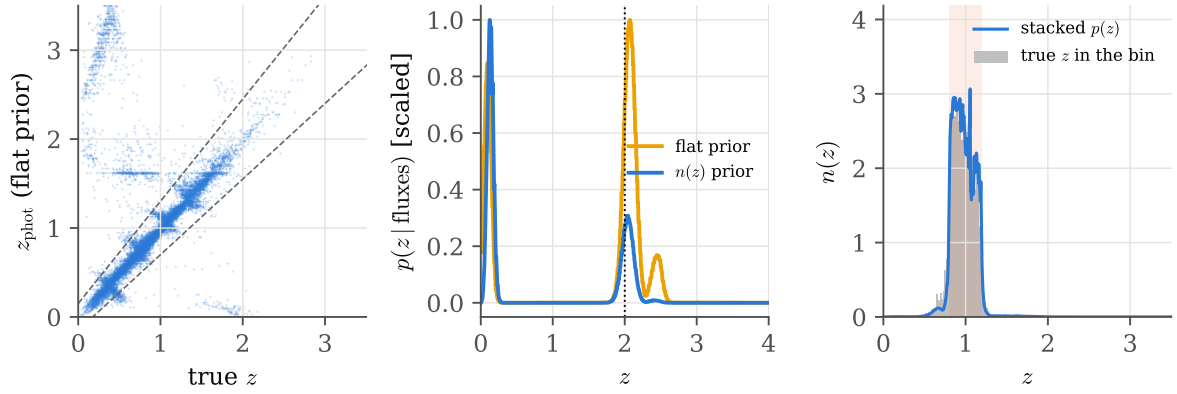


Figure T2.34: Left: photo-z against true redshift (flat prior); the dashed lines bound the core $|z_p - z| < 0.15(1+z)$, and the clumps outside are the degeneracies. Middle: the redshift posterior of one galaxy at $z = 2.0$ (dotted line) with a flat prior and with the $n(z)$ prior. Right: the true redshifts of the galaxies selected by $0.8 < z_p < 1.2$ (grey) and the sum of their posteriors (blue).

Nerfed, honestly. Two toy templates, top-hat filters and a population independent of magnitude make the experiment run in a second. Real pipelines use hundreds of templates or machine-learning estimators trained on spectroscopic samples, fold in magnitude-dependent priors, and calibrate $n(z)$ per bin with spectroscopic subsamples, clustering redshifts and self-organising maps of colour space; the residual uncertainty of each bin's mean is then carried as a nuisance parameter.

Where the statistics comes back in. The photo-z posterior is Bayes' theorem with a nuisance amplitude marginalised (Chapter 5). The core-plus-outliers error distribution is a mixture model, fitted with the likelihood methods of Chapter 3. The shift of the mean of each bin is

the nuisance parameter that weak-lensing likelihoods marginalise (Chapter T1), sampled with MCMC (Chapter 9), and its required accuracy is set by a Fisher forecast (Chapter 10).

Where each item is used. The standardised supernova distance (T2.98) with the noise budget (T2.99) is the datum of the H_0 problem of Chapter 3 and of every supernova likelihood after it. The ladder (T2.106) is a worked generalised least-squares fit whose covariance is the inverse Fisher matrix of Chapter 10. The selection-aware likelihood (T2.111) is the template for any analysis of a selected sample. The bias, shot noise and Kaiser factor (T2.112)–(T2.119) turn the matter $P(k)$ of Chapter 7 into the galaxy $P_g(k,\mu)$ that Chapter 11 fits. The BAO distances (T2.123) and (T2.126) are a compact, nearly Gaussian likelihood used with the samplers of Chapter 9. And the photo-z posterior (T2.128) and its mean-redshift bias are the redshift nuisances of the weak-lensing likelihoods of Chapter T1.

Recap

- Type Ia supernovae are standardisable: $\mu = m_B - M + \alpha x_1 - \beta c$ reduces a 0.4 mag scatter to ≈ 0.12 mag, a 6% distance per supernova; noisy colours dilute the fitted β by $V_c/(V_c+\sigma_c^2)$, and peculiar velocities add $2.17\,\sigma_v/cz$ mag.
- Supernovae measure only $M_B - 5\log_{10} H_0$. The ladder (geometry → Cepheids or TRGB → calibrator supernovae → Hubble flow) is one linear model; each rung adds its variance, and lower-rung errors never average down.
- A flux limit biases a sample: $\langle M\rangle_{\rm det} = M_0 - 1.382\sigma^2$ in Euclidean space, $-\sigma\phi(t)/\Phi(t)$ near the limit; the cure is the likelihood of the selected data, $\mathcal{N}/\Phi(t)$.
- Galaxies trace matter with a bias, $\delta_g = b\delta$, $b = 1 + b_L$ with $b_L = \phi(\nu)/[\sigma(1-\Phi(\nu))]$ for threshold tracers; their power has shot noise $1/\bar{n}$; in redshift space $P_s = b^2(1+\beta\mu^2)^2P$, $\beta = f/b$.
- The BAO bump at $r_d \approx 147$ Mpc measures D_M/r_d and D_H/r_d, or D_V/r_d; in flat ΛCDM DESI's distances fix $\Omega_m = 0.295 \pm 0.015$ and $hr_d = 101.8 \pm 1.3$ Mpc, and with the CMB r_d an inverse ladder gives $H_0 \approx 69$.
- Photometric redshifts have a core of width $\sim 0.02(1+z)$ and catastrophic outliers from the Lyman/4000 Å degeneracy; $\langle z\rangle_{z_p} \neq z_p$ even when $\langle z_p\rangle_z = z$, and a bin's mean redshift must be known to $\lesssim 0.01$.

Chapter summary

Each astrophysical object of the chapter enters the statistics the same way: what is measured, where its noise lives, and which tool turns the measurement into physics.

Compact binaries and gravitational waves

concept	operational meaning	used later in
chirp, chirp mass	$\dot{f} = \frac{96}{5}\pi^{8/3}(G\mathcal{M}/c^3)^{5/3}f^{11/3}$ from energy balance; only $\mathcal{M} = (m_1 m_2)^{3/5}/M^{1/5}$ at leading order	Chapter 10
stationary-phase waveform	$\tilde{h} = \mathcal{A}f^{-7/6}e^{-i\Psi(f)}$, $\mathrm{d}^2\Psi/\mathrm{d}f^2 = 2\pi/\dot{f}$: the signal model of the likelihood	Chapters 0, 5 and 10
LISA noise $S_n(f)$	acceleration noise at low f, metrology and arm response at high f, Galactic foreground near a mHz; the covariance of Gaussian detector noise	Chapters 0 and 10
signal-to-noise ratio	$\rho^2 = 4\int \lvert\tilde{h}\rvert^2/S_n\,\mathrm{d}f$; phase resolution $\sim 1/\rho$ rad	Chapter 10
dark-matter spike	$\rho \propto r^{-(9-2\gamma_i)/(4-\gamma_i)}$ around a growing black hole; fixes the physical meaning of the environmental parameter	Chapter 10
power-law dephasing	extra power $\propto f^{-n}P_{\mathrm{GW}}$ gives $\delta\Psi = \beta u^b$, $b = -5/3 - n$; one Fisher calculation per exponent	Chapters 9 and 10

Standard candles, distances and galaxy surveys

concept	operational meaning	used later in
Tripp standardisation	$\mu = m_B - M + \alpha x_1 - \beta c$; Hubble residual Gaussian in magnitude, $\sigma \approx 0.12$ mag	Chapters 3 and 5
regression dilution	least squares with noisy regressors shrinks a slope by $V/(V+\sigma^2)$; cure: latent variables in a hierarchical model	Chapters 3, 5 and 9
distance ladder	one linear model through shared parameters M_W, M_B; covariance $(A^{\mathsf{T}}\mathbf{C}^{-1}A)^{-1}$; the H_0–M_B degeneracy broken by geometry	Chapters 3, 4 and 10
Malmquist bias	detected candles brighter by $1.382\sigma^2$; near a limit $-\sigma\phi(t)/\Phi(t)$; likelihood of the selected data divides by $\Phi(t)$	Chapters 3, 5 and 6
linear bias	$\delta_g = b\delta$, $P_g = b^2P$ on large scales; threshold and halo-model values; a nuisance in every clustering fit	Chapters 7 and 11
shot noise	$1/\bar{n}$ added to P_g; error $\sqrt{2/N_k}(1 + 1/\bar{n}P)$ per band; one-halo constant	Chapters 2, 7 and 10
Kaiser formula	$P_s(k,\mu) = b^2(1+\beta\mu^2)^2P$; monopole $1 + \frac{2}{3}\beta + \frac{1}{5}\beta^2$; measures $f\sigma_8$	Chapter 11
BAO ruler	$\Delta\theta = r_d/D_M$, $\Delta z = r_d/D_H$, $D_V = (zD_M^2D_H)^{1/3}$; dilation α from a template fit with broadband nuisances	Chapters 4, 7, 9 and 11
DESI BAO likelihood	Gaussian in $(D_M/r_d, D_H/r_d)$ per bin; depends on Ω_m and $H_0 r_d$ only; inverse ladder with r_d	Chapters 5 and 9
photometric redshift	template likelihood with amplitude marginalised, times an $n(z)$ prior; core $\propto 1+z$ plus outliers; $P(z \mid z_p)$ by Bayes	Chapters T1 and 5
mean-redshift nuisance	shift of each bin's $\langle z\rangle$, required to $\lesssim 0.01$; marginalised in lensing likelihoods	Chapters T1, 9 and 10

Reading the literature

- **Inspirals and waveforms.** The quadrupole formula and the Newtonian chirp are in Maggiore's textbook [81]; the stationary-phase waveform and the precision of the chirp mass in Cutler and Flanagan [87]; parametrised deviations from general relativity in Yunes and Pretorius [91].
- **LISA.** The sensitivity curve and its ingredients in Robson, Cornish and Liu [85]; the detector response and Doppler modulation in Barack and Cutler [84].
- **Dark-matter environments.** Spikes around black holes from Gondolo and Silk [96] and Sadeghian et al. [97]; dephasing of intermediate-mass-ratio inspirals in Eda et al. [82], with

halo feedback in Coogan et al. [89] and relativistic corrections in Speeney et al. [88]; wave dark matter in Hui et al. [98, 102], Kim et al. [86], Chase et al. [100] and Hancock and Witek [99]; the broader landscape of environmental effects in Barausse et al. [92]; dynamical friction in Chandrasekhar [101]; halo profiles in Navarro, Frenk and White [93, 94] and Correa et al. [95].

- **Cosmological distances.** Hogg's short note [104] defines every distance used here.
- **Probes of the expansion.** Weinberg et al. [103] review supernovae (Sec. 3), BAO (Sec. 4) and weak lensing with photometric redshifts (Sec. 5.4.3), with their systematics.
- **Supernovae and the ladder.** The Pantheon+ compilation [26] and the SH0ES measurement [24]; the tensions and their proposed resolutions, with sections on ladder systematics and photometric redshifts, in Abdalla et al. [105]; the CMB values in Planck 2018 [25].
- **Galaxy clustering.** The halo model and halo bias in Cooray and Sheth [107]; redshift-space distortions in Kaiser [108].
- **BAO.** The first detection in Eisenstein et al. [42]; the no-wiggle transfer function in Eisenstein and Hu [109]; the DESI first-year cosmology in [106].

Chapter T3

CMB Morphology and Topology

Where this chapter is going

On a map of the microwave sky, some hot regions are isolated while others join into ridges, and a large cold region can surround a hot island. The angular power spectrum C_ℓ of Chapter 7 measures the mean squared amplitude of the fluctuations at each multipole. The spectrum does not specify which hot regions connect on a particular map. For a Gaussian, statistically isotropic sky, however, it determines how often each kind of pattern occurs across the ensemble (Section 7.13). This gives us a reference against which to test shape measurements. We can ask how a small primordial skewness, measured by $f_{\rm NL}$, changes the shapes predicted by inflation, or how gravitational lensing alters them by shifting points on the sky by a few arcminutes. When lensing is a nuisance for the measurement at hand, we want to identify shape statistics that survive those shifts and understand how smoothing and noise affect them.

The polarization of the microwave background adds a second pattern. At every point the polarization is a short stick with a direction but no arrowhead. Where the polarized intensity drops to zero the stick has no direction at all, and around such a point the sticks turn by half a revolution. These points are the polarization singularities. They can be counted, their number per square degree follows from the spectrum, and a smooth remapping such as lensing moves them but cannot create or destroy them one at a time.

The ideas connect as in Figure T3.1. The shapes of a temperature map are read through its *excursion sets*, the regions hotter than a chosen threshold, and through three numbers for each set: its area, the length of its boundary and its Euler characteristic, which counts hot regions minus holes (Section T3.1). For a Gaussian sky all three are known in closed form, and they depend on C_ℓ through just two numbers, the variance of the map and the variance of its gradient. A non-Gaussian sky moves the curves, but part of the move is only a relabelling of the threshold, and only what survives the relabelling is information about non-local physics; lensing, by contrast, leaves the one-point distribution of the map exactly as it was (Section T3.2). The polarization singularities have a half-integer index, and their density follows from the spectrum by a counting argument (Section T3.3). Lensing remaps the sky smoothly and invertibly, which preserves both the topology of every excursion set and every singularity with its index; what breaks the protection is anything applied after the remapping, such as the beam (Section T3.4). *Planck* used exactly these statistics, together with the bispectrum, to test whether the sky is Gaussian and isotropic, and found it consistent with both apart from a handful of large-angle anomalies (Section T3.5). The same counting, applied in three dimensions to the matter distribution, describes the cosmic web (Section T3.6).

In the language of the book's pipeline (Chapter 0): the physical model is the primordial field plus everything that happens to the photons on the way to us. The data set is the map. What we choose to keep is the connectivity of its hot and cold regions and the zeros of its

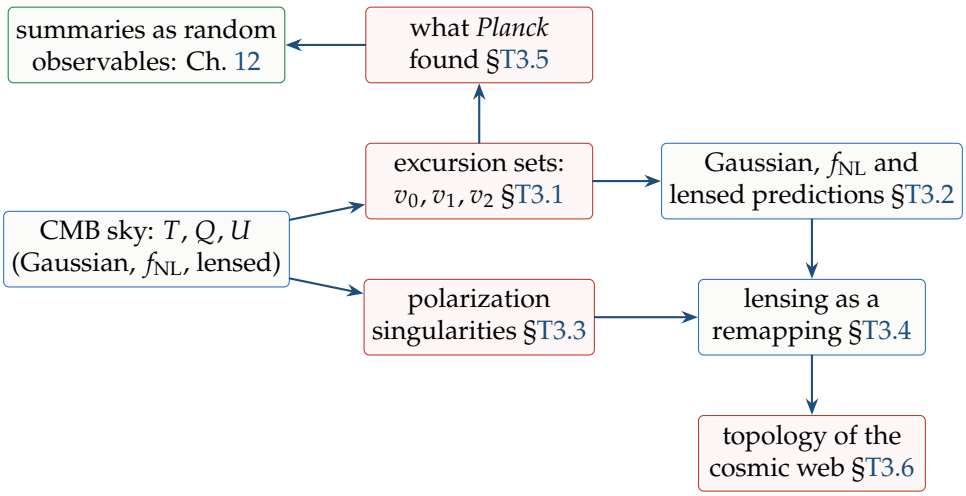


Figure T3.1: How the ideas of the chapter connect. Blue boxes are physics: the sky, what a Gaussian, a non-Gaussian or a lensed sky predicts for its shapes, and what lensing does geometrically. Red boxes are the things one measures on a map: the three Minkowski functionals of the excursion sets, the polarization singularities, the published *Planck* tests, and the same tools applied to the three-dimensional matter distribution. The green box is the statistics: Chapter 12 treats every one of these summaries as a random variable with a sampling distribution, and tests and fits with it.

polarization. Each such summary has a sampling distribution, known in closed form for a Gaussian sky and obtained by simulation otherwise, and Chapter 12 uses those distributions to infer physics. The cosmology that produces each spectrum is in the cosmology notes.

T3.1 The shape of a temperature map

Take a temperature map of a patch of sky, like the left panel of Figure T3.2, and paint every pixel hotter than one standard deviation. What is left is a scatter of islands: 73 more separate hot pieces than holes in this patch. Lower the threshold to one standard deviation *below* the mean and paint again. Now almost everything is painted, and the unpainted pixels form cold lakes inside one connected hot continent: the count of pieces minus holes has become -84. Somewhere in between, at the mean, the painted and unpainted regions are two interlocked networks, neither of which is a sea around the other. This change of connectivity with the threshold is the morphology of the map. Two maps with the same C_ℓ can have different morphologies (the phase-scrambling experiment of Section 7.13 built one), so a statistic of connectivity is a probe that the spectrum does not replace.

The question. If we colour every point of the sky hotter than ν standard deviations, how much of the sky is coloured, how long is the coastline of the coloured region, and how many separate pieces does it have once its holes are subtracted? And for a Gaussian sky with a given C_ℓ, what should each of those three numbers be at each ν?

T3.1.1 Excursion sets and their three numbers

Write the temperature fluctuation in direction $\hat{\boldsymbol{n}}$ as $T(\hat{\boldsymbol{n}})$, measured from the mean, and its standard deviation over the sky as σ_0 (the subscript says it is the variance of the field itself, with no derivatives). The region of the sky where the field exceeds ν standard deviations is the **excursion set** at threshold ν,

$$Q_\nu = \{\hat{\boldsymbol{n}} \,:\, T(\hat{\boldsymbol{n}}) > \nu\,\sigma_0\}. \tag{T3.1}$$

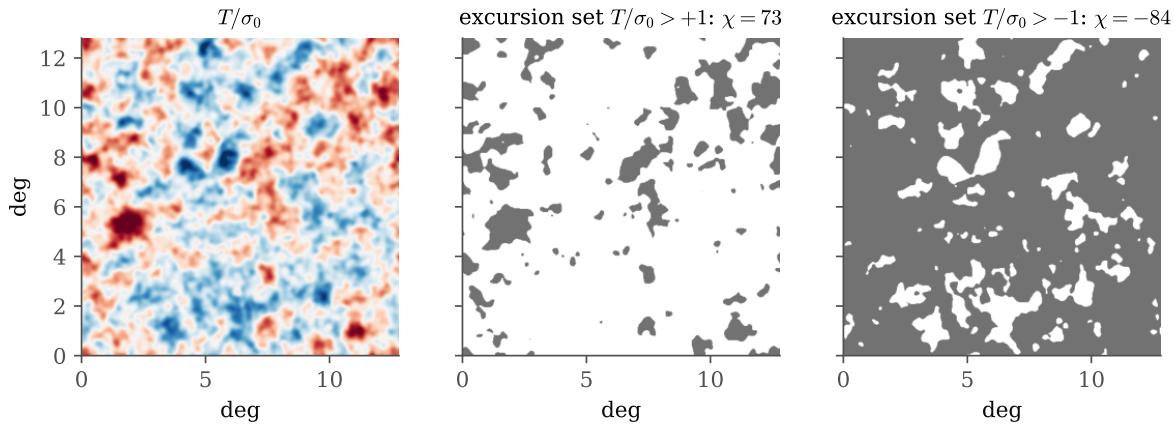


Figure T3.2: Left: a simulated $12.8° \times 12.8°$ patch of the CMB temperature, smoothed by a 10′ beam, in units of its standard deviation σ_0. Middle: the pixels hotter than $+1\sigma_0$ form separate islands, with very few holes. Right: the pixels hotter than $-1\sigma_0$ form one connected region pierced by cold lakes, so holes outnumber pieces and the Euler characteristic is negative.

Figure T3.2 shows two of them. In two dimensions a classical theorem of integral geometry (Hadwiger's) says that any summary of the shape of a region that adds up over disjoint pieces and does not care where the region sits or how it is turned is a combination of just three numbers [110]: the area, the length of the boundary, and the Euler characteristic. These are the three **Minkowski functionals**.

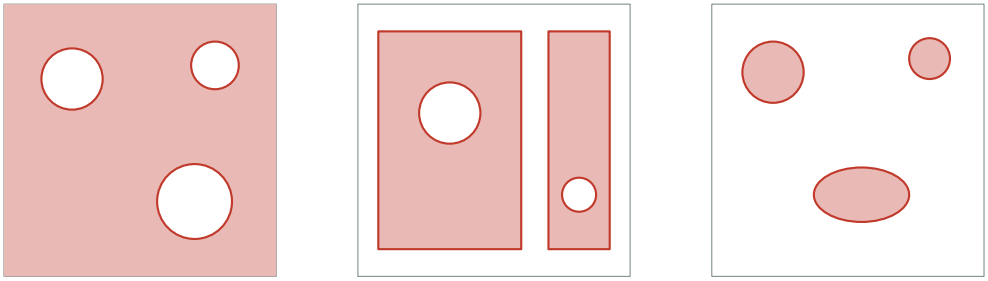
Working definition T3.1 (the Minkowski functionals of a map)

For a map T on a patch of area A and a threshold ν, divide by the area to get densities:

- $v_0(\nu)$, the *area fraction* [the fraction of pixels with $T > \nu\sigma_0$];
- $v_1(\nu) = \frac{1}{4A} \times$ the total length of the boundary of Q_ν [the factor $\frac{1}{4}$ is a convention of integral geometry];
- $v_2(\nu) = \frac{1}{A} \times \chi(Q_\nu)$, where the *Euler characteristic* χ is the number of separate pieces of Q_ν minus the number of holes in them [a hole is a cold region completely surrounded by hot sky].

Operational: v_0 is a pixel count. v_1 is the average of $|\nabla T|\,\delta(T - \nu\sigma_0)$ over the map, with the delta replaced by a narrow bin. χ is counted on the pixel grid as vertices minus edges plus faces of the union of the hot pixels (see the topology notes for why this alternating count gives pieces minus holes).

Why these three and not others? Each answers a physical question about the coloured region. The area fraction says how much of the sky is hotter than the threshold. The boundary length says how wiggly the coastline is, which is a statement about how fast the field changes in space. The Euler characteristic is the one number that does not change when the region is stretched or bent without tearing: it sees only connectivity. Its value is the same whether the island is round or elongated, so it is the most robust of the three to anything that distorts the map smoothly, a point that returns in Section T3.4. On the full sky there is one more fact, from the Gauss–Bonnet theorem: a region that grows until it covers the whole sphere ends with $\chi = 2$ (the Euler characteristic of a sphere), not 0, so the full-sky count contains a small term proportional to the area, $\chi = V_2 + V_0/(2\pi R^2)$ in the notation of [110] (their eq. 7), where V_0 is the area, V_2 the integrated boundary curvature divided by 2π and R the radius of the sphere.



$\nu = -1$: one continent, three lakes
$\chi = 1 - 3 = -2$

$\nu = 0$: two pieces, two holes
$\chi = 2 - 2 = 0$

$\nu = +1$: three islands
$\chi = 3 - 0 = 3$

Figure T3.3: Cartoon of excursion sets at three thresholds (shaded: hotter than the threshold). At a low threshold the hot region is one piece with cold lakes in it; at a high threshold it is a set of separate islands. The Euler characteristic, pieces minus holes, runs from negative to positive as the threshold rises.

T3.1.2 What a Gaussian sky predicts

The area first. The area fraction is the share of the sky where $T > \nu\sigma_0$. On a statistically homogeneous sky the average of a quantity over the sky equals its average over the ensemble of possible skies (ergodicity, Section 7.5). So the expected area fraction is the probability that the field at one point exceeds the threshold. For a Gaussian field of mean zero and variance σ_0^2,

$$\mathbb{E}[v_0(\nu)] \overset{(1)}{=} \mathrm{P}\left(T(\hat{\boldsymbol{n}}) > \nu\sigma_0\right) \overset{(2)}{=} \int_{\nu\sigma_0}^{\infty} \frac{e^{-t^2/2\sigma_0^2}}{\sqrt{2\pi}\,\sigma_0}\,\mathrm{d}t \overset{(3)}{=} \int_{\nu/\sqrt{2}}^{\infty} \frac{e^{-s^2}}{\sqrt{\pi}}\,\mathrm{d}s \overset{(4)}{=} \tfrac{1}{2}\,\mathrm{erfc}\left(\nu/\sqrt{2}\right). \tag{T3.2}$$

(1) The sky average of the indicator of Q_ν equals its ensemble average at one point (ergodicity), and the average of an indicator is a probability.
(2) The one-point density of a Gaussian field.
(3) Substitute $s = t/(\sqrt{2}\sigma_0)$, so $\mathrm{d}t = \sqrt{2}\sigma_0\,\mathrm{d}s$.
(4) The definition $\mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-s^2}\mathrm{d}s$.

Example: the area above one standard deviation. At $\nu = 1$, (T3.2) gives $\frac{1}{2}\,\mathrm{erfc}(0.7071) = 0.1587$: about one sky pixel in six. The 300 simulated patches of Figure T3.4 give 0.1585. Notice what v_0 does *not* depend on: the spectrum. Only the one-point distribution enters, so v_0 is a test of the shape of the histogram of pixel values and nothing else.

Two numbers from the spectrum. The boundary length and the Euler characteristic do depend on the spectrum, but only through two of its moments: the variance of the field, $\sigma_0^2 = \langle T^2\rangle$, and the variance of its gradient, $\sigma_1^2 = \langle|\nabla T|^2\rangle$ [110, 111]. Both follow from C_ℓ. Write $T = \sum_{\ell m} a_{\ell m} Y_{\ell m}$ with $\langle a_{\ell m} a^*_{\ell' m'}\rangle = C_\ell\,\delta_{\ell\ell'}\delta_{mm'}$ (Eq. (7.62)). Then

$$\sigma_0^2 = \langle T(\hat{\boldsymbol{n}})^2\rangle \overset{(1)}{=} \sum_{\ell m} C_\ell\,|Y_{\ell m}(\hat{\boldsymbol{n}})|^2 \overset{(2)}{=} \sum_\ell \frac{2\ell+1}{4\pi}\,C_\ell, \tag{T3.3}$$

$$\begin{aligned}\sigma_1^2 = \langle|\nabla T|^2\rangle &\overset{(3)}{=} \left\langle \frac{1}{4\pi}\int |\nabla T|^2\,\mathrm{d}\Omega\right\rangle \overset{(4)}{=} -\left\langle \frac{1}{4\pi}\int T\,\nabla^2 T\,\mathrm{d}\Omega\right\rangle \\ &\overset{(5)}{=} \frac{1}{4\pi}\sum_{\ell m}\ell(\ell+1)\langle|a_{\ell m}|^2\rangle = \sum_\ell \frac{2\ell+1}{4\pi}\,\ell(\ell+1)\,C_\ell.\end{aligned} \tag{T3.4}$$

(1) Expand both factors of $T^2 = T\,T^*$ (the field is real) and average; only $\ell = \ell'$, $m = m'$ survives.

(2) Addition theorem (7.63) at coincident points, $\sum_m |Y_{\ell m}|^2 = (2\ell+1)/4\pi$.
(3) On an isotropic sky the expectation at one point equals the expectation of the sky average.
(4) Integrate by parts on the sphere; there is no boundary term because the sphere has no edge.
(5) $\nabla^2 Y_{\ell m} = -\ell(\ell+1) Y_{\ell m}$ and orthonormality of the $Y_{\ell m}$.

A beam of window b_ℓ multiplies C_ℓ by b_ℓ^2 in both sums. The ratio has the dimension of an inverse angle squared, so it defines a correlation angle

$$\theta_c \equiv \sigma_0/\sigma_1, \tag{T3.5}$$

the typical distance over which the field changes by one standard deviation. In flat-sky language, $\ell(\ell+1) \to |\boldsymbol{\ell}|^2$ and the sums become integrals over the plane of wave vectors.

Example: the correlation angle of a 10′ map. For the fiducial *Planck* spectrum smoothed by a 10′ beam, the full-sky sums give $\sigma_0 = 104.2\,\mu$K and $\theta_c = 13.4'$. On a 12.8° patch the multipoles below $\ell \approx 2\pi/12.8° \approx 28$ do not fit, so σ_0 drops to 93.8 μK and θ_c to 12.0′. The lowest multipoles carry a large share of the variance but almost none of the gradient: they are the gentle, sky-wide tilts.

The boundary length: counting crossings. The boundary of Q_ν is the set of points where $T = \nu\sigma_0$ exactly. Its length per unit area is the sky average of $|\nabla T|\,\delta(T-\nu\sigma_0)$: the delta function selects the points on the contour, and the factor $|\nabla T|$ converts "a thin slab of threshold values of width $\mathrm{d}T$" into the length of contour it contains, since a steeper field crosses the slab in a narrower strip. By ergodicity the sky average is an ensemble average at one point, and at one point the value and the gradient of a homogeneous field are independent, because $\langle T\,\partial_i T\rangle = \frac{1}{2}\partial_i\langle T^2\rangle = 0$ and both are Gaussian. So the average factorises. On an isotropic sky the total gradient variance σ_1^2 splits equally between the two directions, so each component of ∇T has variance $s^2 = \sigma_1^2/2$. The length $r = |\nabla T|$ of such a two-dimensional Gaussian vector has the Rayleigh density $(r/s^2)e^{-r^2/2s^2}$, with mean $\int_0^\infty (r^2/s^2)e^{-r^2/2s^2}\mathrm{d}r = s\sqrt{\pi/2}$. Hence

$$\begin{aligned}\mathbb{E}[v_1(\nu)] &\overset{(1)}{=} \frac{1}{4}\,\mathbb{E}\left[|\nabla T|\,\delta(T-\nu\sigma_0)\right] \overset{(2)}{=} \frac{1}{4}\,\frac{e^{-\nu^2/2}}{\sqrt{2\pi}\,\sigma_0}\,\mathbb{E}\,|\nabla T| \\ &\overset{(3)}{=} \frac{1}{4}\,\frac{e^{-\nu^2/2}}{\sqrt{2\pi}\,\sigma_0}\,\frac{\sigma_1}{\sqrt{2}}\sqrt{\frac{\pi}{2}} \overset{(4)}{=} \frac{1}{8\sqrt{2}\,\theta_c}\,e^{-\nu^2/2}.\end{aligned} \tag{T3.6}$$

(1) The definition of v_1 (a quarter of the length per unit area) and the crossing-counting argument above.
(2) Independence of T and ∇T at a point; the average of $\delta(T-\nu\sigma_0)$ is the one-point density of T at $\nu\sigma_0$.
(3) The Rayleigh mean with $s = \sigma_1/\sqrt{2}$.
(4) $\frac{1}{\sqrt{2\pi}}\cdot\frac{1}{\sqrt{2}}\sqrt{\frac{\pi}{2}} = \frac{1}{2\sqrt{2}}$, and $\sigma_1/\sigma_0 = 1/\theta_c$.

The Euler characteristic, stated. The same reasoning applied to the number of hot regions minus holes needs the second derivatives of the field as well, and its derivation is in Chapter 12. The result is

$$\mathbb{E}[v_2(\nu)] = \frac{1}{(2\pi)^{3/2}}\,\frac{1}{2\theta_c^2}\,\nu\,e^{-\nu^2/2}, \tag{T3.7}$$

and (T3.6) with (T3.7) are the two curves referred to below as the Gaussian functionals. Here we read off what each factor means. The factor $e^{-\nu^2/2}$ is the reason a coastline exists at all: a boundary needs the field to be exactly at the threshold, and the probability density of that is proportional to $e^{-\nu^2/2}$. So boundaries and islands are rare at large $|\nu|$. The factor $1/\theta_c$ in

v_1 turns "crossings per unit of threshold" into "length per unit area": a field that changes faster has more coastline. The factor $1/\theta_c^2$ in v_2 is one island per correlation area. Finally the factor ν makes v_2 odd. That follows from the symmetry of a Gaussian under $T \to -T$: the complement of the set $\{T > -\nu\sigma_0\}$ is the set $\{-T \geq \nu\sigma_0\}$, which has the same statistics as Q_ν, and on a closed patch (a torus, or the plane far from edges) the Euler characteristic of a region and of its complement are opposite, because the pieces of one are the holes of the other. So $v_2(-\nu) = -v_2(\nu)$, and $v_2(0) = 0$.

Example: how many hot spots above 2σ? At $\nu = 2$ holes inside hot spots are rare, so χ counts the hot spots. On the full sky the Gaussian kinematic formula adds the sphere's own Euler characteristic times the area fraction to 4π times the density (T3.7) [110]:

$$\mathbb{E}[\chi_{\text{sky}}(2)] \overset{(1)}{=} 4\pi \, \frac{1}{(2\pi)^{3/2}} \, \frac{1}{2\theta_c^2} \, 2e^{-2} + 2 \cdot \tfrac{1}{2} \operatorname{erfc}(\sqrt{2})$$
$$\overset{(2)}{=} 4\pi \times 0.0635 \times \frac{1}{2\,(3.90 \times 10^{-3})^2} \times 0.271 + 2 \times 0.0228 \;\approx\; 7.1 \times 10^3 .$$

(1) The density (T3.7) at $\nu = 2$ times the area 4π of the sphere, plus $\chi(S^2) = 2$ times $\mathbb{E}[v_0(2)]$ from (T3.2).
(2) $(2\pi)^{-3/2} = 0.0635$, $\theta_c = 13.4' = 3.90 \times 10^{-3}$ rad, $2e^{-2} = 0.271$. The second term is negligible.

The unrounded evaluation in `01_excursion_mfs.py` gives 7157. A 10′ map of the whole sky has several thousand hot spots above two standard deviations; a sky with a hundred more or a hundred fewer would be a statement about physics.

T3.1.3 Checking the formulas on simulated skies

To see the formulas at work we simulate 300 flat patches of $12.8° \times 12.8°$, with 1.5′ pixels and the fiducial lensed spectrum smoothed by a 10′ beam, and measure all three functionals at 41 thresholds. The area fraction is a pixel count, the boundary length uses the binned-delta average of $|\nabla T|$, and the Euler characteristic is counted on the pixel grid:

Listing T3.1: Counting pieces minus holes on the pixel grid (`code/chT3/lib_t3.py`)

```
def euler_torus(mask):
    """Euler characteristic of the union of closed pixel squares in ‘mask‘, on the periodic
    grid.

    F = pixels in the set; an edge belongs to the complex if either pixel sharing it is in
    the set;
    a vertex belongs if any of the four pixels around it is in the set.  chi = V - E + F.
    (Closed squares touching at a corner are connected: 8-connectivity for the set.)
    """
    m = np.asarray(mask, bool)
    F = m.sum(axis=(-2, -1))
    r = lambda a, s, ax: np.roll(a, s, axis=ax)
    Eh = (m | r(m, 1, -2)).sum(axis=(-2, -1))      # edges between vertical neighbours
    Ev = (m | r(m, 1, -1)).sum(axis=(-2, -1))      # edges between horizontal neighbours
    V = (m | r(m, 1, -2) | r(m, 1, -1) | r(r(m, 1, -2), 1, -1)).sum(axis=(-2, -1))
    return V - (Eh + Ev) + F
```

Each pixel above the threshold is a closed square. The function counts the squares (F), the pixel edges that belong to at least one of them (E), and the pixel corners that belong to at least one of them (V); the alternating sum $V - E + F$ is the Euler characteristic of their union. The rolls make the patch periodic, so the patch is a torus and has no edge to correct for. On one map the

count agrees exactly with `skimage.measure.euler_number` (89 and 89), the library routine the research literature uses.

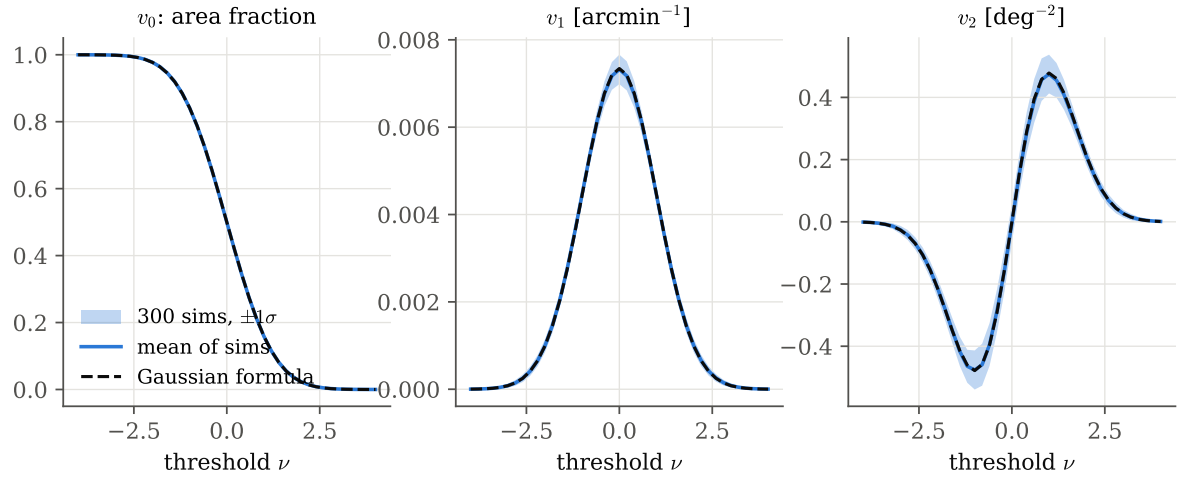


Figure T3.4: The three Minkowski functionals of simulated Gaussian CMB patches (10′ beam) against the threshold. The band is the spread of single patches, the solid line their mean, and the dashed line the Gaussian formulas (T3.2) and (T3.6), (T3.7) with σ_0 and σ_1 computed from the spectrum. The Euler characteristic changes sign at $\nu = 0$: islands above, lakes below.

Figure T3.4 shows the result. The mean of the simulations lies on the Gaussian curves; the largest departure of the mean Euler characteristic from the formula is 1.4% of its peak, against a Monte Carlo error of 0.8% on the mean. The bands are wide: a single 164-square-degree patch has a peak Euler characteristic near 78, and the patch-to-patch scatter is many times larger than the error on the mean. That scatter is the sampling distribution of the statistic, and it is what Chapter 12 estimates by Monte Carlo to turn a measured curve into a test.

Watch out: Thresholds in units of which σ_0?

The formulas use the true σ_0 of the ensemble. A map only gives its own sample standard deviation, which differs from patch to patch (a 164 deg^2 patch misses all the variance at $\ell \lesssim 28$). Normalising each map by its own standard deviation removes most of the scatter in the *amplitude* of the curves and leaves their shape, which is why *Planck* does it (Section T3.5). Whichever choice is made, the simulations must make the same one.

Real maps are not Gaussian random fields drawn on a computer, and the functionals are just as easy to compute on any surface. [110] show them for the Earth's topography binned onto a sky map: the curves change sharply at the depth of the ocean floor and at sea level, the Euler characteristic dips where the mid-ocean ridges emerge as long connected ranges, and nothing about them looks like (T3.6), (T3.7). A strongly non-Gaussian surface announces itself immediately. The cosmological question is the opposite one: how small a departure from the Gaussian curves can be detected, and what would cause it.

T3.2 What a skewed sky and a lensed sky change

The simplest inflationary models predict nearly Gaussian primordial fluctuations. In models with more than one field, the primordial potential can also contain a small term quadratic in a Gaussian field. The *Planck* analysis describes this contribution with the **local model** [112]

$$\Phi(\boldsymbol{x}) = \phi(\boldsymbol{x}) + f_{\mathrm{NL}}\left[\phi^2(\boldsymbol{x}) - \langle\phi^2\rangle\right], \tag{T3.8}$$

where ϕ is a Gaussian field, Φ is the primordial potential (Bardeen's potential, $\Phi = \frac{3}{5}\zeta$ on super-horizon scales, with ζ the comoving curvature perturbation; see the cosmology notes), and f_{NL} is a number that measures how much of the square is mixed in. Subtracting $\langle\phi^2\rangle$ keeps the mean of Φ zero. Because $\phi \sim 10^{-5}$, even $f_{\mathrm{NL}} \sim 10$ makes the quadratic term a part in 10^4 of the linear one. The question is what such a term does to the three curves of Figure T3.4. A second question runs alongside it. Every CMB photon has been deflected by a couple of arcminutes on its way to us; does that change the curves?

The question. If the sky is slightly skewed, which of the three Minkowski functionals move, by how much, and how much of the move tells us about the physics that produced the skewness rather than about the histogram of pixel values? And what does lensing do to the same three numbers?

T3.2.1 The first correction to the Gaussian curves

A mildly non-Gaussian field is described, to first order, by its three-point averages. For the Minkowski functionals in two dimensions only three such averages matter [113] (his eqs. 2.61–2.63 with $d = 2$). Measured in units of σ_0, and with $\theta_c = \sigma_0/\sigma_1$ the correlation angle of (T3.5), they are

$$S^{(0)}\sigma_0 = \frac{\langle T^3\rangle}{\sigma_0^3}, \qquad S^{(1)}\sigma_0 = -\frac{3}{4}\,\frac{\langle T^2\,\nabla^2 T\rangle}{\sigma_0\,\sigma_1^2}, \qquad S^{(2)}\sigma_0 = -3\,\frac{\langle |\nabla T|^2\,\nabla^2 T\rangle\,\sigma_0}{\sigma_1^4}. \tag{T3.9}$$

The first is the ordinary skewness of the histogram of pixel values: positive when the hot tail is longer than the cold one. The other two, the **skewness parameters** of the derivatives, say whether hot regions differ from cold ones in how steep and how curved they are. All three vanish for a Gaussian field. To first order in them, Matsubara's result (his eq. 3.48, $d = 2$) reads

$$\begin{aligned}
\mathbb{E}[v_0(\nu)] &= \tfrac{1}{2}\,\mathrm{erfc}\big(\nu/\sqrt{2}\big) + \frac{e^{-\nu^2/2}}{\sqrt{2\pi}}\,\frac{S^{(0)}\sigma_0}{6}\,H_2(\nu),\\
\mathbb{E}[v_1(\nu)] &= \frac{1}{8\sqrt{2}\,\theta_c}\,e^{-\nu^2/2}\Big\{1 + \Big[\frac{S^{(0)}}{6}H_3(\nu) + \frac{S^{(1)}}{3}H_1(\nu)\Big]\sigma_0\Big\},\\
\mathbb{E}[v_2(\nu)] &= \frac{1}{(2\pi)^{3/2}}\,\frac{1}{2\theta_c^2}\,e^{-\nu^2/2}\Big\{H_1(\nu) + \Big[\frac{S^{(0)}}{6}H_4(\nu) + \frac{2S^{(1)}}{3}H_2(\nu) + \frac{S^{(2)}}{3}\Big]\sigma_0\Big\},
\end{aligned} \tag{T3.10}$$

with the Hermite polynomials $H_n(\nu) = e^{\nu^2/2}(-\mathrm{d}/\mathrm{d}\nu)^n e^{-\nu^2/2}$, that is $H_1 = \nu$, $H_2 = \nu^2 - 1$, $H_3 = \nu^3 - 3\nu$, $H_4 = \nu^4 - 6\nu^2 + 3$. Setting the $S^{(a)}$ to zero gives back (T3.2) and (T3.6), (T3.7). The derivation is the Gaussian one of Chapter 12 with the joint density of the field and its derivatives expanded one order beyond the Gaussian (an Edgeworth expansion).

T3.2.2 A pointwise distortion only relabels the threshold

The claim in words. Suppose the observed map is a fixed increasing function of a Gaussian map, $T = g(u)$, applied pixel by pixel. Then the region where T exceeds t is exactly the region where u exceeds $g^{-1}(t)$. The two maps have the same excursion sets; only the labels on the thresholds differ. So every functional of T at threshold t is the Gaussian functional of u at threshold $g^{-1}(t)$, and the whole non-Gaussian correction is a sideways shift of the curves.

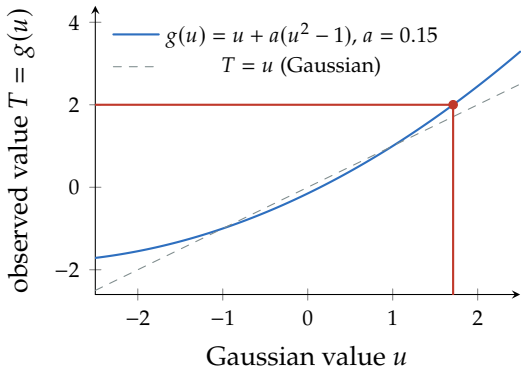


Figure T3.5: A pointwise, increasing distortion of the values. The threshold $T = 2$ on the observed map (horizontal red line) picks out exactly the pixels with $u > g^{-1}(2) \approx 1.71$ on the Gaussian map (vertical red line). The excursion set is the same set of pixels; only the number attached to the threshold has changed.

Example: the relabelling reproduces Matsubara's correction. Take $T = u + a(u^2 - 1)$ with u Gaussian of unit variance and a small (Figure T3.5). Its skewness is

$$\langle T^3 \rangle \overset{(1)}{=} \langle u^3 \rangle + 3a \, \langle u^2(u^2 - 1) \rangle + O(a^2) \overset{(2)}{=} 0 + 3a \, (3 - 1) = 6a,$$

(1) Expand the cube and keep terms of first order in a.
(2) Odd Gaussian moments vanish, $\langle u^4 \rangle = 3$ and $\langle u^2 \rangle = 1$.

so $S^{(0)}\sigma_0 = 6a$, and a calculation of the same kind (or the simulation below) gives $S^{(1)}\sigma_0 = S^{(2)}\sigma_0 = 6a$ as well. Now do it the other way. The threshold ν on T corresponds to $u_\nu = g^{-1}(\nu) = \nu - a(\nu^2 - 1) + O(a^2)$, so

$$\begin{aligned} v_2^T(\nu) = v_2^G(u_\nu) &\overset{(1)}{=} v_2^G(\nu) - a(\nu^2 - 1)\,\frac{\mathrm{d}v_2^G}{\mathrm{d}\nu} \overset{(2)}{=} v_2^G(\nu) + \frac{1}{(2\pi)^{3/2}}\frac{1}{2\theta_c^2}\, e^{-\nu^2/2}\, a\,(\nu^2 - 1)^2 \\ &\overset{(3)}{=} v_2^G(\nu) + \frac{1}{(2\pi)^{3/2}}\frac{1}{2\theta_c^2}\, e^{-\nu^2/2}\left[\frac{6a}{6}H_4 + \frac{2\cdot 6a}{3}H_2 + \frac{6a}{3}\right]. \end{aligned}$$

(1) Taylor expansion to first order in the shift $-a(\nu^2 - 1)$.
(2) $\frac{\mathrm{d}}{\mathrm{d}\nu}(\nu e^{-\nu^2/2}) = (1 - \nu^2)e^{-\nu^2/2}$, so the product of the two minus signs is $+a(\nu^2 - 1)^2$.
(3) $H_4 + 4H_2 + 2 = \nu^4 - 6\nu^2 + 3 + 4\nu^2 - 4 + 2 = (\nu^2 - 1)^2$.

The last line is (T3.10) with all three skewness parameters equal to $6a$. For a pointwise distortion the perturbative formula is nothing but the relabelling.

Removing the relabelling. The relabelling can be undone without knowing g. Instead of the threshold ν, label each excursion set by the area it covers: define the **area threshold** ν_A by $\frac{1}{2}\,\mathrm{erfc}(\nu_A/\sqrt{2})$ = the measured area fraction. A pointwise increasing distortion keeps the order of the pixel values, so the set covering a given fraction of the sky is the same set before and after it. At fixed ν_A the three functionals of $T = g(u)$ are therefore exactly those of u. In Matsubara's expansion this shows up as follows: rewritten in terms of ν_A, the corrections contain only the differences $S^{(1)} - S^{(0)}$ and $S^{(2)} - S^{(0)}$ [113] (his eq. 3.54), which vanish when the three are equal. What is left at fixed area fraction is information about how the skewness is shared between the field and its slopes and curvatures, and that information can only come from physics that couples neighbouring points.

T3.2.3 Testing it on two skewed skies

`02_nongaussian_mfs.py` builds two kinds of skewed patch, both with the smoothed CMB spectrum shape of Section T3.1.3, and with amplitudes exaggerated so that the effect stands out in 300 patches. The *local* sky is $T = u + a(u^2 - 1)$ with $a = 0.05$; the code measures $S^{(0)}\sigma_0, S^{(1)}\sigma_0, S^{(2)}\sigma_0 = 0.298, 0.300, 0.299$, against $6a = 0.30$. The *non-local* sky applies the same quadratic term, with coefficient 0.1, to a scale-invariant field ϕ (equal power per logarithmic interval of ℓ, like the primordial potential) and then filters the result with the transfer function that turns ϕ's spectrum into the CMB one. The filter mixes neighbouring points, as the physics between the end of inflation and last scattering does, and the measured parameters are no longer equal: 0.493, 0.443, 0.470.

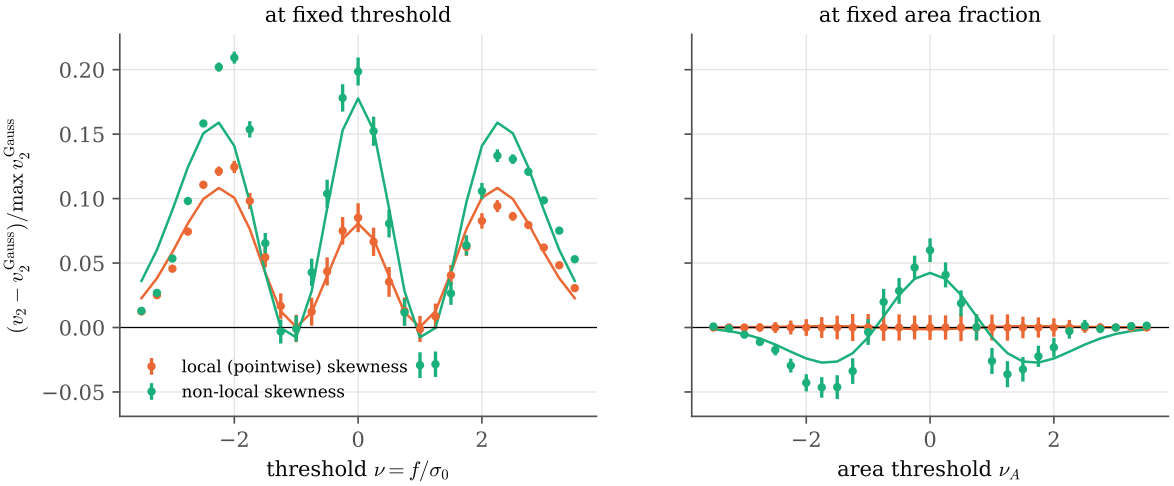


Figure T3.6: Change of the Euler characteristic density of two skewed skies from the Gaussian one, in units of the peak of the Gaussian curve. Points: 300 simulated patches; lines: Matsubara's first-order formula with the measured skewness parameters. Left, at fixed threshold ν: the local sky moves by up to about a tenth of the peak, the non-local one by up to about a fifth. Right, at fixed area fraction: the pointwise (local) distortion disappears exactly, the non-local one does not.

Figure T3.6 shows the result. At fixed threshold the local sky shifts by up to 12% of the peak, with the shape $(\nu^2 - 1)^2$ predicted above. At fixed area fraction the shift is 0.0%: the same pixels, the same sets, the same counts, realisation by realisation. The non-local sky keeps a shift of up to 6.0% of the peak at fixed area fraction, against a Monte Carlo error of 1.0% (the non-local and Gaussian patches are independent draws, so both errors enter), and the first-order formula follows it roughly. Near $\nu_A = -2$ the points and the formula differ by -1.9% of the peak, 2.8 times that error, too few for 300 patches to pin down. A difference there would be no surprise: $S^{(0)}\sigma_0 \approx 0.5$ is not small, and the formula keeps only the first order.

Example: what the real f_{NL} does on large scales. On the largest angular scales the temperature is set by the potential at last scattering alone, $\Delta T/T = -\Phi/3$ (the Sachs–Wolfe limit; see the cosmology notes). Insert the local model (T3.8) and write the Gaussian part of the temperature as $t = -\phi/3$, so that $\phi = -3t$:

$$\frac{\Delta T}{T} = -\frac{\Phi}{3} \overset{(1)}{=} -\frac{\phi}{3} - \frac{f_{\text{NL}}}{3}\left(\phi^2 - \langle\phi^2\rangle\right) \overset{(2)}{=} t - 3f_{\text{NL}}\left(t^2 - \langle t^2\rangle\right) \overset{(3)}{=} \sigma_t\left[u - 3f_{\text{NL}}\sigma_t\,(u^2 - 1)\right].$$

(1) Substitute (T3.8).

(2) $\phi = -3t$, so $\phi^2 = 9t^2$ and $-\frac{f_{\text{NL}}}{3}\cdot 9 = -3f_{\text{NL}}$.

(3) Write $t = \sigma_t u$ with u of unit variance.

This is the local toy with $a = -3 f_{\rm NL}\sigma_t$. Two conclusions follow. First, the sign: a positive $f_{\rm NL}$ gives a negative skewness $6a$, so the cold spots of the Sachs–Wolfe sky are the more extreme ones. Second, the size: the large-angle temperature has $\sigma_t \approx 2\times10^{-5}$, so $f_{\rm NL} = 5$ gives $|6a| \approx 18\times5\times2\times10^{-5} \approx 2\times10^{-3}$, a shift of the Euler characteristic of order 0.1–0.2% of its peak, and at fixed area fraction none at all. On scales where the temperature is a pointwise function of the potential, the Minkowski functionals at fixed area fraction are blind to local $f_{\rm NL}$; the signal that survives comes from the acoustic scales, where the transfer is non-local, and it is small there too. This is why *Planck* measures $f_{\rm NL}$ with the bispectrum, the optimal statistic for a known three-point shape, and uses the Minkowski functionals as a model-independent check (Section T3.5).

T3.2.4 Lensing leaves the one-point distribution alone

Lensing moves the photons arriving in direction $\hat{\boldsymbol{n}}$ so that we see the last-scattering surface at a neighbouring point, $\tilde{T}(\hat{\boldsymbol{n}}) = T(\hat{\boldsymbol{n}} + \boldsymbol{\alpha})$, with a deflection $\boldsymbol{\alpha}$ of a couple of arcminutes that is coherent over a couple of degrees [47] (their §1.2). It does not change the photons' energies, so it conserves surface brightness: every temperature we see in the lensed sky is a temperature that existed on the unlensed one, seen in a slightly different direction. For a statistically homogeneous sky the lensed one-point distribution is therefore the unlensed one, and in particular the variance is unchanged, $\tilde{\sigma}_0 = \sigma_0$ (their eq. 4.30). By (T3.2) the area functional $v_0(\nu)$ is untouched. The gradient variance is not: lensing shears and magnifies the image, moving power to smaller scales, so $\tilde{\sigma}_1 > \sigma_1$. Plugged into (T3.6), (T3.7), that would predict more coastline and more islands. Whether there really are more islands is a question about topology, and Section T3.4 answers it: a smooth remapping cannot change how many islands there are.

Example: lensing that looks like $f_{\rm NL}$. The deflection is produced by the same large-scale structure that, through the integrated Sachs–Wolfe effect, adds a little to the large-angle temperature. The two are correlated, and the correlation produces a three-point signal in the observed temperature whose shape resembles the local one. *Planck*'s temperature-only estimate of local $f_{\rm NL}$ is 6.7 ± 5.6 before subtracting the expected lensing contribution and -0.5 ± 5.6 after it [112] (their Table 5, KSW estimator, `SMICA`). A lensed sky is, for three-point statistics, a slightly non-Gaussian sky with a known shape, and an analysis that forgets it reports a spurious $f_{\rm NL}$ of about one standard deviation.

T3.3 Polarization patterns and their singularities

The CMB is linearly polarized at the level of a few per cent of its anisotropy. At each point of the sky the polarization is described by two Stokes parameters, Q and U, measured with respect to a pair of axes drawn on the sky at that point. Draw a short stick at each point, along the direction of the polarization. The stick has a direction but no arrowhead: turning it by 180° gives the same stick, because a polarization direction is the orientation of an oscillating electric field, not the direction of a vector. In the language of [1], the sky is covered with headless rods, like the molecules of a thin film of nematic liquid crystal, and such patterns have point defects. Wherever Q and U both vanish, the stick has no direction at all. Walk around such a point and the sticks turn steadily, by half a turn in one sense or the other. These points are the **polarization singularities**.

The question. Where does the polarization vanish? What does the pattern of sticks look like around such a point, and why only half a turn? How many such points does a Gaussian polarization map with given C_ℓ^{EE} and C_ℓ^{BB} have per square degree, and does their arrangement carry information that the spectra do not?

T3.3.1 Headless sticks and a half-integer index

From Q and U to a stick. The polarization at a point is a symmetric, trace-free 2×2 matrix [1] (their eq. 1),

$$\mathsf{P} = \begin{pmatrix} Q & U \\ U & -Q \end{pmatrix}, \qquad \frac{\mathsf{P}}{2P} = \hat{\boldsymbol{m}}\hat{\boldsymbol{m}}^{\mathsf{T}} - \tfrac{1}{2}\mathbb{1}, \qquad P = \sqrt{Q^2+U^2}, \tag{T3.11}$$

with $\hat{\boldsymbol{m}} = (\cos\alpha, \sin\alpha)$ the unit vector along the stick. Multiplying out $\hat{\boldsymbol{m}}\hat{\boldsymbol{m}}^{\mathsf{T}} - \frac{1}{2}\mathbb{1}$ gives the matrix with entries $\frac{1}{2}\cos 2\alpha$ and $\frac{1}{2}\sin 2\alpha$, so $Q = P\cos 2\alpha$ and $U = P\sin 2\alpha$ [2] (their eq. 2). The vector $\hat{\boldsymbol{m}}$ enters twice, so $\hat{\boldsymbol{m}}$ and $-\hat{\boldsymbol{m}}$ give the same matrix: that is the missing arrowhead. In complex notation,

$$Q + iU = P\, e^{2i\alpha}, \qquad \alpha = \tfrac{1}{2}\arg(Q+iU), \tag{T3.12}$$

and rotating the reference axes by an angle ψ changes α by $-\psi$ and the pair (Q,U) by -2ψ. That factor two is what is meant by calling the polarization a spin-2 quantity.

Working definition T3.2 (polarization singularity and its index)

A *polarization singularity* is a point where $Q = U = 0$ [the polarized intensity P vanishes, so the stick direction α is undefined]. Its *index* is the number of turns the stick makes along a small loop Γ around it,

$$W_\Gamma = \frac{1}{2\pi}\oint_\Gamma \mathrm{d}\alpha,$$

with $\mathrm{d}\alpha$ the change of the stick angle along the loop [1] (their eq. 5). *Operational:* on a pixel grid, add the changes of $\arg(Q+iU)$ around each 2×2 block of pixels, each change taken the short way round the circle; the total is $2\pi n$, and the index is $n/2$.

Why the index is a half-integer. Go once around the loop. The complex number $Q+iU$ comes back to its starting value, so its phase 2α has changed by a whole number of turns, $2\pi n$. The stick angle α has therefore changed by πn, and $W_\Gamma = n/2$. A stick that has turned by π looks exactly as it did at the start, so half-integers are allowed, which they would not be for arrows. *Why $\pm\frac{1}{2}$ is the usual value.* Near a generic zero both Q and U vanish linearly. The simplest such patterns are $Q+iU = x+iy$ and $Q+iU = x-iy$, with (x,y) measured from the zero. In polar coordinates $x+iy = re^{i\theta}$, the first gives $\alpha = \theta/2$: the stick turns half a turn in the same sense as the loop, index $+\frac{1}{2}$. The second gives $\alpha = -\theta/2$: half a turn against it, index $-\frac{1}{2}$. These two patterns, drawn in Figure T3.7, are the fundamental singularities; a general linear zero, $Q+iU = c_1(x+iy)+c_2(x-iy)$, has the index of whichever coefficient is larger in modulus. The index ± 1 patterns often drawn in the CMB literature (radial and circular patterns, and saddles) are pairs of fundamental singularities sitting on top of each other [1, 2], which a random field produces only by accident.

The total index. On a closed surface the indices of all singularities of a line field add up to the Euler characteristic of the surface (the Poincaré–Hopf theorem; see the topology notes). The sky is a sphere, so the indices of all the singularities on the full sky add up to $+2$ [1]: it is impossible to comb the polarization smoothly over the whole sky. With thousands of singularities of each sign this excess of four half-charges is negligible [2]. On a periodic flat patch, a torus, the total is exactly 0.

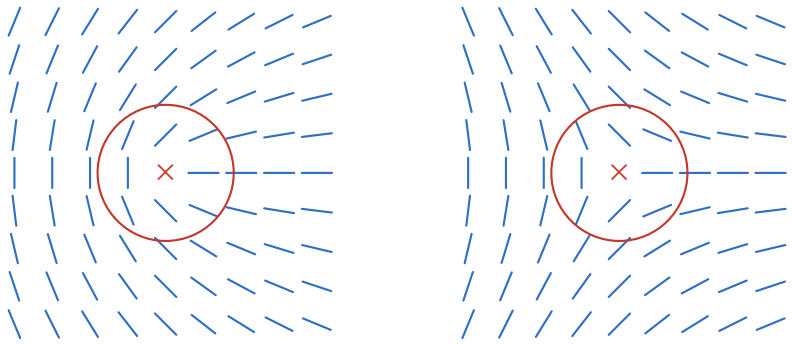


Figure T3.7: The two fundamental polarization singularities. Each stick is drawn along $\alpha = \pm\frac{1}{2}\arg(x + iy)$ at its grid point, and the cross marks the zero. As we follow the red loop once around the zero, the orientation of the sticks rotates through half a turn, in the direction of traversal on the left and in the opposite direction on the right.

T3.3.2 How many singularities a Gaussian sky has

The counting idea. Look at a tiny square of sky of area $\mathrm{d}A$. The map $(x, y) \mapsto (Q, U)$ sends it to a tiny parallelogram in the (Q, U) plane of area $|\det J|\,\mathrm{d}A$, where J is the Jacobian matrix with rows ∇Q and ∇U. The square contains a zero exactly when that parallelogram contains the origin of the (Q, U) plane. So the expected number of zeros per unit area is the probability density of (Q, U) at the origin, times the average area factor:

$$n \overset{(1)}{=} p_{Q,U}(0,0)\ \mathbb{E}\,|\det J| \overset{(2)}{=} \frac{1}{2\pi\sigma_0^2}\ \mathbb{E}\,|\det J| \overset{(3)}{=} \frac{1}{2\pi\sigma_0^2}\cdot\frac{\sigma_1^2}{2} = \frac{\sigma_1^2}{4\pi\sigma_0^2}, \tag{T3.13}$$

(1) For a statistically homogeneous Gaussian field the value and the gradient at the same point are uncorrelated, hence independent: $\langle Q\,\partial_x Q\rangle = \frac{1}{2}\partial_x\langle Q^2\rangle = 0$, and similarly for the other pairs.

(2) For isotropic polarization Q and U are uncorrelated at a point with equal variances σ_0^2, so their joint density at the origin is $1/(2\pi\sigma_0^2)$.

(3) Shown just below, with $\sigma_1^2 = \langle|\nabla Q|^2\rangle = \langle|\nabla U|^2\rangle$.

The average area factor. Write $\det J = |\nabla Q|\,|\nabla U|\,\sin\theta$, with θ the angle between the two gradients. For isotropic polarization the four derivatives are independent Gaussians of equal variance $s^2 = \sigma_1^2/2$ (a direction-averaged statement checked by `03_polarization.py`). Each gradient is then a two-dimensional Gaussian vector, whose length has the Rayleigh mean $s\sqrt{\pi/2}$ (the integral done for (T3.6)), and whose direction is uniform and independent of the other. So $\mathbb{E}\,|\det J| = (s\sqrt{\pi/2})^2 \cdot \mathbb{E}\,|\sin\theta| = s^2\frac{\pi}{2}\cdot\frac{2}{\pi} = s^2 = \sigma_1^2/2$.

Both variances come from the spectra exactly as in (T3.3) and (T3.4), now with $C_\ell^P = C_\ell^{EE} + C_\ell^{BB}$ in place of C_ℓ (Q and U each carry half of the total E plus B power, and the half cancels in the ratio). The ratio is an average of $\ell(\ell+1)$ weighted by the polarization power, so the expected number of singularities on the full sky is

$$N_{\mathrm{sky}} = 4\pi n = \frac{\sum_\ell (2\ell+1)\,\ell(\ell+1)\,C_\ell^P}{\sum_\ell (2\ell+1)\,C_\ell^P} \equiv \big\langle \ell(\ell+1)\big\rangle_P. \tag{T3.14}$$

In words: one singularity per patch of area $4\pi\theta_P^2$, where $\theta_P = \sigma_0/\sigma_1$ is the correlation angle of the polarization. (On the sphere the spin-2 nature of $Q + iU$ changes $\ell(\ell+1)$ to $\ell(\ell+1) - 4$ at the lowest multipoles, which does not matter for the counts below.)

Example: a 10′ polarization map. For the fiducial $C_\ell^{EE} + C_\ell^{BB}$ smoothed by a 10′ beam, (T3.13) predicts 1510 singularities in a $12.8^\circ \times 12.8^\circ$ patch, about 9.2 per square degree. The 200 simulated patches of Figure T3.9 have 1501 on average, off the prediction by -0.6%, which is 3.0 standard errors of the mean but tiny next to the spread from patch to patch. That spread is 39 ± 2, close to the Poisson value $\sqrt{N} = 39$. The net index of every patch is exactly zero, as Poincaré–Hopf requires on a torus.

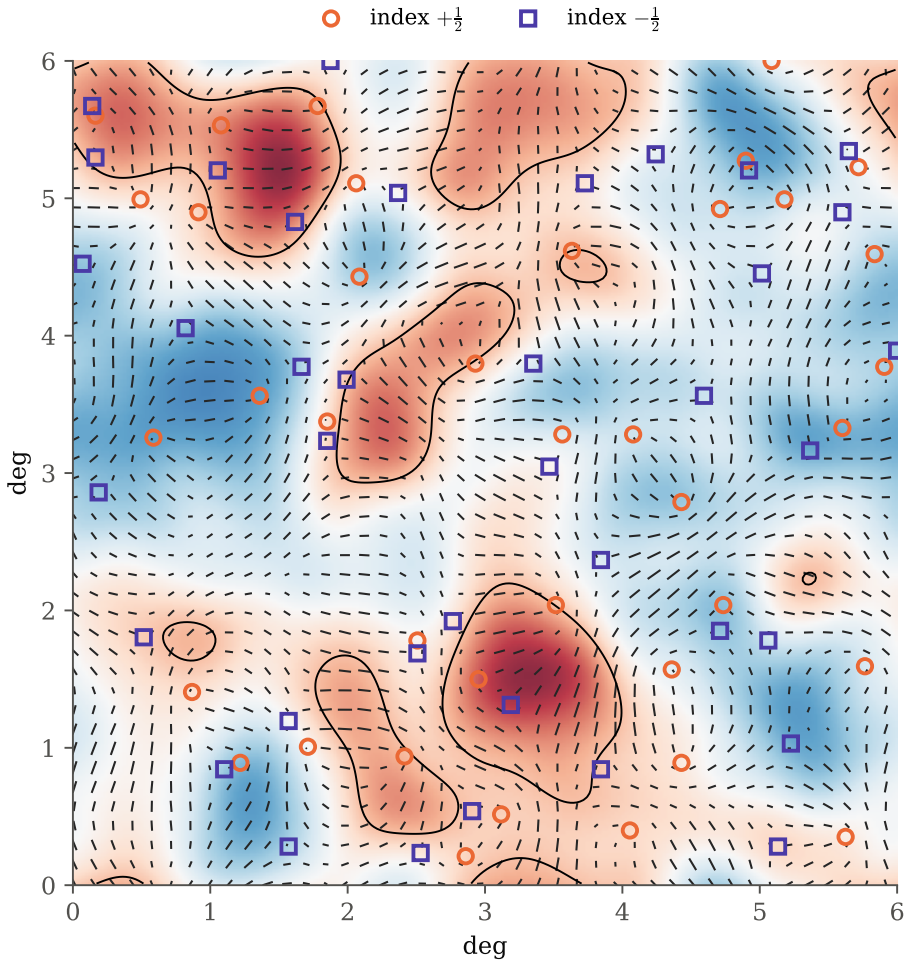


Figure T3.8: A simulated $6^\circ \times 6^\circ$ patch of the CMB with a 30′ beam: temperature in colour (units of its standard deviation), the boundary of the excursion set $T > +1\sigma_0$ as black contours, the polarization as headless sticks whose length grows with P, and the polarization singularities, $+\frac{1}{2}$ as circles and $-\frac{1}{2}$ as squares. The patch holds 76 singularities, 38 of each index. Near every marker the sticks are short, and they turn by half a turn around it.

In Figure T3.8 the colour and the black contours are the temperature and its excursion set of Section T3.1; the sticks and markers are the polarization. The singularities sit where the sticks are shortest, they come in comparable numbers of each index, and they are scattered fairly evenly, with a tendency for a circle and a square to sit close together. The finder is a few lines:

Listing T3.2: Finding polarization singularities by the winding of $Q + iU$ (`code/chT3/lib_t3.py`)

```
def singularities(Q, U):
    """Zeros of P = Q + iU on the periodic grid, one per plaquette, with their index.

    Around each plaquette (i,j)->(i+1,j)->(i+1,j+1)->(i,j+1)->(i,j) add the changes of
    arg(Q + iU), each wrapped into (-pi, pi]. The total is 2 pi n with n an integer; the
    polarization angle alpha = arg/2 turns by n pi, so the index is n/2.
```

```
    Returns (i, j, index) arrays; positions are plaquette centres (i + 1/2, j + 1/2).
    """
    a = np.angle(Q + 1j * U)
    def d(x, y):
        return np.angle(np.exp(1j * (y - x)))
    a10 = np.roll(a, -1, 0)
    a11 = np.roll(a10, -1, 1)
    a01 = np.roll(a, -1, 1)
    w = d(a, a10) + d(a10, a11) + d(a11, a01) + d(a01, a)
    n = np.rint(w / (2 * np.pi)).astype(int)
    i, j = np.nonzero(n)
    return i, j, n[i, j] / 2
```

`np.angle` gives $\arg(Q + iU)$ at every pixel. The helper `d` takes the change of phase between two neighbours the short way round, wrapping it into $(-\pi, \pi]$, which is the pixelized version of following the stick continuously [2] (their Appendix A, there written for α with jumps wrapped into $\pm\pi/2$). Summing the four changes around a 2×2 block gives $2\pi n$, and the block holds a singularity of index $n/2$ when $n \neq 0$.

T3.3.3 Reproducing the Huterer–Vachaspati counts

Their question. When Huterer and Vachaspati wrote, polarization maps were about to arrive, and they asked what the singularities of a ΛCDM polarization map should look like: how many, how they cluster, and how robust those statistics are to changes in the model [2]. *Their setup.* They computed spectra for a flat ΛCDM model ($\Omega_m = 0.3$, $\Omega_m h^2 = 0.127$, $\Omega_b h^2 = 0.021$, $n_s = 1$, no tensors), drew Gaussian E and B coefficients on the full sky with power up to a cutoff $\ell_{\max}$, made HEALPix maps at $N_{\rm side} = 256$, and counted singularities with the winding rule. They report about 6000 singularities for $\ell_{\max} = 100$ and about 150 000 for $\ell_{\max} = 500$, growing roughly as $\ell_{\max}^2$. *Their key equation, here.* The count is (T3.14): for a spectrum cut sharply at $\ell_{\max}$ and rising with ℓ, as C_ℓ^{EE} does below $\ell \approx 1000$, the average of $\ell(\ell+1)$ is a fixed fraction of $\ell_{\max}^2$, which is the scaling they found. *What we compute.* The fiducial *Planck* spectra of this book, cut at the same two values, in (T3.14), and a direct count on 20° flat patches cut at $\ell_{\max} = 500$, scaled to the full sky.

$\ell_{\max}$	H&V 2005 (their maps)	(T3.14), our spectra	our flat patches, scaled
100	≈ 6000	5310	–
500	$\approx 150\,000$	133315	133453

Why they differ. The formula and our direct count agree to a fraction of a per cent. The published counts are about 12% higher, and their spectrum accounts for the direction: with $n_s = 1$ instead of 0.965 their model has relatively more power at high ℓ, which raises the power-weighted average of $\ell(\ell+1)$; their numbers are also quoted as round figures, and a pixel of 14′ at $N_{\rm side} = 256$ resolves the $\ell_{\max} = 500$ pattern only coarsely. *The lesson.* The *number* of singularities is one moment of the polarization spectrum; it carries nothing that C_ℓ^{EE} and C_ℓ^{BB} do not. Whatever new information the singularities hold must be in their *arrangement*.

What the arrangement shows. Huterer and Vachaspati measured three things. The net index inside a closed loop of length L fluctuates with a root-mean-square value growing as L^β with $\beta \approx 0.5$ (between 0.54 for $\ell_{\max} = 100$ and 0.45 for 500): the index enclosed is a sum of roughly L/ξ independent contributions from the stretches of the loop, one per correlation length ξ, and a sum of that many random terms grows as their square root [1]. Singularities of the same index repel and those of opposite index attract: for $\ell_{\max} = 100$ the mean distance to the nearest neighbour of the same index is 3.24° and of the opposite index 2.69°. Beyond about 10° the

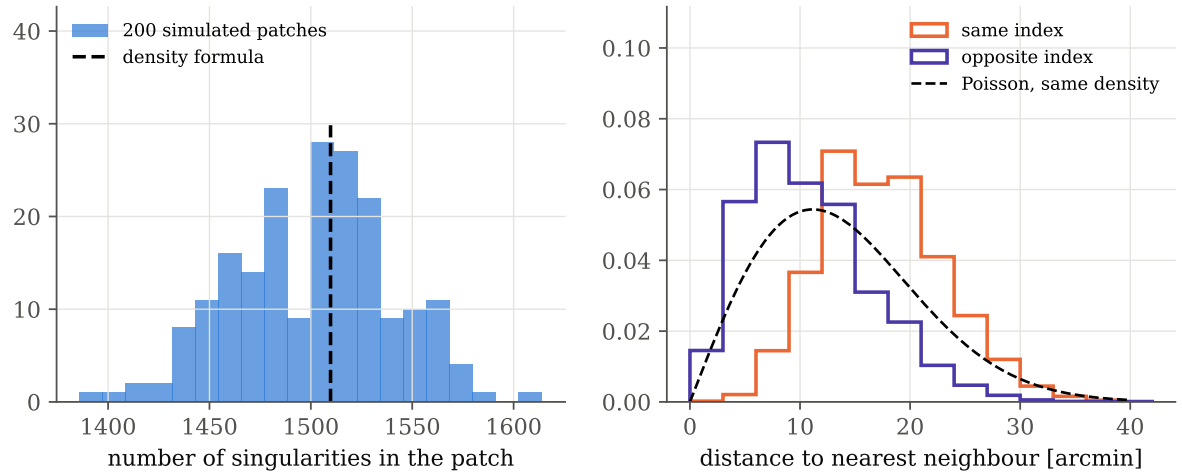


Figure T3.9: Left: number of polarization singularities in 200 simulated 10′-beam patches; the dashed line is the density formula (T3.13). Right: distance from each singularity to its nearest neighbour of the same index and of the opposite index, against the distribution for points scattered at random with the density of one index. Opposite indices sit closer than random points would; equal indices keep apart.

singularities are uncorrelated. The right panel of Figure T3.9 shows the same attraction and repulsion at 10′ resolution: mean nearest-neighbour distances of 17.3′ for the same index and 10.6′ for the opposite one, against 14.0′ for random points. Physically, a $+\frac{1}{2}$ and a $-\frac{1}{2}$ are the two ends of a short region where the field passes close to zero; they are born together.

Huterer and Vachaspati also found these statistics remarkably robust. Tensor modes, strongly tilted spectra and even strongly non-Gaussian, exponentially distributed E and B coefficients left the scaling exponent, the correlation function and the nearest-neighbour distances unchanged; only skies with grossly broken isotropy moved them. That robustness cuts both ways: the singularity distribution is a poor probe of primordial non-Gaussianity, but any departure from it in real data would point to something that happened to the photons after last scattering, and lensing is the obvious candidate [2]. Section T3.4 shows what lensing actually does.

T3.4 Lensing moves the pattern but keeps its topology

Between last scattering and us, every CMB photon passes a few dozen potential wells and hills of the large-scale structure, each of which bends its path by about 10^{-4} radians. The bends add up like a random walk to a deflection of about 2.7′, coherent over a couple of degrees, because the potentials that matter are a few hundred megaparsecs across [47] (their §1.2). The image of the last-scattering surface is therefore not blurred but remapped: each patch is shifted, slightly stretched and slightly sheared. A degree-sized hot spot appears a few per cent larger or smaller than it was. For the shapes of the previous sections this is a nuisance. The question is which of them it can change.

The question. If the observed sky is the unlensed sky seen through a smooth, slightly distorted window, which morphological numbers come through unchanged, and what could change them anyway?

T3.4.1 The remapping, for temperature and for polarization

For the temperature the remapping is the one already used in Section T3.2.4: the lensed field in direction $\hat{\boldsymbol{n}}$ is the unlensed field at the displaced point, $\tilde{T}(\hat{\boldsymbol{n}}) = T(\hat{\boldsymbol{n}} + \boldsymbol{\alpha})$, where the deflection is the gradient of a lensing potential, $\boldsymbol{\alpha} = \nabla\psi$ [3] (their eq. 1). On the sphere "the displaced point" means: walk a distance $|\boldsymbol{\alpha}|$ along the great circle that leaves $\hat{\boldsymbol{n}}$ in the direction of $\boldsymbol{\alpha}$.

For polarization there is one more step, and it is the point of [3]. The Stokes parameters are components of the matrix (T3.11) in a pair of axes at each point, usually the local north and east directions. Copying the numbers Q and U from the displaced point to $\hat{\boldsymbol{n}}$ would compare components in two different sets of axes, and on a curved sky those axes are rotated relative to each other by an angle that depends on the coordinates. The physical statement is that the polarization *matrix* at the displaced point is carried back to $\hat{\boldsymbol{n}}$ by parallel transport along the great circle, and only then read in the local axes (their eq. 4). The naive copy differs from the transported one by a factor $e^{-2i\chi}$ in $Q + iU$, with χ a rotation angle of the order of the deflection times a coordinate factor that grows towards the poles (their eqs. 7–8); in the flat-sky limit it vanishes, and in any case the residual rotation of the polarization basis is of order an arcminute and negligible [47] (their §3). The upshot for us is simple: the lensed polarization at $\hat{\boldsymbol{n}}$ is the unlensed polarization at $\hat{\boldsymbol{n}} + \boldsymbol{\alpha}$, the same headless stick carried along.

T3.4.2 Why the topology survives

The excursion sets. The lensed hot region is $\{\hat{\boldsymbol{n}} : T(\hat{\boldsymbol{n}} + \boldsymbol{\alpha}) > \nu\sigma_0\}$. Call the remapping $f(\hat{\boldsymbol{n}}) = \hat{\boldsymbol{n}} + \boldsymbol{\alpha}(\hat{\boldsymbol{n}})$. The lensed excursion set is exactly the set of points that f sends into the unlensed excursion set, $\tilde{Q}_\nu = f^{-1}(Q_\nu)$, with the same threshold, because lensing does not change σ_0 (Section T3.2.4). If f is smooth and has a smooth inverse, it neither tears regions apart nor glues them together, so $\tilde{Q}_\nu$ has as many pieces and as many holes as Q_ν. The Euler characteristic at every threshold is unchanged, realisation by realisation. The area and the boundary length are not protected in the same way: they change wherever the image is magnified or sheared, although on the full sky the first-order changes average out because the convergence integrates to zero.

The singularities. A lensed singularity sits wherever the unlensed polarization at $f(\hat{\boldsymbol{n}})$ vanishes, so the lensed zeros are the images under f^{-1} of the unlensed ones, one for one. Their indices are unchanged too. A small loop around a lensed zero is mapped by f to a small loop around its parent, and the stick along one loop is the stick along the other; if f keeps the orientation of the sky, the two loops are traversed in the same sense and the number of half-turns is the same. In general a continuous change can only move a singularity, or create or destroy a $+\frac{1}{2}$ and $-\frac{1}{2}$ together, since the total index inside any fixed large loop is an integer multiple of $\frac{1}{2}$ that cannot jump [1].

When is the remapping invertible? Locally, f stretches a small patch by the magnification matrix $\mathsf{A} = \mathbb{1} + \partial\boldsymbol{\alpha}$, with entries $\delta_{ij} + \partial_i\alpha_j$ [47] (their eq. 1.4). As long as $\det\mathsf{A} > 0$ everywhere, f is a local diffeomorphism that keeps orientation; it fails only where $\det\mathsf{A}$ passes through zero, on the caustics of strong lensing. With a convergence $\kappa = -\frac{1}{2}\nabla\cdot\boldsymbol{\alpha}$ whose rms is below 0.1 even on arcminute scales, CMB lensing is nowhere near that.

Example: how far from a caustic? In the 20 simulated 10° patches of Section T3.4.4 the deflection has an rms of 1.81′ (lower than the full-sky 2.7′, because the patch cannot hold the degree-scale modes that carry most of the deflection, and a uniform shift of the whole patch would not matter anyway), the convergence has an rms of 0.071, and the smallest value of $\det\mathsf{A}$ anywhere in any patch is 0.38. It never comes close to zero.

T3.4.3 What breaks the protection

The protection is exact for the remapping alone. Each of the following steps falls outside it.

- *Anything applied after lensing.* The telescope beam convolves the *lensed* sky. Smoothing then lensing is a pure remapping; lensing then smoothing is not, because lensing has already moved power to small scales that the beam partly removes. Two singularities of opposite index closer than the beam can be merged and annihilate; a thin hot bridge can be smoothed away. Instrumental noise, added last, creates and destroys features at the noise scale.
- *A finite patch, and local densities.* On a patch with an edge, features cross the edge as they move. Magnification also changes the number of features per unit *observed* area, even though the total is fixed, which is how lensing could leave the small-scale imprint on the singularity distribution that [2] anticipated.
- *Splitting into E and B.* Lensing turns part of the E pattern into B. The singularities of the full (Q, U) field are protected; those of an E-only or B-only map, built non-locally from it, are not.
- *The Gaussian formula itself.* The lensed C_ℓ has a larger $\tilde{\sigma}_1$, so plugging it into (T3.6), (T3.7) predicts more islands. The prediction is wrong for the remapped sky, which is no longer Gaussian: its islands are the old ones, deformed.

T3.4.4 Testing it: lensing simulated patches

`04_lensing_remap.py` draws 20 unlensed $10^\circ \times 10^\circ$ patches of T, Q and U (pixels of 0.59′) from CAMB's unlensed spectra, a lensing potential ψ from its spectrum $C_\ell^{\psi\psi}$ (CAMB's $C_\ell^{\phi\phi}$), and the deflection $\boldsymbol{\alpha} = \nabla\psi$; the lensed maps are the unlensed ones evaluated at $\boldsymbol{x} + \boldsymbol{\alpha}(\boldsymbol{x})$ by cubic-spline interpolation. Case A smooths with a 5′ beam and then lenses: a pure remapping of a smooth sky. Case B lenses the raw sky and then applies the beam, which is the order in which a real telescope sees it. For every lensed singularity in case A, the code follows the deflection back to $\boldsymbol{x} + \boldsymbol{\alpha}(\boldsymbol{x})$ and asks whether an unlensed singularity of the same index is there.

Figure T3.10 shows the outcome. In case A, 99.8% of the lensed singularities (never fewer than 99.6% in any patch) sit within a pixel and a half of the point the deflection points back to, on a parent of the same index; the rest are pixel-scale ambiguities in where a zero sits inside its 2×2 block. The Euler characteristic at every threshold and the singularity count change by at most 0.16% of their size, which is the accuracy of counting on a grid. In case B the singularity count changes by $-0.21 \pm 0.19\%$ and the Euler characteristic by up to $0.66 \pm 0.69\%$ of its peak, where each error is the patch-to-patch scatter of the change divided by $\sqrt{20}$. (In case A the same errors are at most 0.19%.) Lensing followed by the beam is no longer a pure remapping, so nothing protects the counts. But at a 5′ resolution the change is small: a few tenths of a per cent, no more than two or three times its error, so 20 patches cannot pin it down. The Gaussian formula with the lensed gradient variance predicts a change of +1.0%, which these patches can neither confirm nor rule out; and the formula has no right to be used here anyway, because the lensed sky is not a new Gaussian sky.

What this bought us. For the analysis of real maps the lesson is practical. A topological count (the Euler characteristic, the number of singularities and their indices) is insensitive to lensing to the extent that lensing acts as a remapping, and the residual comes from the instrument's smoothing and noise, which simulations of the instrument capture. The nuisance operator $\mathcal{N}$ of Chapter 12 here is "lens, then smooth, then add noise"; the first factor is harmless for topology, the last two are not, and that is where the Monte Carlo of Chapter 12 has to be careful. The

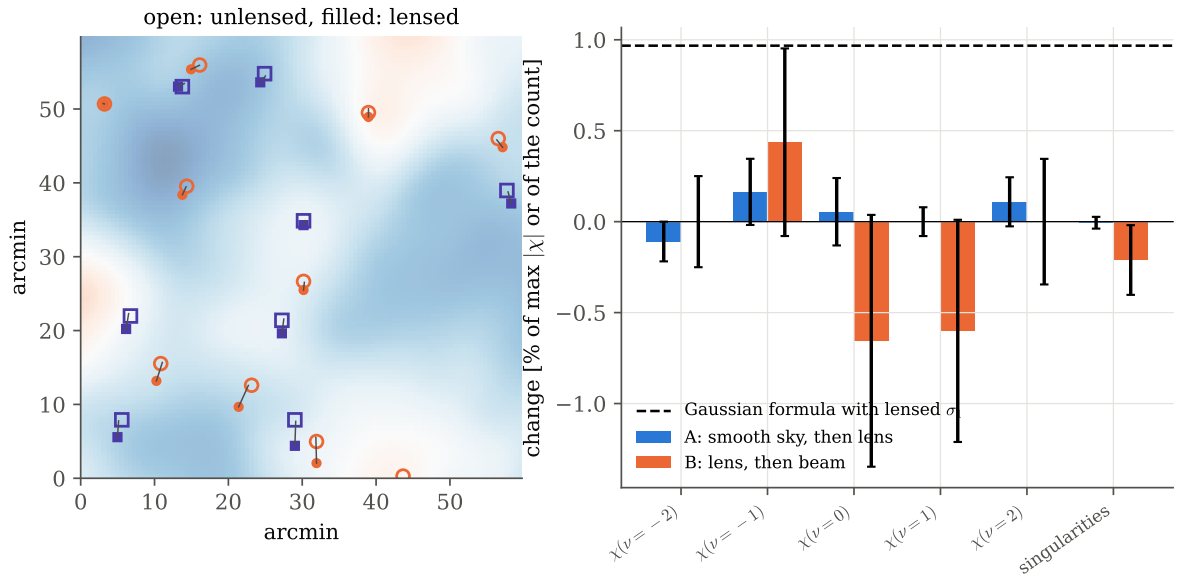


Figure T3.10: Left: a $1° \times 1°$ corner of one patch; open markers are unlensed singularities, filled markers the lensed ones (circles $+\frac{1}{2}$, squares $-\frac{1}{2}$), and each grey line joins a lensed singularity to the point its deflection traces back to. Every one lands on a parent of the same index, an arcminute or two away. Right: change of the Euler characteristic at five thresholds and of the singularity count. Case A (pure remapping) changes nothing beyond pixel-level effects; case B (beam after lensing) changes them by less than one per cent. Error bars: the patch-to-patch scatter of each change divided by $\sqrt{20}$. The dashed line is what the Gaussian formula with the lensed σ_1 would predict for the Euler characteristic.

full research version would lens full-sky HEALPix maps at $N_{\mathrm{side}} \geq 2048$ with the curved-sky remapping and parallel transport of [3], include noise, and use thousands of realisations; the flat $10°$ patches here miss the degree-scale deflection modes, which only shift a patch as a whole, and probe the effect at a single $5'$ resolution.

T3.5 What *Planck* found

By 2018 the *Planck* temperature maps were limited by cosmic variance over most of the sky and most angular scales: a better instrument would see the same sky. The final release asked, in two papers, the question that the shapes of the previous sections make sharp. Is the CMB a Gaussian, statistically isotropic field, as the simplest inflationary models predict [114]? And if a primordial three-point signal is there, how large can f_{NL} be [112]?

The question. Measured against simulations of the full *Planck* pipeline, do the shapes of the real maps look like the shapes of Gaussian, isotropic skies? If not, where do they differ, and how surprised should we be?

T3.5.1 The Minkowski-functional test

Planck computes the three functionals of Section T3.1.1 on masked maps at resolutions from $N_{\mathrm{side}} = 16$ to 2048: the area fraction as a pixel count, the boundary length from the contour lengths of all hot and cold spots, and the Euler characteristic as $(N_{\mathrm{hot}} - N_{\mathrm{cold}})/(2\pi A_{\mathrm{tot}})$, the number of compact hot spots minus compact cold spots per area, which is the count of pieces minus holes in another guise [114] (their eqs. 13–15). Two choices are worth noticing, because

Chapter 12 has to make them too. First, each functional is divided by its Gaussian amplitude, the factor A_k that carries all the dependence on σ_0 and σ_1, so that a Gaussian sky would give exactly $v_k(\nu) = e^{-\nu^2/2} H_{k-1}(\nu)$ (their eqs. 16–19, our (T3.2)–(T3.7) rearranged); the thresholds are in units of the map's own standard deviation, as in the pitfall of Section T3.1.3. This removes the large cosmic-variance fluctuation of the overall amplitude and keeps the shape. Second, polarization is tested through maps of the scalar E mode, because the functionals of Q and U depend on the orientation of the axes once a mask breaks the symmetry of the sky.

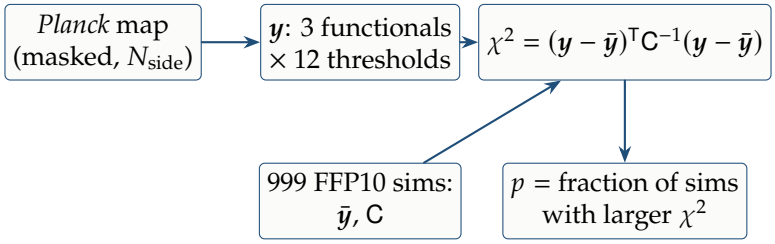


Figure T3.11: The *Planck* Minkowski-functional test. The data vector stacks the three normalised functionals at twelve thresholds between $-3\sigma_0$ and $3\sigma_0$. Its mean and covariance come from end-to-end simulations of Gaussian, isotropic skies processed exactly like the data, and the simulations also supply the distribution of χ^2 under that hypothesis.

The data vector $\boldsymbol{y}$ stacks v_0, v_1, v_2 at twelve thresholds from -3 to 3: 36 numbers per map (Figure T3.11). Its mean $\bar{\boldsymbol{y}}$ and covariance C are estimated from 999 FFP10 simulations, which contain the lensed Gaussian CMB, the instrument's noise, beams and scanning, and the component-separation pipeline. The statistic is $\chi^2 = (\boldsymbol{y} - \bar{\boldsymbol{y}})^{\mathsf{T}} \mathsf{C}^{-1} (\boldsymbol{y} - \bar{\boldsymbol{y}})$ (their eqs. 20–22), and the reported number is the fraction of simulations with a larger χ^2 than the data. This is a hypothesis test in exactly the sense of Chapter 4: the null hypothesis is the one whose sampling distribution we can generate, and the question is whether a χ^2 as large as the observed one, or larger, would be unremarkable if it were true.

Example: the scale of a χ^2 with 36 components. If $\boldsymbol{y}$ were exactly Gaussian with the estimated mean and covariance, χ^2 would follow a χ^2 distribution with 36 degrees of freedom: mean 36, standard deviation $\sqrt{2 \times 36} \approx 8.5$. A value of 53 would then sit at the 97th percentile. *Planck* does not rely on that shortcut: the simulations give the distribution directly, which also absorbs the noise in C estimated from a finite number of simulations (Chapter 12).

The result. For temperature, across four component-separation methods and eight resolutions, the probabilities of a larger χ^2 range from about 24% to 99.8%: no map is unusual (their Table 5). For E-mode polarization most values are similarly unremarkable; a few low ones (down to 1% for one method at $N_{\rm side} = 16$) appear for one method and not for the others, which points to residual systematics and noise in the polarization data rather than to the sky. Measured scale by scale, on needlet-filtered maps, the temperature functionals on the largest scales ($\ell \lesssim 32$) give probabilities of a few per cent (their Table 6), a mild echo of the large-angle anomalies below. Together with one-point moments, N-point correlation functions, peak statistics and the stacking of hot and cold spots, the verdict is: no significant departure from Gaussianity on intermediate and small scales [114].

T3.5.2 $f_{\rm NL}$ from the bispectrum

For a specific three-point shape the bispectrum is the optimal statistic, and *Planck* uses it for the local model (T3.8) and for two shapes typical of single-field models with unusual kinetic terms, the equilateral and orthogonal templates. The temperature-plus-E results, after subtracting the

expected lensing contribution, are [112]

$$f_{\rm NL}^{\rm local} = -0.9 \pm 5.1, \qquad f_{\rm NL}^{\rm equil} = -26 \pm 47, \qquad f_{\rm NL}^{\rm ortho} = -38 \pm 24 \qquad (68\%). \tag{T3.15}$$

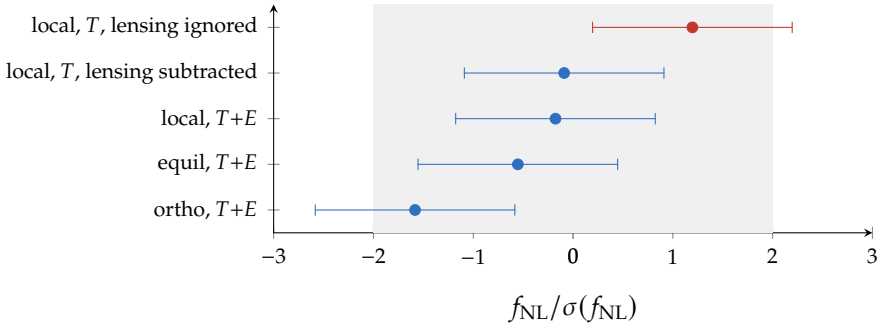


Figure T3.12: *Planck* 2018 $f_{\rm NL}$ estimates in units of their own standard deviations, with $\pm1\sigma$ bars; the shaded band is $\pm2\sigma$. All are consistent with zero. The red point is the temperature-only local estimate before the lensing contribution is subtracted: lensing alone moves it by more than one standard deviation.

Example: what ±5.1 means for the map. By the Sachs–Wolfe estimate of Section T3.2.3, $f_{\rm NL} = 5$ skews the large-angle temperature by $|6a| \approx 2 \times 10^{-3}$. The bispectrum detects a three-point signal of this size because it adds up the contributions of millions of triangles of multipoles, each weighted by how well it matches the template; the three Minkowski functionals compress the same map into 36 numbers that are not tuned to any one shape. That is the division of labour: the bispectrum to measure a known shape, the functionals to look for anything unexpected.

T3.5.3 The anomalies

A handful of large-angle features, first seen by WMAP, are confirmed in the *Planck* temperature maps [114] (their §§6–7). The variance of the map at low resolution is low: about 1% of the simulations have a variance as small. The angular correlation function $C(\theta)$ is almost zero beyond 60°: the statistic $S_{1/2} = \int_{-1}^{1/2} [C(\theta)]^2 \, {\rm d}\cos\theta$ is smaller in the data than in all but about 1% of the simulations, and partly because the quadrupole is low (removing it lowers the significance from about 99% to 86% of simulations having a larger value). One half of the sky has more power than the other, a dipolar modulation of the fluctuation amplitude. Odd multipoles carry more power than even ones at the lowest ℓ, a preference for odd point-parity. And there is the Cold Spot, an unusually deep and wide cold region in the southern hemisphere.

Each of these is a deviation at roughly the 2–3σ level, and each was noticed *before* a statistic was designed to measure it. That makes the quoted significances hard to interpret: among the many statistics one could have computed, some are bound to look unusual by chance (the look-elsewhere effect). Since the temperature on these scales is cosmic-variance limited, more temperature data cannot settle it. Polarization can, because it is generated by different modes. The 2018 polarization maps show no corresponding anomaly (no lack of large-angle correlation, no hemispherical asymmetry, no point-parity violation, no polarization signature of the Cold Spot), but their signal-to-noise on these scales is low, and the absence is not yet strong evidence either way [114].

How this is used in real research: verification and research with the same machinery

The Gaussian checks of Sections T3.1.3 and T3.3.2 *verify* known formulas: the simulations had to reproduce (T3.6), (T3.7) and (T3.13), and they did. In the *Planck* test the same machinery is the *research instrument*: for a masked sky with anisotropic noise, component-separation residuals and lensing, no formula gives the mean, the covariance or the χ^2 distribution of the functionals, and a thousand end-to-end simulations are the only way to get them. Chapter 12 builds that test on simulated skies; Chapter 13 applies the summary statistics to real *Planck* maps.

T3.6 The topology of the cosmic web

The same questions can be asked one dimension up. Galaxy surveys and simulations show the matter of the late universe arranged in a web: dense clusters at the nodes, filaments joining them, thin walls, and large empty voids. In 1986, before the web had a name, Hamilton, Gott and Weinberg asked whether its connectivity looked like that of a Gaussian field: were the dense regions isolated clumps in an empty background (a "meatball" topology), isolated voids in a dense background ("Swiss cheese"), or two interpenetrating networks, like a sponge in which both the material and the holes are connected [115]?

The question. If we take the density contrast δ of a three-dimensional field, smoothed on some scale, and keep the region with $\delta > \nu\sigma$, how connected is it at each ν, and does a Gaussian field predict a sponge at the median density?

T3.6.1 The genus of a Gaussian field

In three dimensions the boundary of the excursion set is a surface, and its connectivity is measured by its **genus**: roughly, the number of tunnels through the set minus the number of separate pieces of the surface. A doughnut contributes +1 more than a ball. Per unit volume the convenient definition is $g = -\chi/V$, with χ the Euler characteristic of the set itself [5]: every tunnel adds +1 to g, while every isolated clump and every isolated void adds −1. For a Gaussian field the expectation is [113, 115]

$$g(\nu) = \frac{1}{4\pi^2}\left(\frac{\langle k^2\rangle}{3}\right)^{3/2}(1-\nu^2)\,e^{-\nu^2/2}, \qquad \langle k^2\rangle = \frac{\langle|\nabla\delta|^2\rangle}{\langle\delta^2\rangle}, \tag{T3.16}$$

the three-dimensional counterpart of (T3.6), (T3.7): again the spectrum enters only through the ratio of the gradient variance to the field variance, and the shape in ν is a Hermite polynomial times the Gaussian, now $-H_2(\nu) = 1-\nu^2$. Read it from the middle outwards. Near $\nu = 0$ the genus is positive: many tunnels, a sponge. For $|\nu| > 1$ it is negative: at high thresholds isolated clusters (meatballs), at low thresholds isolated voids (Swiss cheese). A Gaussian field is symmetric between the two, so the curve is even in ν. Gravitational evolution breaks that symmetry, shifting the curve towards a meatball topology, and measuring that shift, at fixed volume fraction as in Section T3.2.2, is the classic use of the genus statistic [113].

Example: a simulated box. `05_web_genus.py` draws 50 periodic boxes of 128^3 cells of side $2\,h^{-1}$Mpc, with a power spectrum $P(k) \propto k^{-1.5}$ smoothed on $6\,h^{-1}$Mpc, and counts the Euler characteristic of the excursion sets as vertices minus edges plus faces minus cubes of the union of voxels, the three-dimensional version of the counter of Section T3.1.3 (on one box it agrees with scikit-image: 128 and 128). At the median density a box has a genus of 259, against 260

from (T3.16); over all thresholds the mean of the boxes stays within 1.6% of the peak of the formula, with a Monte Carlo error of 1.3% (Figure T3.13). A Gaussian field is a sponge.

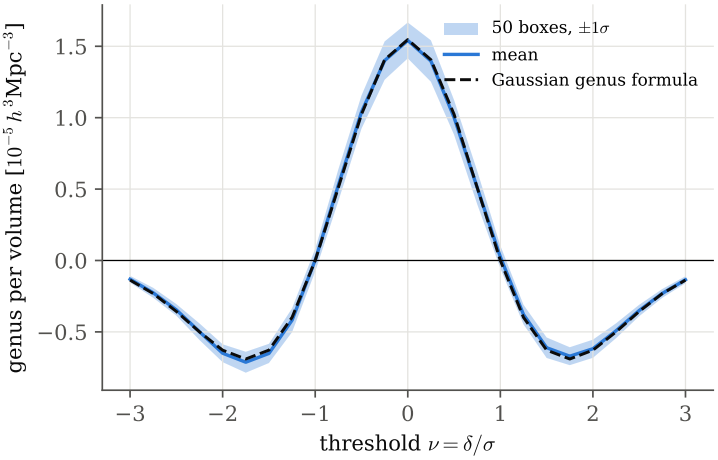


Figure T3.13: Genus per unit volume of the excursion sets of simulated three-dimensional Gaussian fields against the threshold. Band: box-to-box scatter; solid: mean of 50 boxes; dashed: the Gaussian formula (T3.16). Positive genus near the median density means many tunnels (a sponge); negative genus at $|\nu| > 1$ means isolated clumps or isolated voids.

T3.6.2 Betti numbers and persistence

The Euler characteristic is an alternating sum, $\chi = \beta_0 - \beta_1 + \beta_2$, of three counts, the **Betti numbers** [5]: β_0 counts separate pieces (clusters), β_1 independent tunnels or loops (the loops formed by filaments), and β_2 enclosed cavities (voids sealed inside the set); see the topology notes for why they are well defined. A sum with alternating signs can hide its terms: two isolated clusters and one hollow shell around a void both have $\chi = 2$ (Figure T3.14). Pranav and collaborators showed that the Betti numbers separately, as functions of the threshold, distinguish web-like models that the genus alone does not, and they introduced *persistence*: follow each feature as the threshold is lowered, record the level at which it appears and the level at which it merges or fills in, and plot the pairs [5]. A feature that persists over a wide range of thresholds is a robust structure; one that lives only briefly is noise. Chapter 12 builds these tools from scratch.

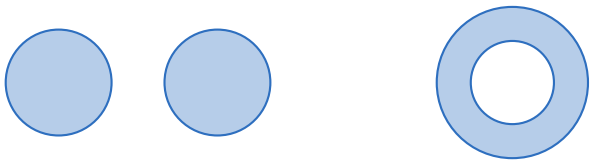


Figure T3.14: Cross-sections of two three-dimensional sets with the same Euler characteristic. Left: two separate solid balls. Right: a single hollow ball, whose interior cavity is a sealed void. The Euler characteristic cannot tell them apart; the Betti numbers can.

Three later studies show what this buys in practice. In ΛCDM simulations, the persistence diagrams of clusters, filament loops and voids have sharp apexes, and the apex for clusters sits at the density at which matter typically detaches from the Hubble expansion and begins to

collapse; persistence diagrams track the hierarchical build-up of the web more informatively than the Euler characteristic or the Betti numbers alone [4]. In halo catalogues from N-body simulations, persistent homology detects a local primordial non-Gaussianity of $f_{\rm NL}^{\rm loc} = 10$ at 97.5% confidence in about 85% of 40 $(h^{-1}\mathrm{Gpc})^3$ volumes, using information at $\sim 10\,h^{-1}$Mpc that is complementary to the scale-dependent bias of the power spectrum [116]. And in weak-lensing mass maps, persistent Betti numbers fed to an emulator and an MCMC give constraints on $S_8 = \sigma_8\sqrt{\Omega_m/0.3}$ that are 5% tighter than peak counts for a KiDS-like survey and, for a 100 deg^2 Euclid-like mock, 18% tighter on S_8 and 10% on Ω_m [56]. In each case the topological summary is an ordinary random observable, with a mean and a covariance taken from simulations and a likelihood built on them, which is the machinery of Chapters 9, 10 and 12.

The full research version of the experiment above would replace the Gaussian boxes by N-body simulations or galaxy catalogues of a few $h^{-3}\mathrm{Gpc}^3$, add redshift-space distortions and a survey mask, and compute Betti curves and persistence diagrams; the 128^3 Gaussian boxes here only verify the Gaussian baseline that those analyses measure departures from.

Chapter summary

The microwave sky carries information in its pattern, not only in its amplitudes. Its temperature map is cut at thresholds into excursion sets whose area, coastline and connectivity are three numbers per threshold; a Gaussian sky fixes all three through σ_0 and σ_1, a pointwise skewness merely relabels the threshold, and what survives at fixed area fraction is non-local physics. Its polarization has half-integer singularities whose number is a moment of the spectrum and whose arrangement is robust. Lensing remaps both patterns smoothly, which leaves their topology intact until the instrument smooths and adds noise. *Planck* found the sky consistent with Gaussianity and isotropy apart from a few large-angle anomalies of debated significance, and $f_{\rm NL}$ consistent with zero. Table T3.1 collects the pieces.

Reading the literature

- J. Schmalzing and K. M. Górski, Minkowski functionals used in the morphological analysis of CMB anisotropy maps [110]: the three functionals on the sphere, their Gaussian expectations and pixel estimators.
- T. Matsubara, Statistics of smoothed cosmic fields in perturbation theory [113]: the first non-Gaussian correction to the genus and the Minkowski functionals, and the area-fraction threshold.
- J. M. Bardeen, J. R. Bond, N. Kaiser and A. S. Szalay, The statistics of peaks of Gaussian random fields [111]: spectral moments and the geometry of Gaussian fields.
- A. J. S. Hamilton, J. R. Gott and D. Weinberg, The topology of the large-scale structure of the universe [115]: the genus curve and the sponge.
- T. Vachaspati and A. Lue [1] and D. Huterer and T. Vachaspati [2]: polarization singularities, their index, and their distribution in simulated maps.
- A. Challinor and G. Chon, Geometry of weak lensing of CMB polarization [3], and A. Lewis and A. Challinor, Weak gravitational lensing of the CMB [47]: the lensing remapping on the sphere.
- *Planck* 2018 results VII and IX [112, 114]: tests of isotropy and Gaussianity, the anomalies, and $f_{\rm NL}$.

Concept	Operational meaning	Used in
excursion set Q_ν (T3.1)	pixels with $T > \nu\sigma_0$	Chapter 12
v_0, v_1, v_2 (Section T3.1.1)	area fraction; $\frac{1}{4}$ boundary length per area; pieces minus holes per area	Chapters 12 and 13
$\sigma_0, \sigma_1, \theta_c$ (T3.3)–(T3.5)	variance of the map and of its gradient from C_ℓ; their ratio is the correlation angle	Chapter 12
Gaussian functionals (T3.2), (T3.6), (T3.7)	$\frac{1}{2}$ erfc, $e^{-\nu^2/2}$, $\nu e^{-\nu^2/2}$ curves; derived in Chapter 12	Chapter 12
skewness parameters, Matsubara correction (T3.9)–(T3.10)	first-order non-Gaussian change of the curves	Chapter 12
area threshold ν_A (Section T3.2.2)	label sets by area fraction; removes pointwise distortions	Chapter 12
local $f_{\rm NL}$ (T3.8)	quadratic piece of the potential; $a = -3f_{\rm NL}\sigma_t$ in the Sachs–Wolfe limit	Chapters 12 and 13
polarization index (T3.12)	half-turns of the stick around a zero of $Q + iU$; $\pm\frac{1}{2}$	Chapter 12
singularity density (T3.13), (T3.14)	$\sigma_1^2/4\pi\sigma_0^2$; full sky $\langle\ell(\ell+1)\rangle_P$	Chapter 12
lensing remapping (Section T3.4.1)	$\tilde{X}(\hat{\boldsymbol{n}}) = X(\hat{\boldsymbol{n}} + \boldsymbol{\alpha})$, polarization parallel-transported	Chapters 12 and 13
protected / unprotected (Sections T3.4.2 and T3.4.3)	topology kept by an invertible remapping; broken by beam, noise, edges, E/B split	Chapter 12
MF χ^2 test (Section T3.5.1)	36-vector against end-to-end simulations; p from simulations	Chapters 4, 12 and 13
Planck $f_{\rm NL}$ (T3.15)	bispectrum constraints; lensing bias	Chapters 12 and 13
genus, Betti numbers (Sections T3.6.1 and T3.6.2)	$g = -\chi/V$; $\chi = \beta_0 - \beta_1 + \beta_2$; persistence	Chapters 11 and 12

Table T3.1: Every concept of the chapter, how to compute or recognise it, and where the book uses it next.

- P. Pranav et al. [5], G. Wilding et al. [4], M. Biagetti, A. Cole and G. Shiu [116], S. Heydenreich, B. Brück and J. Harnois-Déraps [56]: Betti numbers and persistence in the cosmic web, in halo catalogues and in weak lensing.

Recap

- The Minkowski functionals of the excursion sets (area fraction, boundary length, Euler characteristic) are the complete set of additive, motion-invariant shape numbers in two dimensions.
- For a Gaussian sky: $v_0 = \frac{1}{2}\,\mathrm{erfc}(\nu/\sqrt{2})$ needs no spectrum; $v_1 \propto e^{-\nu^2/2}/\theta_c$ and $v_2 \propto \nu e^{-\nu^2/2}/\theta_c^2$ need only $\theta_c = \sigma_0/\sigma_1$.

- A pointwise increasing distortion only relabels thresholds; at fixed area fraction it vanishes exactly. Local $f_{\rm NL}$ is such a distortion on Sachs–Wolfe scales.
- Polarization singularities have index $\pm\frac{1}{2}$; their density is $\sigma_1^2/4\pi\sigma_0^2$, about $\langle\ell(\ell+1)\rangle_P$ on the full sky; opposite indices attract, equal ones repel.
- Lensing is a smooth invertible remapping: it moves singularities and islands but keeps their number and index; the beam and noise applied afterwards break the protection slightly.
- *Planck*: Gaussian and isotropic within errors on intermediate and small scales; a few large-angle anomalies at the 2–3σ level; $f_{\rm NL}^{\rm local} = -0.9 \pm 5.1$.
- In three dimensions the genus curve of a Gaussian field is a sponge at the median density; Betti numbers and persistence separate what the genus adds together.

Chapter T4

Particle Physics for the Data Analyst

Where this chapter is going

Two beams of protons circle the Large Hadron Collider in opposite directions and cross each other forty million times a second. At every crossing a few dozen pairs of protons collide. Almost all of these collisions are ordinary: the protons break up into sprays of light particles that the theory has described for decades. Roughly once in a billion and a half collisions something rarer happens, and a Higgs boson is made. It lives for about 10^{-22} s, far too short to travel a measurable distance, and in about two decays out of a thousand it turns into two photons. The detector never sees the Higgs boson. It sees two photons, measures their energies and their directions, and writes these numbers to disk together with everything else that happened in that crossing.

The measured energies and opening angle of the two photons determine their *invariant mass*, a quantity independent of the motion of their parent particle. Photon pairs from Higgs decays cluster around 125 GeV with a width set largely by detector resolution, while background pairs populate a smooth, falling spectrum. Combining millions of events in a mass histogram therefore lets a small Higgs peak emerge above the background. This is the diphoton measurement used in the 2012 discovery and the one we reconstruct from public ATLAS data in Chapter 13.

The statistics of the search, worked out in Chapter 4, needs one thing from physics: the *expected* number of events in each bin of that histogram, as a function of a few parameters. The cross section σ sets the production rate per unit instantaneous luminosity, so multiplying by the recorded integrated luminosity L gives the expected number produced. The acceptance A and efficiency ε determine how many events enter the analysis. A template from detector simulation then distributes those events among the mass bins. In the fit, the signal strength μ scales the nominal signal prediction, while nuisance parameters $\boldsymbol{\theta}$ carry calibration uncertainties into the predicted counts (Figure T4.1).

A collider produces collisions at a known rate, and the expected number of events of any kind is its cross section times the integrated luminosity, $N = \sigma L$; the formula follows from the overlap of two bunches of protons, and the number actually recorded is a Poisson count (Section T4.2). What is counted is a recorded event: a list of reconstructed particles, each with a transverse momentum p_T, a pseudorapidity η and an azimuth ϕ (Section T4.1). From two of these particles comes the invariant mass, which is the same in every decay of a given particle; its spread comes from the lifetime of the particle and from the resolution of the detector, and both shapes can be derived (Section T4.3). Simulated events and the side regions of the data give the expected shapes of signal and background, the templates, and the signal strength μ scales the predicted signal (Section T4.4). Every input is known only to some precision; each imprecision becomes a nuisance parameter with its own calibration measurement (Section T4.5). All of

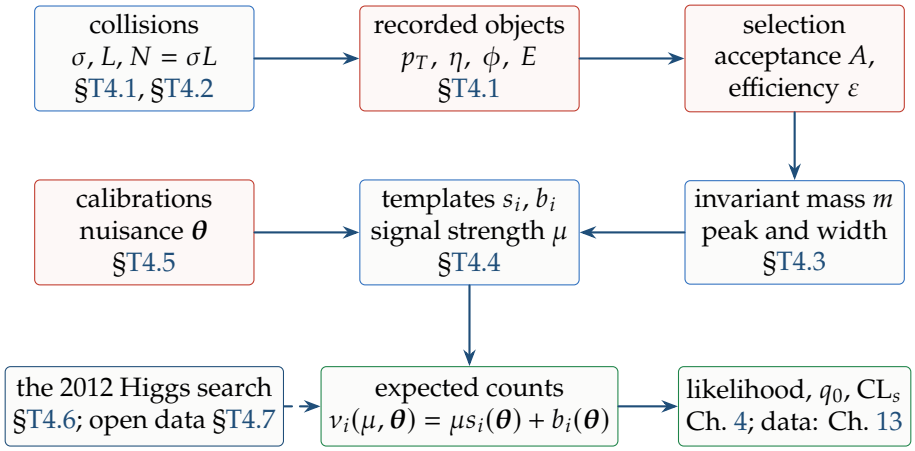


Figure T4.1: The path from collisions to the numbers the statistics uses. Blue boxes are physics that theory and simulation predict: how often a process happens, and where its events land on the mass axis. Red boxes are what the detector and the analysis do to the events: record them, keep some of them, and get calibrated. The two green boxes are where the statistics of Chapter 4 takes over: every arrow above ends in the expected count ν_i of one histogram bin, and the likelihood compares those expectations with the observed counts.

these were at work in the two searches that found the Higgs boson in 2012 (Section T4.6), and every one has a column, or a combination of columns, in the open-data files that Chapter 13 analyses (Section T4.7). Table T4.3 lists each symbol with its typical size and the place in the book where it is used.

The likelihood of a histogram, the treatment of nuisance parameters, the discovery statistic q_0, the look-elsewhere effect and the CL_s limits are built on these expected counts in Chapter 4 (Sections 4.14 to 4.16); undoing the detector's blurring is Section 6.14. In the language of the book's pipeline (Chapter 0), the collider case reads as follows. The physical model is the Standard Model with or without the Higgs signal. The data set is the list of recorded events. What matters in it is the invariant mass of the photon pair, and the summary is a histogram of that mass. The sampling distribution of each bin is Poisson, with a mean that physics supplies. Inference then asks the question of Chapter 5 in its collider form: which signal strengths, and which values of the calibration constants, could have produced the histogram we recorded?

T4.1 What a collider does, and what an event is

A proton is not a hard ball. It is a bound state of quarks and gluons, together called **partons**, which share its momentum among themselves in a way that changes from one moment to the next. When two bunches of protons cross at the LHC, most protons pass each other untouched. A few pairs come close enough for one parton of each to collide hard. Those two partons, not the whole protons, make the interesting new particles. Each parton carries an unknown fraction of its proton's momentum, so the collision we care about happens in a frame that moves along the beam line at a speed we do not know. Everything else in the two protons sprays out as low-energy debris, mostly close to the beam.

Around the collision point sits the detector, a set of nested cylinders, each built to measure one kind of particle. A crossing of two bunches lasts a fraction of a nanosecond, and the next one comes 25 ns later. Only a few hundred crossings per second can be written to disk, so fast electronics, the **trigger**, decide within microseconds whether a crossing looks interesting enough to keep, for example because two large deposits of electromagnetic energy appeared in the calorimeter [117]. What is kept is the raw response of millions of detector channels.

Reconstruction programs then turn these responses into a short list of particles with measured momenta. That list is what an analyst works with.

The questions in words.

1. What does the detector record, and how does a photon look different from an electron, a muon, a jet or a neutrino?
2. Which numbers describe a recorded particle, and why does a hadron collider describe directions by a pseudorapidity η instead of the polar angle θ?
3. What is the energy of a photon in terms of those numbers?

T4.1.1 The detector, layer by layer

The ATLAS detector used for the Higgs search has four layers [118, Sec. 2], [117]. Going outwards from the beam:

- The **inner detector**: layers of silicon sensors and straw tubes inside a 2 T magnetic field. A charged particle leaves a trail of hits, and the curvature of the trail in the field gives its momentum. Neutral particles leave nothing here.
- The **electromagnetic calorimeter**: liquid argon between lead plates. Photons and electrons dump all their energy in it, in a narrow shower. Its first layer is finely segmented in angle, so that a single photon can be told apart from the two nearby photons of a decaying neutral pion.
- The **hadronic calorimeter**: iron and scintillating tiles. Protons, pions and the other strongly interacting particles shower here.
- The **muon spectrometer**: tracking chambers in a toroidal magnetic field outside everything else. Muons are the only charged particles that cross the calorimeters.

Each kind of particle leaves a different pattern across the layers, and the pattern is its identity. Figure T4.2 draws the five patterns that matter.

Working definition T4.1 (event and reconstructed object)

An **event** is everything recorded in one bunch crossing that the trigger kept. A **reconstructed object** is one particle candidate built from it, a photon, electron, muon, jet or hadronic tau decay, with four measured numbers: its transverse momentum p_T, pseudorapidity η, azimuth ϕ and energy E, plus quality flags. *Operationally:* an event in the files of Section T4.7 is one row of a table; each kind of object fills a set of columns that hold one entry per object of that kind [a row with three photons has three numbers in the photon-p_T column].

Two of the identity rules need a word more. A photon can convert into an electron–positron pair in the material in front of the calorimeter; it then leaves one or two tracks that start in the middle of the inner detector instead of at the collision point, and is recognised as a *converted* photon. In the open data, about a third of the selected photons are converted (34.66%). And a jet is not one particle but the collimated spray of hadrons that a quark or gluon becomes; its energy is the sum of everything in a cone around its axis.

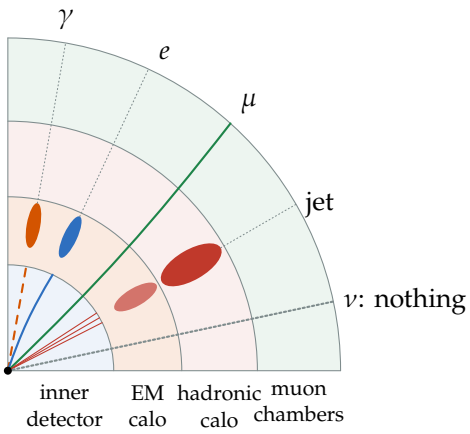


Figure T4.2: A quarter of the detector seen along the beam, with the beam at the lower left corner, and the five signatures. A photon leaves no track and a narrow shower in the electromagnetic calorimeter. An electron leaves a track and the same narrow shower. A muon crosses everything. A jet, the spray of hadrons from a quark or gluon, leaves several tracks and showers in both calorimeters. A neutrino leaves nothing at all; it shows up only as an imbalance of the momenta of everything else.

T4.1.2 Coordinates: why the transverse plane

ATLAS uses a right-handed frame with the origin at the collision point and the z axis along the beam [118, footnote 2]. The polar angle θ is measured from the z axis, and the azimuth ϕ around it, in the plane perpendicular to the beam called the **transverse plane**. A particle's momentum splits into a part along the beam, $p_z = |\boldsymbol{p}| \cos\theta$, and a part in the transverse plane, of size

$$p_T = \sqrt{p_x^2 + p_y^2} = |\boldsymbol{p}| \sin\theta. \tag{T4.1}$$

Why the transverse part is special. Before the collision, the two partons move along the beam. Their momenta along z are $x_1 E_\mathrm{b}$ and $-x_2 E_\mathrm{b}$, where E_b is the beam energy and x_1, x_2 are the momentum fractions the two partons happened to carry. Neither fraction is measured, so the total z momentum of the hard collision, $(x_1 - x_2)E_\mathrm{b}$, is unknown in every event. Their transverse momenta, on the other hand, are close to zero. Momentum conservation then says that the transverse momenta of everything produced add up to zero, while it says nothing useful about the z momenta. Two consequences follow. A particle that escapes unseen, like a neutrino, shows up as an imbalance of the transverse momenta of the visible particles, the **missing transverse momentum** $-\sum_i \boldsymbol{p}_{T,i}$ [119, (49.49)]; and the description of a particle should use quantities that do not change when the whole event is boosted along z by an unknown amount.

A variable that adds under boosts. The **rapidity** of a particle is [119, (49.40)]

$$y = \frac{1}{2} \ln \frac{E + p_z}{E - p_z}. \tag{T4.2}$$

What happens to it under a boost along z with velocity β (in units of $c = 1$) and $\gamma = (1-\beta^2)^{-1/2}$? The Lorentz transformation of [119, (49.1)] gives $E' = \gamma(E - \beta p_z)$ and $p_z' = \gamma(p_z - \beta E)$. Then

$$\begin{aligned} y' &\overset{(1)}{=} \frac{1}{2} \ln \frac{\gamma(E - \beta p_z) + \gamma(p_z - \beta E)}{\gamma(E - \beta p_z) - \gamma(p_z - \beta E)} \overset{(2)}{=} \frac{1}{2} \ln \frac{(1-\beta)(E + p_z)}{(1+\beta)(E - p_z)} \\ &\overset{(3)}{=} y + \frac{1}{2} \ln \frac{1-\beta}{1+\beta} \overset{(4)}{=} y - \tanh^{-1}\beta. \end{aligned} \tag{T4.3}$$

(1) Insert the boosted energy and momentum into (T4.2).
(2) Collect terms: the numerator is $\gamma(E+p_z)(1-\beta)$ and the denominator $\gamma(E-p_z)(1+\beta)$; the γ's cancel.
(3) The logarithm of a product is the sum of the logarithms.
(4) $\tanh^{-1}\beta = \frac{1}{2}\ln\frac{1+\beta}{1-\beta}$, the inverse of $\tanh u = (e^u - e^{-u})/(e^u + e^{-u})$.

A boost along the beam shifts every rapidity in the event by the same amount. The *difference* of two rapidities is therefore the same in every frame, and so is the shape of a distribution of rapidities [119, §49.5.2]. The unknown motion of the parton pair along the beam drops out of every rapidity difference.

From rapidity to an angle. Rapidity needs the energy and p_z, and so the mass. For a massless particle, or any particle with $|\boldsymbol{p}| \gg m$, $E = |\boldsymbol{p}|$ and $p_z = |\boldsymbol{p}|\cos\theta$, and the rapidity becomes a function of the angle alone:

$$y \overset{(1)}{=} \frac{1}{2}\ln\frac{1+\cos\theta}{1-\cos\theta} \overset{(2)}{=} \frac{1}{2}\ln\frac{2\cos^2(\theta/2)}{2\sin^2(\theta/2)} \overset{(3)}{=} -\ln\tan\frac{\theta}{2} \;\equiv\; \eta. \tag{T4.4}$$

(1) Divide numerator and denominator of (T4.2) by $|\boldsymbol{p}| = E$.
(2) The half-angle identities $1+\cos\theta = 2\cos^2(\theta/2)$ and $1-\cos\theta = 2\sin^2(\theta/2)$.
(3) $\frac{1}{2}\ln(\cot^2 u) = \ln\cot u = -\ln\tan u$.

This is the **pseudorapidity** [119, (49.47)]. It is defined for every particle as $-\ln\tan(\theta/2)$, whatever its mass, because the angle is always measured; for photons it is exactly the rapidity, and for light particles at high momentum it is very nearly so.

Working definition T4.2 (pseudorapidity)

The **pseudorapidity** of a particle at polar angle θ from the beam is $\eta = -\ln\tan(\theta/2)$, and inversely $\theta = 2\arctan e^{-\eta}$. It is 0 perpendicular to the beam and grows without bound towards the beam. *Operationally:* $\eta = \sinh^{-1}(p_z/p_T)$ from the measured momentum components [a form that follows from $\sinh\eta = \cot\theta$, derived below]. For a massless particle, differences of η do not change under boosts along the beam.

Energy and momentum from (p_T, η, ϕ). Write $t = \tan(\theta/2)$, so $e^\eta = 1/t$ by (T4.4). Then

$$\sinh\eta \overset{(1)}{=} \frac{1}{2}\left(\frac{1}{t} - t\right) \overset{(2)}{=} \frac{1-t^2}{2t} \overset{(3)}{=} \cot\theta, \qquad \cosh\eta \overset{(1)}{=} \frac{1}{2}\left(\frac{1}{t} + t\right) \overset{(2)}{=} \frac{1+t^2}{2t} \overset{(3)}{=} \frac{1}{\sin\theta}. \tag{T4.5}$$

(1) The definitions $\sinh\eta = (e^\eta - e^{-\eta})/2$ and $\cosh\eta = (e^\eta + e^{-\eta})/2$, with $e^\eta = 1/t$.
(2) Put both terms over the common denominator $2t$.
(3) The double-angle formulas in terms of $t = \tan(\theta/2)$: $\tan\theta = 2t/(1-t^2)$ and $\sin\theta = 2t/(1+t^2)$.

Combining these with (T4.1), a particle's momentum components and, for a massless particle, its energy are

$$p_x = p_T\cos\phi, \qquad p_y = p_T\sin\phi, \qquad p_z = p_T\sinh\eta, \qquad E = |\boldsymbol{p}| = \frac{p_T}{\sin\theta} = p_T\cosh\eta. \tag{T4.6}$$

The first two are the definition of the azimuth; the third is $p_z = |\boldsymbol{p}|\cos\theta = p_T\cot\theta$ with $\cot\theta = \sinh\eta$; the last uses $\cosh\eta = 1/\sin\theta$. These four numbers are all that the invariant mass of Section T4.3 needs.

Example: angles and pseudorapidities. At $\theta = 90°$, $\tan 45° = 1$ and $\eta = 0$. At $\theta = 45°$, $\tan 22.5° = 0.4142$ and $\eta = -\ln 0.4142 = 0.881$. At $\theta = 10°$, $\tan 5° = 0.08749$ and $\eta = 2.436$. The

pseudorapidity grows slowly over most of the detector and fast near the beam: the last 10° next to the beam cover more than $\eta = 2.4$ onwards. Going the other way, the ATLAS photon acceptance ends at $|\eta| = 2.37$, which is $\theta = 2\arctan e^{-2.37} = 2\arctan 0.0935 = 10.7°$ from the beam, and the gap between the barrel and the end-cap calorimeters, $1.37 < |\eta| < 1.52$, is the thin cone between $\theta = 24.7°$ and 28.5° (on each side), where cables and services pass and photons are poorly measured [118, Sec. 5.1]. Figure T4.3 draws these directions.

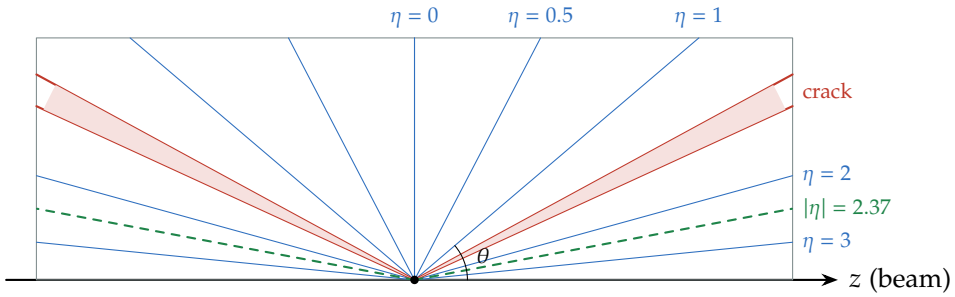


Figure T4.3: Directions of constant pseudorapidity in the plane containing the beam, drawn at their true angles $\theta = 2\arctan e^{-\eta}$. Equal steps in η crowd towards the beam. The shaded red wedges are the barrel/end-cap gap $1.37 < |\eta| < 1.52$, and the green dashed lines mark the edge $|\eta| = 2.37$ of the region where photons are accepted.

Example: a photon's energy from its three numbers. A photon is recorded with $p_T = 50\,\text{GeV}$ and $\eta = 1.2$. Then $\sinh 1.2 = 1.5095$ and $\cosh 1.2 = 1.8107$, so by (T4.6) $p_z = 75.5\,\text{GeV}$ and $E = 90.5\,\text{GeV}$. As a check, $p_T^2 + p_z^2 = 2500 + 5697 = 8197\,\text{GeV}^2 = E^2$, as it must be for a massless particle. A photon going straight up, at $\eta = 0$, would have $E = p_T$; the same p_T further forward carries more energy, because most of its momentum then points along the beam.

Example: why the beam direction is lost. Suppose the two partons carry $x_1 = 0.02$ and $x_2 = 0.005$ of their protons' momenta at $E_\text{b} = 6.5\,\text{TeV}$. The parton pair then has energy $(x_1 + x_2)E_\text{b} = 162.5\,\text{GeV}$ and momentum $(x_1 - x_2)E_\text{b} = 97.5\,\text{GeV}$ along $+z$, so it moves with $\beta = 97.5/162.5 = 0.6$ and every particle it makes has its rapidity shifted by $\tanh^{-1} 0.6 = \frac{1}{2}\ln 4 = 0.693$, by (T4.3) read backwards. In the next event the fractions are different and so is the shift. A search that looked at η itself would see the decay products smeared by this random shift; one that looks at p_T, at differences of η, or at the invariant mass, does not. The collision energy available to make new particles is $\sqrt{x_1 x_2}\,\sqrt{s} = \sqrt{10^{-4}} \times 13\,\text{TeV} = 130\,\text{GeV}$ in this example, which is why a 125 GeV Higgs boson needs only a small slice of the 13 TeV of the protons.

Watch out: Pseudorapidity is an angle, rapidity is a boost

Only rapidity differences are exactly unchanged by a boost along the beam. For a massive particle, such as a heavy jet or a slow proton, η and y differ, and η differences do change under the boost. For photons and electrons at the energies of this chapter the two are equal to better than one part in 10^6, and the distinction never matters.

Where this is used. The photon columns of the open data (Section T4.7) are exactly p_T, η, ϕ and E. The acceptance cuts $|\eta| < 2.37$ outside 1.37–1.52 appear in every diphoton selection of the book, from the toy of Section T4.2 to the capstone of Chapter 13. The identity $E = p_T \cosh\eta$ is checked on a real event in Section T4.7.

T4.2 Cross section, luminosity and the expected number of events

Before a search can ask whether the Higgs boson is in the data, it has to say how many Higgs bosons the data *should* contain if the Standard Model is right. The production cross section measures the rate of the process per unit luminosity and has dimensions of area. Accelerator physicists measure the integrated luminosity, which accounts for the exposure supplied by the beams and has dimensions of inverse area. Their product is therefore the expected number of Higgs bosons produced. A third factor comes from the experiment itself: of the events produced, only some land in the detector where photons can be measured, and only some of those survive the selection.

The questions in words.

1. What is a cross section, as a physical picture we could draw?
2. How many collisions do two crossing bunches of protons give, and how does that turn into the luminosity?
3. Of the events produced, what fraction is kept, and why is the number kept a Poisson count with mean $\sigma L \varepsilon A$?

T4.2.1 A cross section is an area

Shoot one small projectile through a thin slab of material of thickness $\mathrm{d}x$ containing n target particles per unit volume. Suppose each target behaves like a small disc of area σ facing the projectile: if the projectile's path crosses the disc, the reaction happens; otherwise nothing happens. *What is the probability of a reaction?* Look at a patch of the slab of area S as seen by the projectile. It holds $nS\,\mathrm{d}x$ targets, whose discs cover the area $nS\,\mathrm{d}x\,\sigma$, and for a thin enough slab the discs do not overlap. The projectile arrives at a random point of the patch, so the probability that it hits a disc is the covered fraction,

$$\mathrm{P}(\text{reaction in } \mathrm{d}x) = \frac{nS\,\mathrm{d}x\,\sigma}{S} = n\sigma\,\mathrm{d}x. \tag{T4.7}$$

This is the operational meaning of the **cross section** σ: the effective area a target presents for one particular reaction. It has nothing to do with the geometric size of the proton. A process with a small quantum-mechanical amplitude has a small σ, even though the protons are the same. The unit is the **barn**, $1\,\mathrm{b} = 10^{-24}\,\mathrm{cm}^2$, with $1\,\mathrm{mb} = 10^{-3}\,\mathrm{b}$, $1\,\mathrm{pb} = 10^{-12}\,\mathrm{b}$ and $1\,\mathrm{fb} = 10^{-15}\,\mathrm{b}$. In natural units, $(\hbar c)^2 = 0.3894\,\mathrm{GeV}^2\,\mathrm{mb}$ converts a cross section computed in GeV^{-2} into millibarn [119, §49].

Two cross sections set the scale of the whole chapter. At $\sqrt{s} = 13\,\mathrm{TeV}$, the probability that two crossing protons interact at all corresponds to the **inelastic cross section** $\sigma_{\mathrm{inel}} = 78.1 \pm 2.9\,\mathrm{mb}$ [120]. The probability that they make a Higgs boson by gluon fusion that then decays to two photons corresponds to $\sigma \cdot B(H \to \gamma\gamma) = 0.102\,\mathrm{pb}$, the number stored in the simulated signal files of Section T4.7 [121]. Their ratio is $0.102 \times 10^{-12}/78.1 \times 10^{-3} = 1.3 \times 10^{-12}$: one Higgs-to-diphoton decay for every 7.7×10^{11} collisions.

T4.2.2 Two bunches crossing: the luminosity

At a collider the targets are not at rest in a slab; they are the protons of the other bunch. *How many reactions happen when one bunch crosses another?* Describe each bunch by its number of protons, N_1 and N_2, and by how they are spread over the plane transverse to the beam: $\rho_1(x, y)$

and $\rho_2(x,y)$, each normalised to $\int \rho \,\mathrm{d}x\,\mathrm{d}y = 1$. The length of the bunches along the beam does not matter, because every proton of one bunch passes through the whole of the other.

Take one proton of bunch 1 at transverse position (x,y). As it crosses bunch 2 it meets an areal density of targets $N_2\rho_2(x,y)$ (protons per unit transverse area). Bunch 2 is then a thin slab whose targets cover the fraction $\sigma N_2\rho_2(x,y)$ of the area, so by the same argument as (T4.7) this proton reacts with probability $\sigma N_2\rho_2(x,y)$. Bunch 1 has $N_1\rho_1(x,y)\,\mathrm{d}x\,\mathrm{d}y$ protons in a small patch around (x,y). Adding up over the patches, the expected number of reactions in one crossing is

$$\bar{n}_{\text{crossing}} = \sigma\, N_1 N_2 \int \rho_1(x,y)\,\rho_2(x,y)\,\mathrm{d}x\,\mathrm{d}y. \tag{T4.8}$$

Everything except σ is a property of the beams. With n_b pairs of bunches, each pair crossing f times per second (once per turn of the ring), the reaction rate is σ times the **instantaneous luminosity**

$$L_{\text{inst}} = n_b f\, N_1 N_2 \int \rho_1\rho_2 \,\mathrm{d}x\,\mathrm{d}y \qquad [\mathrm{cm}^{-2}\,\mathrm{s}^{-1}]. \tag{T4.9}$$

Gaussian bunches. The transverse profile of a bunch is close to Gaussian, $\rho(x,y) = g(x;\sigma_x)\,g(y;\sigma_y)$ with $g(x;s) = e^{-x^2/2s^2}/(\sqrt{2\pi}\,s)$. Let the two beams have the same sizes and let their centres be separated by Δ along x. The x part of the overlap integral is

$$\begin{aligned}\int_{-\infty}^{\infty} g(x;\sigma_x)\,g(x-\Delta;\sigma_x)\,\mathrm{d}x &\overset{(1)}{=} \frac{1}{2\pi\sigma_x^2}\int \exp\!\Big[-\frac{x^2+(x-\Delta)^2}{2\sigma_x^2}\Big]\mathrm{d}x \\ &\overset{(2)}{=} \frac{e^{-\Delta^2/4\sigma_x^2}}{2\pi\sigma_x^2}\int \exp\!\Big[-\frac{(x-\Delta/2)^2}{\sigma_x^2}\Big]\mathrm{d}x \overset{(3)}{=} \frac{e^{-\Delta^2/4\sigma_x^2}}{2\sqrt{\pi}\,\sigma_x}.\end{aligned} \tag{T4.10}$$

(1) Write out the two Gaussians and multiply them.
(2) Complete the square: $x^2+(x-\Delta)^2 = 2(x-\Delta/2)^2 + \Delta^2/2$, so the exponent splits into a part that depends on x and the constant $-\Delta^2/4\sigma_x^2$.
(3) The Gaussian integral $\int e^{-u^2/a^2}\mathrm{d}u = \sqrt{\pi}\,a$ with $a = \sigma_x$.

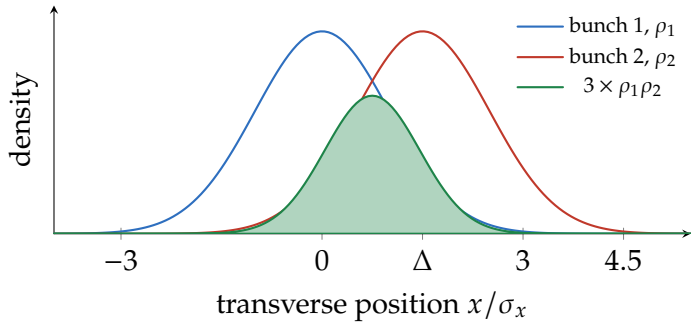


Figure T4.4: The overlap that sets the luminosity, in one transverse direction, for two bunches of equal width whose centres are $\Delta = 1.5\,\sigma_x$ apart. The shaded curve is the product of the two profiles (magnified three times); its area is the integral (T4.10). It is centred half-way between the bunches, and its area falls as $e^{-\Delta^2/4\sigma_x^2}$ when they are pulled apart, here to 57% of its head-on value.

The y part is the same with $\Delta = 0$, that is $1/(2\sqrt{\pi}\,\sigma_y)$. Multiplying the two parts and inserting into (T4.9),

$$L_{\text{inst}} = \frac{n_b f\, N_1 N_2}{4\pi\,\sigma_x\sigma_y}\, e^{-\Delta^2/4\sigma_x^2}\, F. \tag{T4.11}$$

At $\Delta = 0$ this is the formula of the PDG review [119, (31.3)], where a factor F of order one, written in here as well, corrects for the small crossing angle of the beams and for the finite

length of the bunches. The Gaussian dependence on the separation will be the basis of the luminosity calibration in Section T4.5.

Working definition T4.3 (cross section and luminosity)

The **cross section** σ of a process is the reaction rate it produces divided by the instantaneous luminosity; for a target at rest it is the effective area per target, (T4.7). The **instantaneous luminosity** L_{inst} [collision opportunities per unit area per second] is a property of the beams alone, (T4.11). The **integrated luminosity** is $L = \int L_{\text{inst}}\,\mathrm{d}t$, quoted in inverse femtobarn, $1\,\text{fb}^{-1} = 10^{39}\,\text{cm}^{-2}$. *Operationally:* the expected number of events of a process in a data set is $N = \sigma L$ [119, (31.2)].

The relation $N = \sigma L$ follows by integrating the rate $\sigma L_{\text{inst}}(t)$ over the running time. It is the reason cross sections are quoted in barns and luminosities in inverse barns: their product needs no conversion factor, as long as the prefixes are matched ($1\,\text{pb} \times 1\,\text{fb}^{-1} = 10^3$).

Example: the LHC design luminosity, and how many collisions share one crossing. The LHC was designed for $N_1 = N_2 = 1.15 \times 10^{11}$ protons per bunch, $n_b = 2808$ bunches per beam, a revolution frequency $f = 11\,245\,\text{Hz}$, beams of r.m.s. size $\sigma_x = \sigma_y = 16.7\,\mu\text{m}$ at the collision point, and $F = 0.836$ [122]. Then (T4.11) with $\Delta = 0$ gives

$$L_{\text{inst}} \overset{(1)}{=} \frac{2808 \times 11\,245\,\text{s}^{-1} \times (1.15 \times 10^{11})^2}{4\pi \times (16.7 \times 10^{-4}\,\text{cm})^2} \times 0.836$$
$$\overset{(2)}{=} 1.192 \times 10^{34}\,\text{cm}^{-2}\text{s}^{-1} \times 0.836 = 9.961 \times 10^{33}\,\text{cm}^{-2}\text{s}^{-1}, \tag{T4.12}$$

(1) Insert the design values; $16.7\,\mu\text{m} = 16.7 \times 10^{-4}\,\text{cm}$.
(2) Numerator $4.18 \times 10^{29}\,\text{s}^{-1}$, denominator $3.50 \times 10^{-5}\,\text{cm}^2$.

the famous $10^{34}\,\text{cm}^{-2}\text{s}^{-1}$. The rate of inelastic collisions is $\sigma_{\text{inel}} L_{\text{inst}} = 7.780 \times 10^8$ per second, spread over $n_b f = 3.16 \times 10^7$ bunch crossings per second, so each crossing contains on average

$$\bar{n}_{\text{pu}} = \frac{\sigma_{\text{inel}} L_{\text{inst}}}{n_b f} = 24.64 \tag{T4.13}$$

simultaneous collisions. One of them may be interesting; the others, called **pile-up**, add soft particles everywhere in the detector. The 8 TeV data of 2012 had about 20 such collisions per crossing [118, Sec. 1], and the PDG quotes up to 50 for recent runs [119, §31.4.1]. Pile-up is why the analyses ask for photons with little energy around them (Section T4.7) and why the efficiency of that requirement is itself a calibrated quantity (Section T4.5).

Example: how many Higgs bosons in the open data. The 2016 data released by ATLAS correspond to $L = 10\,\text{fb}^{-1}$ [123]. The expected number of Higgs bosons made by gluon fusion and decaying to two photons is

$$N = \sigma B\, L \overset{(1)}{=} 0.102\,\text{pb} \times 10\,\text{fb}^{-1} \overset{(2)}{=} 0.102 \times 10^3\,\text{fb} \times 10\,\text{fb}^{-1} = 1020. \tag{T4.14}$$

(1) The cross section times branching fraction stored in the simulation files, and the luminosity of the release.
(2) $1\,\text{pb} = 10^3\,\text{fb}$, so the femtobarns cancel.

In the same $10\,\text{fb}^{-1}$ the number of inelastic collisions is $\sigma_{\text{inel}} L = 78.1 \times 10^{12}\,\text{fb} \times 10\,\text{fb}^{-1} = 7.8 \times 10^{14}$. Almost none of them were recorded: the trigger kept 7 798 424 events with two energetic photon candidates in the released diphoton stream (Section T4.7).

T4.2.3 The events we keep: acceptance and efficiency

Not every Higgs boson produced becomes an event in our histogram. Its photons may fly too close to the beam, outside the calorimeter. They may hit the gap between barrel and end-cap. One of them may be too soft to pass the trigger, may fail the identification, or may have too much energy around it. The fraction of produced events that survive is split, by convention, into two factors [Cowan §11.1]. The **acceptance** A is the fraction that is geometrically and kinematically inside the region the analysis looks at (here: both photons at $|\eta| < 2.37$ outside the crack, with large enough p_T). The **efficiency** ε is the fraction of those that the detector and the selection actually record as two good photons. Cowan uses "efficiency" for the product; the split matters only because the two factors are calibrated in different ways. The expected number of selected events is then [Cowan (2.12)], [29, (15)]

$$\boxed{\nu = \sigma \, L \, \varepsilon \, A} \tag{T4.15}$$

Cranmer writes the same relation as "rate = flux × cross section × efficiency × acceptance", and calls the product $\sigma\varepsilon A$ an effective cross section [29, (16)].

Example: the acceptance of a diphoton selection, by simulation. We generate 200 000 Higgs decays with a short script (`01_mass_lineshape.py`), isotropic in the rest frame of the boson, and give each boson a random transverse momentum (exponential, mean 25 GeV) and a random rapidity (Gaussian, width 1.5) to mimic production by gluon fusion. Requiring both photons inside $|\eta| < 2.37$ and outside the crack keeps a fraction 0.631 of the decays; adding $p_T > 40$ and 30 GeV for the two photons leaves $A = 0.551 \pm 0.001$. Almost half of the Higgs bosons are lost before the detector has done anything wrong, simply because the photons go where the detector cannot measure them. (The production model is a toy; the number is meant to show the size of the effect, not to replace the full simulation used below.)

Example: the cut flow of the open-data signal. For the real simulation of gluon fusion in the 2016 detector, the script `02_open_data.py` applies a selection modelled on the 2012 analysis and follows the expected number of events step by step; Figure T4.5 shows the result next to the number of recorded data events that survive the same steps. The files contain only events that passed a loose two-photon preselection, which already removes almost half of the produced signal. Requiring both photons to pass the *tight identification* [a shower shape like that of one photon, not of two photons from a neutral pion] removes another fifth. The acceptance cuts in η cost almost nothing more, because the preselection had already applied them; the p_T cuts and the isolation requirement [at most 4 GeV of energy in a cone around each photon] cost a few per cent each. In the end $\varepsilon A = 0.3872$ and $\nu = 395$ selected signal events are expected. The data lose far more at the same steps (from 7 798 424 to 151 767 events): most recorded photon pairs are fakes from jets, and identification and isolation are designed to remove exactly those.

T4.2.4 Why the number of selected events is a Poisson count

The claim is plain: *the number of selected signal events in a data set is a Poisson random variable with mean* $\nu = \sigma L \varepsilon A$. The everyday question behind it is "if the LHC had delivered the same data set again, how many would we have got?" That question hides an assumption, namely what is held fixed when the data set is "delivered again". The answer is: the integrated luminosity. Not the running time, which depends on how the beams behaved, and not the total number of recorded events. With that fixed, we can derive the Poisson law from three physical inputs, as we derived it for photons from a star in Section 2.5.1.

The inputs (all physical).

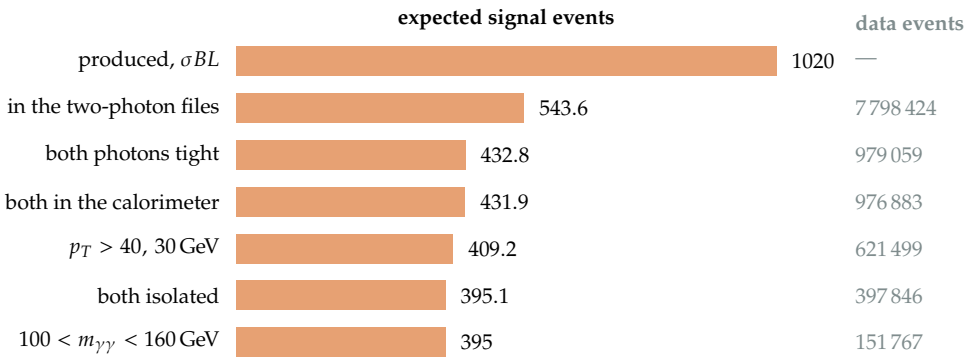


Figure T4.5: Where the Higgs bosons go. Bars: the expected number of gluon-fusion $H \to \gamma\gamma$ events in $10\,\text{fb}^{-1}$ after each step of the selection, from the weighted simulation. Grey column: the number of recorded data events left after the same steps. Signal survives the identification and isolation steps far better than the data, which are dominated by jets faking photons; the final ratio of the two columns is what the mass histogram has to overcome.

(C1) *Crossings are independent.* What happens in one bunch crossing does not change the probabilities in another.

(C2) *The process is rare in each crossing.* In crossing k the probability of producing the process is $\sigma\,\ell_k$, where $\ell_k = N_1 N_2 \int \rho_1 \rho_2$ is that crossing's luminosity, (T4.8); it is tiny, a few times 10^{-11} for the Higgs signal, and the chance of two such events in one crossing is negligible.

(C3) *Each produced event is kept independently.* It is selected with probability εA, independently of everything else.

The luminosity of a crossing may vary freely from one crossing to the next, as it does during a fill when the beams lose protons. Nothing is assumed about it.

What is the probability of the total count? The count in crossing k is 1 with probability $q_k = \sigma \ell_k \varepsilon A$ and 0 otherwise (inputs C2, C3: produced and kept). By C1 the total is a sum of independent Bernoulli variables with different probabilities. Its probability generating function $G(z) = \mathbb{E}\, z^N$ is the product of the individual ones, $1 - q_k + q_k z$, so

$$\ln G(z) \overset{(1)}{=} \sum_k \ln\bigl[1 + q_k(z-1)\bigr] \overset{(2)}{=} \sum_k q_k(z-1) + O\Bigl(\sum_k q_k^2\Bigr)$$
$$\overset{(3)}{=} (z-1)\,\sigma\varepsilon A \sum_k \ell_k \overset{(4)}{=} (z-1)\,\sigma L \varepsilon A = (z-1)\,\nu. \tag{T4.16}$$

(1) Independent factors: the generating function of a sum is the product of the generating functions, and the logarithm of a product is a sum.
(2) $\ln(1+u) = u + O(u^2)$, with $u = q_k(z-1)$ small because every q_k is tiny.
(3) Insert $q_k = \sigma \ell_k \varepsilon A$; the neglected term is at most $(\max_k q_k)\,\nu$, some $10^{-11}\nu$.
(4) The integrated luminosity is the sum of the luminosities of all crossings, $L = \sum_k \ell_k$.

So $G(z) = e^{\nu(z-1)}$, which is the generating function of the Poisson distribution with mean ν (with $z = e^t$ it is (2.13)). The count is Poisson, $\mathrm{P}(N = n) = e^{-\nu}\nu^n/n!$, with $\nu = \sigma L \varepsilon A$. This is the Poisson limit of the binomial of Section 2.4, generalised to trials with unequal probabilities; and the factor εA is the thinning of a Poisson stream from Section 2.5.

The clock of a collider. In Section 2.5.1 a constant rate (input P2) was needed so that the mean count in a window depended only on its length. Here no constant rate is needed. The mean depends only on the sum of the crossing luminosities, so the integrated luminosity plays

the role that elapsed time played for the star: it is the collider's clock, and a fill in which the luminosity decays is just a clock that ticks unevenly.

A counterexample: change what is held fixed. Suppose instead the experiment had decided to stop after recording exactly M selected diphoton events, signal and background together. Then the number of signal events among them is no longer Poisson. Each of the M events is signal with probability $\nu_S/(\nu_S + \nu_B)$, and the count is binomial with M trials, bounded by M and with a smaller variance than a Poisson count of the same mean. The same physics gives a different distribution because the sampling rule changed. The general version of this, a Poisson total split into bins, is the multinomial-to-Poisson step of Section 4.15.1.

Key idea: Counts at a collider

With the integrated luminosity held fixed, the number of selected events of a process is Poisson with mean $\nu = \sigma L \varepsilon A$: cross section from theory, luminosity from the machine, acceptance and efficiency from the detector. Independent processes add their means, and the counts in different bins of a histogram are independent Poisson variables.

What would break it. Input C1 fails if the selection in one crossing depends on the previous one, for example through detector dead time after a large deposit; this is small and corrected for in the luminosity. C3 fails if a whole batch of events is lost together, for example a run with a faulty calorimeter section; such runs are removed from the data before the luminosity is computed, so that L refers only to the data actually used [123]. And ν itself is uncertain, because L, ε and A are measured: that is the subject of Section T4.5.

Where this is used. The expected count in bin i of a histogram, $\nu_i = \mu s_i + b_i$, is (T4.15) applied to signal and background separately and shared among the bins (Section T4.4); its Poisson probability is the building block of the search likelihood in Section 4.15.1. The number 395 is the starting point of the capstone in Chapter 13.

T4.3 Invariant mass: why a decaying particle makes a peak

A Higgs boson is made with some momentum that we do not control and do not measure: a sideways kick from the gluons that made it, and a motion along the beam inherited from the unequal momentum fractions of the two partons (Section T4.1.2). When it decays, its two photons share its energy in a way that depends on that motion and on the direction of the decay. In one event they may carry 40 and 90 GeV, in the next 300 and 35 GeV. A histogram of photon energies would show no trace of the Higgs boson. Yet one combination of the two photons' energies and momenta comes out the same in every decay, and that combination turns a rare particle into a narrow peak.

The questions in words.

1. Which combination of the photons' energies and momenta is the same in every decay, and why?

2. How is it computed from the recorded p_T, η and ϕ?

3. Why is the peak not infinitely sharp? What shape does the finite lifetime of the particle give it, what shape does the detector give it, and which of the two wins for the Z boson and for the Higgs boson?

T4.3.1 The invariant mass of a pair

The energy E and momentum $\boldsymbol{p}$ of a particle form a four-vector $p = (E, \boldsymbol{p})$. The combination $p^2 \equiv E^2 - |\boldsymbol{p}|^2$ is the same in every inertial frame, and equals the square of the particle's mass, $p^2 = m^2$ [119, §49.1]. More generally the product $p \cdot q = E_p E_q - \boldsymbol{p} \cdot \boldsymbol{q}$ of any two four-vectors is frame independent. That is the only fact of special relativity we need.

The event. A particle of mass M decays into two particles with four-momenta p_1 and p_2, and the detector measures both. *What do p_1 and p_2 know about M?* Energy and momentum are conserved in the decay, so the parent's four-momentum is $P = p_1 + p_2$, in every event and in every frame. Squaring,

$$M^2 \overset{(1)}{=} P^2 \overset{(2)}{=} (p_1 + p_2)^2 \overset{(3)}{=} p_1^2 + p_2^2 + 2\,p_1 \cdot p_2 \overset{(4)}{=} m_1^2 + m_2^2 + 2\big(E_1 E_2 - \boldsymbol{p}_1 \cdot \boldsymbol{p}_2\big). \tag{T4.17}$$

(1) The parent's own mass: $P^2 = M^2$ for the parent's four-momentum, in any frame.
(2) Four-momentum conservation in the decay.
(3) Expand the square of a sum; the product is symmetric, $p_1 \cdot p_2 = p_2 \cdot p_1$.
(4) $p_j^2 = m_j^2$ for each daughter, and the definition of the four-product.

The right-hand side is built from measured quantities only. Its square root is the **invariant mass** of the pair, m_{12}, and (T4.17) says that it equals the parent's mass in every decay, however fast and in whichever direction the parent was moving [119, (49.2)]. The parent's motion changes E_1, E_2 and the angle between the daughters, but only in the combination that leaves m_{12} unchanged.

Two photons. Photons are massless, $m_1 = m_2 = 0$ and $|\boldsymbol{p}_j| = E_j$, so with θ_{12} the opening angle between them,

$$m_{\gamma\gamma}^2 = 2E_1 E_2\,(1 - \cos\theta_{12}). \tag{T4.18}$$

In the variables the detector records. Insert the components (T4.6), $E_j = p_{Tj}\cosh\eta_j$ and $\boldsymbol{p}_j = (p_{Tj}\cos\phi_j,\ p_{Tj}\sin\phi_j,\ p_{Tj}\sinh\eta_j)$:

$$\begin{aligned} E_1E_2 - \boldsymbol{p}_1\cdot\boldsymbol{p}_2 &\overset{(1)}{=} p_{T1}p_{T2}\big[\cosh\eta_1\cosh\eta_2 - \cos\phi_1\cos\phi_2 - \sin\phi_1\sin\phi_2 - \sinh\eta_1\sinh\eta_2\big] \\ &\overset{(2)}{=} p_{T1}p_{T2}\big[\cosh(\eta_1-\eta_2) - \cos(\phi_1-\phi_2)\big], \\ m_{\gamma\gamma}^2 &\overset{(3)}{=} 2\,p_{T1}p_{T2}\big[\cosh\Delta\eta - \cos\Delta\phi\big]. \end{aligned} \tag{T4.19}$$

(1) Multiply the energies and take the scalar product of the momenta; every term carries $p_{T1}p_{T2}$.
(2) Two addition formulas: $\cosh a\cosh b - \sinh a\sinh b = \cosh(a-b)$ and $\cos a\cos b + \sin a\sin b = \cos(a-b)$.
(3) Insert into (T4.17) with $m_1 = m_2 = 0$; $\Delta\eta = \eta_1 - \eta_2$, $\Delta\phi = \phi_1 - \phi_2$.

This is the PDG's (49.45) for massless particles [119], and it is the formula every diphoton analysis in the book uses. Only the differences $\Delta\eta$ and $\Delta\phi$ appear: a rotation about the beam adds the same angle to both ϕ_j, and a boost along the beam adds the same amount to both η_j (Section T4.1.2), so neither can change the mass.

Working definition T4.4 (invariant mass)

The **invariant mass** of a set of particles is $m = \sqrt{(\sum_j E_j)^2 - |\sum_j \boldsymbol{p}_j|^2}$, the mass of a single particle that would have their total energy and momentum [119, (49.26)]. *Operationally*, for two photons: $m_{\gamma\gamma} = \sqrt{2p_{T1}p_{T2}(\cosh\Delta\eta - \cos\Delta\phi)}$ from the recorded columns. If the

pair came from the decay of one particle, m is that particle's mass in every event, up to measurement errors.

Example: the Higgs boson at rest, and moving. At rest, the two photons share the energy equally, $E_1 = E_2 = M/2 = 62.5\,\text{GeV}$, and fly back to back, $\theta_{12} = \pi$, so (T4.18) gives $m^2 = 2 \times 62.5^2 \times 2 = 125^2\,\text{GeV}^2$. Now let the boson move with momentum $p_H = 100\,\text{GeV}$, so $E_H = \sqrt{125^2 + 100^2} = 160.1\,\text{GeV}$, and let it decay with the photons along its line of flight, one forwards and one backwards. In the rest frame each has energy $M/2$; boosting with $\gamma = E_H/M$ and $\beta = p_H/E_H$ gives the forward photon $\gamma\frac{M}{2}(1+\beta) = \frac{1}{2}(E_H + p_H) = 130.0\,\text{GeV}$ and the backward one $\frac{1}{2}(E_H - p_H) = 30.0\,\text{GeV}$. They are still back to back, so

$$m^2 \overset{(1)}{=} 2 \cdot \tfrac{1}{2}(E_H + p_H) \cdot \tfrac{1}{2}(E_H - p_H) \cdot 2 \overset{(2)}{=} E_H^2 - p_H^2 \overset{(3)}{=} M^2. \tag{T4.20}$$

(1) (T4.18) with $1 - \cos\pi = 2$.
(2) $(a+b)(a-b) = a^2 - b^2$.
(3) The boson's own mass-shell relation.

Energies of 130 and 30 GeV instead of 62.5 and 62.5, and the same mass. Figure T4.6 draws a decay at an angle, where the opening angle changes too.



Figure T4.6: The same decay in two frames, with arrow lengths proportional to momentum. In the rest frame of the Higgs boson the photons are back to back with 62.5 GeV each. In the lab, where the boson moves to the right, the forward photon gains energy, the backward one loses it, and the opening angle closes to $\theta_{12} = 114°$. Both pictures give $2E_1E_2(1 - \cos\theta_{12}) = 125^2\,\text{GeV}^2$.

Example: two hundred thousand decays in a computer. The script `01_mass_lineshape.py` makes the same check on 200 000 simulated decays with random directions and random boosts. The lines that matter are these:

Listing T4.1: Boost the rest-frame photons and recompute the mass from (p_T, η, ϕ) (`code/chT4/01_mass_lineshape.py`)

```
def boost(p, P, M):
    """Lorentz-boost four-vectors p (given in the rest frame of P) into the frame where P
    is measured."""
    gamma = P[:, 0] / M
    bvec = P[:, 1:] / P[:, :1]                                  # velocity of the
    parent
    b2 = np.sum(bvec**2, axis=1)
    bp = np.sum(bvec * p[:, 1:], axis=1)
    E = gamma * (p[:, 0] + bp)
    coef = np.where(b2 > 0, (gamma - 1) * bp / np.where(b2 > 0, b2, 1), 0.0) + gamma * p[:,
    0]
    return np.column_stack([E, p[:, 1:] + coef[:, None] * bvec])

```

```
60
61 p1, p2 = boost(p1_rest, PH, MH), boost(p2_rest, PH, MH)
62
63
64 def pt_eta_phi(p):
65     pt = np.hypot(p[:, 1], p[:, 2])
66     eta = np.arcsinh(p[:, 3] / pt)                                    # eta =
       asinh(pz/pT), massless
67     return pt, eta, np.arctan2(p[:, 2], p[:, 1])
68
69
70 pt1, eta1, phi1 = pt_eta_phi(p1)
71 pt2, eta2, phi2 = pt_eta_phi(p2)
72 m_lab = np.sqrt(2 * pt1 * pt2 * (np.cosh(eta1 - eta2) - np.cos(phi1 - phi2)))   # PDG
       (49.45), massless
```

The function `boost` applies the Lorentz transformation for a general velocity $\boldsymbol{\beta} = \boldsymbol{P}/E_P$ of the parent: $E' = \gamma(E + \boldsymbol{\beta}\cdot\boldsymbol{p})$ and $\boldsymbol{p}' = \boldsymbol{p} + \left[(\gamma-1)\boldsymbol{\beta}\cdot\boldsymbol{p}/\beta^2 + \gamma E\right]\boldsymbol{\beta}$, the vector form of [119, (49.1)]. Lines 64–67 convert each photon to the detector variables, with $\eta = \sinh^{-1}(p_z/p_T)$ from (T4.5), and line 72 is (T4.19). The photon energies spread from 36.4 to 660.5 GeV (5th to 95th percentile), while the largest deviation of $m_{\gamma\gamma}$ from 125 GeV in all the decays is 2.211×10^{-11} GeV, the rounding error of double precision. Panel (a) of Figure T4.7 shows the two histograms side by side.

T4.3.2 The natural width: a Breit–Wigner from a finite lifetime

The Breit–Wigner shape has been with us since Chapter 2. In Section 2.10.5 it was the line shape of the Z boson, with mass $M = 91.1876$ GeV and width $\Gamma = 2.4952$ GeV, and its lack of a mean defeated a running average. In Section 2.16.2 it was smeared with a Gaussian detector error of 2.0 GeV, and the tails of the resulting Voigt profile stayed those of the Breit–Wigner. In Section 3.3.4 the average of N draws from it was shown to be no better than one draw. In all three places the shape was taken as given, with the detector ignored or modelled crudely. What is new here is where the shape comes from, what the width means physically, and how it competes with a realistic detector resolution.

The question. If the parent's mass is exactly M, (T4.17) gives a spike, not a peak. *What spreads the invariant mass of real decays?* A particle that decays does not have a sharp energy. It exists only for a time of order its lifetime τ, and a state that lasts a time τ can only have its energy defined to about $\hbar/\tau$. We can make this exact in three short steps, with $\hbar = c = 1$.

The decaying state as a damped oscillation. In its rest frame, a stable particle's quantum amplitude oscillates as e^{-iMt}. An unstable one survives to time t with probability $e^{-\Gamma t}$, with $\Gamma = 1/\tau$ [119, (49.14)]. The probability is the squared modulus of the amplitude, so the amplitude must also decay, at half the rate:

$$\psi(t) = e^{-iMt}\, e^{-\Gamma t/2} \qquad (t \geq 0), \qquad |\psi(t)|^2 = e^{-\Gamma t}. \tag{T4.21}$$

Which energies does it contain? The amplitude for energy E is the Fourier component of ψ at E:

$$\tilde{\psi}(E) \overset{(1)}{=} \int_0^\infty e^{iEt}\,\psi(t)\,\mathrm{d}t \overset{(2)}{=} \int_0^\infty e^{[\,i(E-M)-\Gamma/2\,]t}\,\mathrm{d}t \overset{(3)}{=} \frac{1}{\Gamma/2 - i(E-M)},$$
$$|\tilde{\psi}(E)|^2 \overset{(4)}{=} \frac{1}{(E-M)^2 + \Gamma^2/4}. \tag{T4.22}$$

(1) Fourier analysis of a signal that starts at $t = 0$: the weight of the frequency E in $\psi(t)$.
(2) Insert (T4.21) and combine the exponentials.

(3) $\int_0^\infty e^{-at}\,\mathrm{d}t = 1/a$ for $\mathrm{Re}\,a > 0$; here $a = \Gamma/2 - i(E - M)$, whose real part $\Gamma/2$ is positive, so the integrand vanishes at infinity.
(4) The squared modulus of $1/(x - iy)$ is $1/(x^2 + y^2)$.

Normalise it. With $u = 2(E - M)/\Gamma$, $\int \mathrm{d}E/[(E - M)^2 + \Gamma^2/4] = (2/\Gamma)\int \mathrm{d}u/(1 + u^2) = 2\pi/\Gamma$. Dividing by this,

$$p(E \mid M, \Gamma) = \frac{1}{\pi}\,\frac{\Gamma/2}{(E - M)^2 + \Gamma^2/4}, \tag{T4.23}$$

which is exactly the Breit–Wigner (2.58) of Chapter 2, and the bracket of the PDG's resonance cross section [119, (51.1)]. The density falls to half its peak value where $(E - M)^2 = \Gamma^2/4$, so the full width at half maximum is Γ: the **width** of a particle is its decay rate, $\Gamma = \hbar/\tau$. The invariant mass of the decay products is the energy of the decaying state in its rest frame, so it follows (T4.23).

Relativistic form. The PDG writes the amplitude of a resonance near its mass as $1/(m^2 - M^2 + iM\Gamma)$ [119, (49.10)]. Near the peak, $m^2 - M^2 = (m - M)(m + M) \approx 2M(m - M)$, so its squared modulus is $1/[4M^2(m - M)^2 + M^2\Gamma^2] = (1/4M^2)/[(m - M)^2 + \Gamma^2/4]$: the same Breit–Wigner, up to a constant. The two forms differ only far in the tails, where (as Section 2.16 warned) the models break down anyway.

Example: lifetimes from widths. With $\hbar = 6.582 \times 10^{-25}\,\mathrm{GeV\,s}$, the Z boson's width $\Gamma_Z = 2.4952\,\mathrm{GeV}$ means $\tau_Z = 6.582 \times 10^{-25}/2.4952 = 2.64 \times 10^{-25}\,\mathrm{s}$. The Standard Model predicts a width of a few MeV for a Higgs boson of 125 GeV [124, Sec. 4]; with $\Gamma_H = 4.1\,\mathrm{MeV}$, $\tau_H = 6.582 \times 10^{-25}/4.1 \times 10^{-3} = 1.6 \times 10^{-22}\,\mathrm{s}$. The Higgs boson lives six hundred times longer than the Z, and its natural line is six hundred times narrower.

T4.3.3 The detector's blur: resolution, Voigt and Crystal Ball

The question. The calorimeter does not measure a photon's energy exactly. *How large is the error, and how does it spread the invariant mass?*

The energy error. A photon makes an electromagnetic shower, and the calorimeter samples the number of charged particles in it. That number grows in proportion to the energy, and it fluctuates like a Poisson count, so its relative fluctuation falls like $1/\sqrt{E}$. On top of this, imperfect calibration from cell to cell adds a relative error that does not shrink with energy. The two are independent and add in quadrature,

$$\frac{\sigma_E}{E} = \frac{a}{\sqrt{E/\mathrm{GeV}}} \oplus c, \qquad x \oplus y \equiv \sqrt{x^2 + y^2}, \tag{T4.24}$$

with the design values $a \approx 10\%$ and $c \approx 0.7\%$ for the ATLAS electromagnetic calorimeter [117].

From photon energies to the mass. The photon directions are measured well enough that the error on the opening angle contributes negligibly [118, Sec. 5.2]; we keep only the energy errors. Take the logarithm of (T4.18) and vary E_1 and E_2:

$$\begin{aligned} \ln m_{\gamma\gamma} &\overset{(1)}{=} \tfrac{1}{2}\ln E_1 + \tfrac{1}{2}\ln E_2 + \tfrac{1}{2}\ln\bigl[2(1 - \cos\theta_{12})\bigr], \\ \frac{\delta m}{m} &\overset{(2)}{=} \frac{1}{2}\,\frac{\delta E_1}{E_1} + \frac{1}{2}\,\frac{\delta E_2}{E_2}, \qquad \frac{\sigma_m}{m} \overset{(3)}{=} \frac{1}{2}\sqrt{\left(\frac{\sigma_{E_1}}{E_1}\right)^2 + \left(\frac{\sigma_{E_2}}{E_2}\right)^2}. \end{aligned} \tag{T4.25}$$

(1) $m^2 = 2E_1E_2(1 - \cos\theta_{12})$; the logarithm turns the product into a sum, and the square root gives the factors $\frac{1}{2}$.
(2) Differentiate: $\delta(\ln x) = \delta x/x$, with the angle held fixed.

(3) The two energy errors are independent, so the variances of the two terms add (Section 3.6).

For two equal photons of 62.5 GeV, (T4.24) gives $\sigma_E/E = \sqrt{0.10^2/62.5 + 0.007^2} = 1.45\%$, and (T4.25) gives $\sigma_m/m = 1.45\%/\sqrt{2} = 1.02\%$, about 1.3 GeV at 125 GeV. In the boosted toy decays of the previous subsection the photons are usually more energetic and so better measured; smearing each energy with (T4.24) gives a mass spread of 1.170 GeV, and averaging (T4.25) over the same events predicts 1.169 GeV. Panel (b) of Figure T4.7 shows the smeared peak and the Gaussian of that width.

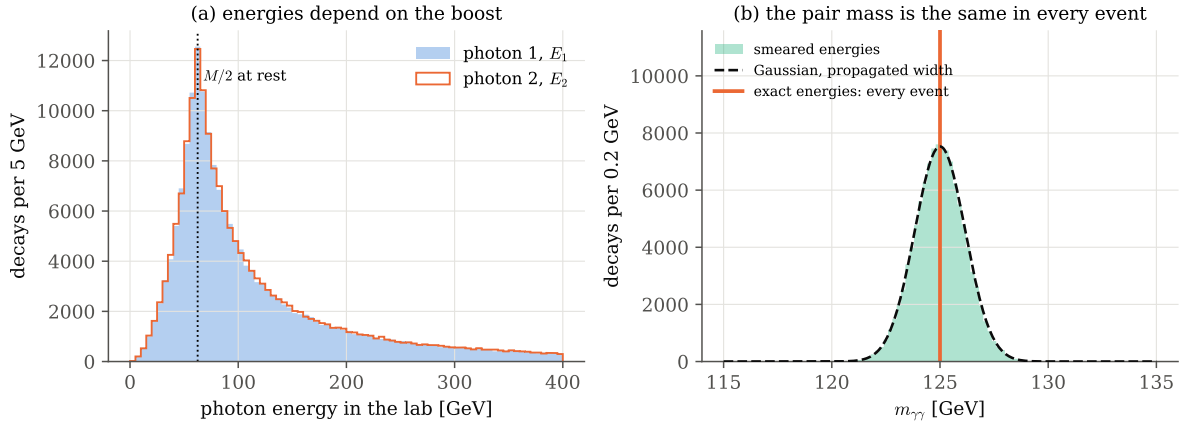


Figure T4.7: (a) Photon energies in 200 000 simulated Higgs decays with random boosts: anywhere from a few tens to several hundred GeV, nowhere near the 62.5 GeV of a decay at rest (dotted). (b) The invariant mass of the same pairs. With exact energies every event sits at 125 GeV (the vertical line). With energies smeared by the calorimeter resolution the pairs form a peak whose width is predicted by the error propagation (T4.25) (dashed).

Natural width and resolution together. The measured mass is the true mass of the decaying state, distributed as (T4.23), plus an independent measurement error, roughly Gaussian with width σ_m. The density of a sum of independent variables is the convolution of their densities, so the observed peak is the Voigt profile of (2.120). *Which of the two shapes wins?* It depends on the ratio Γ/σ_m.

- *The Z boson.* $\Gamma_Z = 2.50$ GeV and a dielectron mass resolution of about 1.5 GeV (CMS quotes 1.56 GeV in its barrel [124, Sec. 3]) are comparable. The full widths at half maximum are 2.50 GeV for the Breit–Wigner alone, 3.53 GeV for the Gaussian alone (2.355 σ), and 5.05 GeV for their convolution. Both effects matter, and a fit to the Z peak can separate them because their tails differ.
- *The Higgs boson.* $\Gamma_H \approx 4$ MeV against $\sigma_m \approx 1.6$ GeV. The Voigt and the Gaussian of the same σ differ by at most $0.0 \times 100\%$ of the peak height, with half-maximum widths 3.770 and 3.768 GeV. The width of the observed Higgs peak is entirely the detector's [124, Sec. 4]; the particle's own width cannot be read off it.

Key idea: Peak position and peak width

The position of a resonance peak in an invariant-mass histogram measures the particle's mass, because $m_{12}^2 = (p_1 + p_2)^2 = M^2$ in every decay. Its width combines the natural width $\Gamma = \hbar/\tau$ (Breit–Wigner) and the detector resolution σ_m (Gaussian core), and is dominated by whichever is larger; for the Higgs boson in two photons it is the resolution, about 1.6 GeV, against a natural width of 4 MeV.

Why the Gaussian is not enough: the Crystal Ball. A Gaussian error assumes many small, symmetric contributions. Some photons, however, lose a sizeable part of their energy before or outside the calorimeter: in the material in front of it, by converting, or by leaking out of the back or the sides of the shower. These losses only go one way, so the measured mass acquires a tail on the low side. ATLAS modelled its signal with a **Crystal Ball** function, a Gaussian core joined to a power-law tail, with $\sigma_{\mathrm{CB}} = 1.6\,\mathrm{GeV}$ and a full width at half maximum of 3.9 GeV for the inclusive sample [118, Sec. 5.4]. In the standardised variable $t = (m - \bar{m})/\sigma$ it is

$$f(t) \propto \begin{cases} e^{-t^2/2}, & t > -\alpha, \\ A\,(B - t)^{-n}, & t \le -\alpha, \end{cases} \tag{T4.26}$$

where $\alpha > 0$ says how many standard deviations below the peak the tail begins, and n how slowly it falls. *What fixes A and B?* We require the two pieces to join smoothly at $t = -\alpha$, with the same value and the same slope. The Gaussian piece has value $e^{-\alpha^2/2}$ and slope $-t\,e^{-t^2/2} = \alpha e^{-\alpha^2/2}$ there; the tail piece has value $A(B+\alpha)^{-n}$ and slope $nA(B-t)^{-n-1} = nA(B+\alpha)^{-n-1}$. So

$$\begin{aligned} A(B+\alpha)^{-n} &\overset{(1)}{=} e^{-\alpha^2/2}, & \frac{n}{B+\alpha}\,e^{-\alpha^2/2} &\overset{(2)}{=} \alpha\,e^{-\alpha^2/2} \\ \Longrightarrow \quad B &\overset{(3)}{=} \frac{n}{\alpha} - \alpha, & A &\overset{(4)}{=} \left(\frac{n}{\alpha}\right)^{n} e^{-\alpha^2/2}. \end{aligned} \tag{T4.27}$$

(1) Equal values at $t = -\alpha$.
(2) Equal slopes; the tail slope $nA(B+\alpha)^{-n-1}$ equals $\frac{n}{B+\alpha}A(B+\alpha)^{-n}$, and the first condition replaces $A(B+\alpha)^{-n}$.
(3) Solve the second condition for B.
(4) Insert $B + \alpha = n/\alpha$ into the first condition.

For $\alpha = 1.5$ and $n = 5$, $B = 1.833$ and $A = 133.6$. The tail then carries 10.4% of the probability, compared with 6.7% for the same region of a Gaussian, while the core, and so the full width at half maximum ($2.355\,\sigma$), is unchanged. Panel (c) of Figure T4.8 shows the difference on a logarithmic scale; it is invisible on a linear one, and yet a few per cent of the signal sits in the tail, where a narrow mass window would lose it.

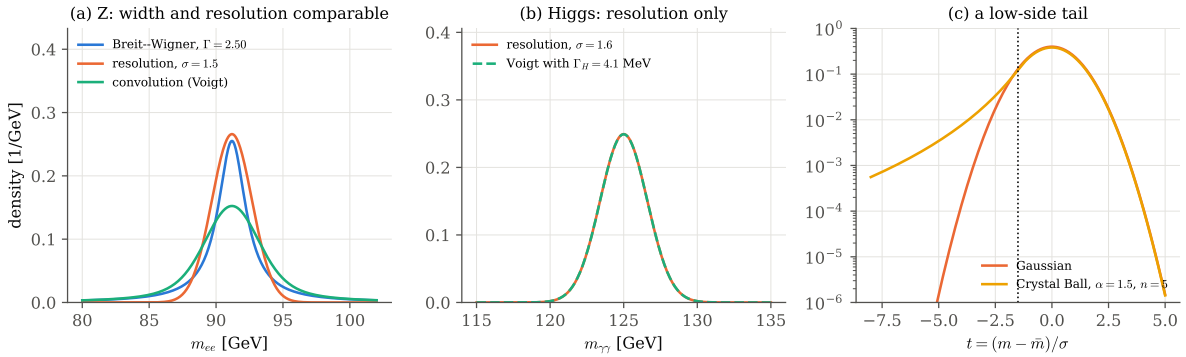


Figure T4.8: Line shapes. (a) For the Z boson, the natural Breit–Wigner width and a 1.5 GeV Gaussian resolution are comparable, and the observed peak is their convolution, wider than either. (b) For the Higgs boson the natural width is four hundred times smaller than the resolution; the convolution (dashed) cannot be told apart from the Gaussian. (c) A Crystal Ball function and a Gaussian with the same core, on a logarithmic scale: below $t = -\alpha$ (dotted) the Crystal Ball falls as a power law, modelling photons that lost energy before reaching the calorimeter.

Where this is used. The signal templates of Section T4.4 and of the capstone (Chapter 13) are peaks of exactly this kind, a Gaussian or Crystal Ball core at the hypothesised mass whose

width is set by (T4.25). The simulated signal of the open data (Section T4.7) has a core of 2.40 GeV and a visible low-side tail. The idea that the detector replaces a true value y by a measured value x with a known probability density, here a Gaussian or a Crystal Ball centred on y, is the *resolution function* of the unfolding problem in Section 6.14, where the question is reversed: given the blurred histogram, what was the true one?

T4.4 Signal and background: where the templates come from

The histogram of the diphoton mass $m_{\gamma\gamma}$ from Section T4.3 has, say, sixty bins of 1 GeV between 100 and 160 GeV. The statistics of Chapter 4 needs, for every bin i, the number of events ν_i we *expect* there. Two kinds of events fill the bins: Higgs decays (the signal) and everything else that produces two photon candidates (the background). Section T4.2 gave the total number of each kind, $\nu = \sigma L \varepsilon A$; Section T4.3 said roughly where the signal lands, in a narrow peak. What is still missing is the full predicted *shape* of each kind, bin by bin, and an honest normalisation of the background. Nobody can write those down from a formula: the shape of a detector's response to millions of different events is not a closed-form function.

The smallest example: two bins. Take only two bins. Bin 1 is a 10 GeV window around 125 GeV, bin 2 is everything else between 100 and 160 GeV. Write the expected counts as

$$\nu_1 = S\, f_1 + B\, g_1, \qquad \nu_2 = S\, f_2 + B\, g_2, \tag{T4.28}$$

where S and B are the expected total numbers of signal and background events, and f_j and g_j are the fractions of each that fall in bin j (so $f_1 + f_2 = 1$ and $g_1 + g_2 = 1$). There are two kinds of dial. The *normalisations* S and B say how many events of each kind there are; the *shapes* f_j and g_j say how they are shared among the bins. For the open data of Section T4.7 the numbers are roughly $S \approx 395$, $f_1 \approx 0.94$, $B \approx 151\,767$ and $g_1 \approx 0.17$: about 371 signal events and 26 254 background events in the window.

A naive choice: simulate everything. Suppose we took both S and B from the Standard Model prediction and a simulation of the detector. The background in the window then carries the uncertainty of the background prediction. Most of the background is genuine photon pairs, but about a quarter is jets that the detector mistook for photons (Section T4.4.3), and how often a jet fakes a photon depends on details of jet formation and of the calorimeter response that simulations reproduce far less precisely than the data count events [29, Sec. 4.2]. Even a 10% error on $B g_1$ is 2600 events, seven times the whole expected signal. The peak would drown in the uncertainty of the floor it sits on.

A better choice, and its price. The data themselves measure the background: away from 125 GeV the histogram contains almost no signal, so the bins on either side of the peak, the **sidebands**, show how many background events there are and how their number falls with mass. With 151 767 events the statistical error on the background level is a fraction of a per cent, not ten per cent. The price is an assumption: to carry the background from the sidebands into the window we must assume a smooth shape for it, and a wrong shape can create or hide a few events of fake signal (Section T4.4.3). The signal, on the other hand, has no sidebands; its shape and its expected size can only come from simulation, and its uncertainty is carried by the calibrations of Section T4.5.

So the division of labour is: signal from simulation, background from the data. Both templates can now be built, and with them the one number the search reports, the signal strength μ. In the language of Cranmer, these are the "simulation narrative" and the "data-driven narrative" of an analysis [29, Secs. 4.1–4.2].

T4.4.1 Simulated events: from the theory to a detector record

The question. Theory predicts the probability of producing two photons with given momenta. What we need is the probability that the *detector records* a photon pair with a given measured mass. How do we get from one to the other?

The answer is to simulate events one at a time, following each step a real event goes through, and then to histogram the results. A simulated event passes through four programs in turn (Figure T4.9) [29, Sec. 4.1], [118, Sec. 3].

1. *The hard process.* A **matrix-element generator** draws the momenta of the outgoing particles of the short-distance collision (for the signal, $gg \to H \to \gamma\gamma$) with the probability density predicted by perturbation theory, multiplied by the parton densities of the protons. ATLAS used Powheg for gluon fusion and vector-boson fusion in 2012 [118, Table 1].
2. *Parton shower and hadronisation.* Quarks and gluons radiate further gluons and then bind into hadrons; programs such as Pythia and Herwig model this with tuned phenomenological parameters. The remnants of the two protons, and the pile-up collisions of Section T4.2.2, are added here.
3. *Detector simulation.* Geant4 tracks every particle through a model of the detector material and records the energy it deposits in each channel [118, Sec. 3].
4. *Reconstruction.* The same programs that process real data turn the simulated deposits into photons, electrons, jets, each with its p_T, η and ϕ (Section T4.1).

Steps 1 and 2 are a model of nature, steps 3 and 4 a model of the experiment. The output has exactly the format of real data, plus the true momenta from step 1, so the same selection can be applied to both.

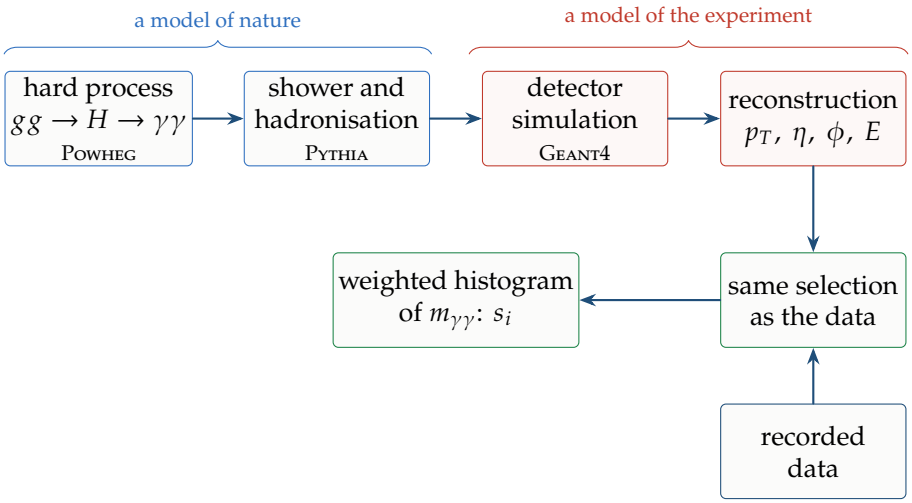


Figure T4.9: How a signal template is made. Each simulated event passes through the programs that model the collision (blue) and the programs that model the detector (red), and comes out in the same format as a recorded event. The selection used on the data is applied to it unchanged, and the surviving events, each with its weight, are histogrammed to give the expected signal count s_i in every bin.

From simulated events to a shape. If the detector were perfect, the distribution of an observable x (here $x = m_{\gamma\gamma}$) would be the normalised differential cross section, $f(x) = (1/\sigma_{\text{eff}})\,\mathrm{d}\sigma_{\text{eff}}/\mathrm{d}x$, with $\sigma_{\text{eff}} = \sigma\varepsilon A$ the effective cross section of (T4.15) [29, (17)]. With a real detector there is no formula, but there is a sample: the values $x_1, \ldots, x_N$ of the simulated events that pass the

selection are draws from the true $f(x)$, and a histogram of them estimates it [29, (18)]. This is the histogram estimate of a density of Section 6.13, with its bin-width trade-off; in a template it is usually harmless, because the bins are chosen narrower than the resolution σ_m of Section T4.3.3 and the simulated sample is large.

Example: what the generator list of 2012 says about the backgrounds. The ATLAS list of generators [118, Table 1] contains one program for the diphoton continuum ($q\bar{q}/gg \to \gamma\gamma$, Sherpa) and several for processes with a W, a Z or top quarks, which matter for the four-lepton and WW channels. It contains no simulation of a photon plus a jet faking a second photon. That background was taken from the data (Section T4.4.3). The list tells us which parts of a search trusted the simulation and which did not.

T4.4.2 Normalising a weighted simulation

The question. A simulated sample has a certain number of events, chosen by whoever generated it; it has nothing to do with the luminosity of the data. *How many real events does one simulated event stand for?*

The simplest case: unweighted events. Suppose the generator produced N_{gen} events of a process with cross section σ, each an independent draw from the produced events. The data contain on average σL produced events of this process. If $N_{\text{sel,i}}$ of the simulated events pass the selection and land in bin i, the fraction $N_{\text{sel,i}}/N_{\text{gen}}$ estimates the probability that a produced event does so (a Monte Carlo average of an indicator, as in the acceptance example of Section 6.11.4). Then $s_i = \sigma L\, N_{\text{sel,i}}/N_{\text{gen}}$: each simulated event counts for $\sigma L/N_{\text{gen}}$ real ones.

Weighted events. Modern generators do not produce equal events. A next-to-leading-order generator gives each event a weight w_e, sometimes negative, so that the *weighted* sample reproduces the predicted distribution. Later, small corrections that make the simulation agree with calibration data multiply each weight by **scale factors** (Section T4.7). The fraction of produced events in a region is then the fraction of the total weight in it, and

$$\begin{aligned} s_i &\overset{(1)}{=} \sigma L \ \mathrm{P}(\text{a produced event is selected and lands in bin } i) \\ &\overset{(2)}{=} \sigma L \,\frac{\sum_{e\in i} w_e}{\sum_{\text{all } e} w_e} \overset{(3)}{=} \sum_{e\in i}\Big(\frac{\sigma L}{W}\Big) w_e, \qquad W \equiv \sum_{\text{all } e} w_e. \end{aligned} \tag{T4.29}$$

(1) The expected number of produced events is σL (Section T4.2); thinning a Poisson count by a probability gives the expected count in bin i.

(2) The probability is estimated by the weighted fraction of simulated events, the self-normalised estimator of Section 6.12.4. The denominator runs over *all* generated events, before any selection; otherwise the efficiency would be lost.

(3) Bring the constant inside the sum: every selected event contributes $\sigma L\, w_e/W$ expected real events.

With all $w_e = 1$, $W = N_{\text{gen}}$ and (T4.29) is the unweighted rule. The capstone uses exactly this formula (Section 13.21.1).

How precise is a simulated template? The simulated sample is finite, so s_i has a statistical error of its own. Think of the simulated events landing in bin i as a Poisson number K of events with mean λ, each carrying an independent weight w. The sum $S_i = \sum_{k=1}^{K} w_k$ then has

$$\begin{aligned} \operatorname{Var}(S_i) &\overset{(1)}{=} \mathbb{E}\big[\operatorname{Var}(S_i \mid K)\big] + \operatorname{Var}\big[\mathbb{E}(S_i \mid K)\big] \overset{(2)}{=} \mathbb{E}[K]\operatorname{Var}(w) + \operatorname{Var}(K)\big(\mathbb{E}\, w\big)^2 \\ &\overset{(3)}{=} \lambda\big[\operatorname{Var}(w) + (\mathbb{E}\, w)^2\big] \overset{(4)}{=} \lambda\ \mathbb{E}[w^2] \ \approx\ \sum_{e\in i} w_e^2. \end{aligned} \tag{T4.30}$$

(1) The law of total variance (1.52), conditioning on the number of events K.
(2) Given K, the sum of K independent weights has variance $K\,\mathrm{Var}(w)$ and mean $K\,\mathbb{E}\,w$.
(3) K is Poisson, so $\mathbb{E}\,K = \mathrm{Var}\,K = \lambda$.
(4) $\mathrm{Var}(w) + (\mathbb{E}\,w)^2 = \mathbb{E}[w^2]$; the expected sum of squared weights $\lambda\,\mathbb{E}[w^2]$ is estimated by the observed one.

So the Monte Carlo error of s_i is $(\sigma L/W)\sqrt{\sum_{e\in i} w_e^2}$. For unit weights it is $(\sigma L/N_{\mathrm{gen}})\sqrt{N_{\mathrm{sel,i}}}$, the familiar $1/\sqrt{N}$ of a count. Unequal weights make it larger, as if the sample contained only $N_{\mathrm{eff}} = \left(\sum w_e\right)^2/\sum w_e^2$ equal events.

Example: the expected Higgs signal in the open data. The gluon-fusion file of the open data (Section T4.7) stores $\sigma B = 0.102\,\mathrm{pb}$ and the sum of the weights of all generated events, $W = 55\,922\,616$. It contains 1 054 711 events, those that survived a two-photon preselection. With $L = 10\,\mathrm{fb}^{-1} = 10\,000\,\mathrm{pb}^{-1}$ each unit of weight stands for $\sigma BL/W = 0.102 \times 10\,000/55\,922\,616 = 1.82\times10^{-5}$ real events. The selection of Section T4.2.3 keeps a weighted fraction $\varepsilon A = 0.3872$, so

$$\nu_S = \sigma B\, L\, \varepsilon A = 1020 \times 0.3872 = 395$$

selected signal events between 100 and 160 GeV, with a Monte Carlo error from (T4.30) of 0.48 events. That error is a tenth of a per cent: with a million simulated events, the template's own statistical error is negligible next to the Poisson fluctuation of the data in the same region (about 160 events, Section T4.4.3). For small simulated samples it is not negligible, and HistFactory gives each bin a nuisance parameter for it, constrained by a Poisson term whose effective count is $(s_i/\delta s_i)^2$ [125, (9)].[1]

Example: several processes in one template. The signal has four production modes (gluon fusion, vector-boson fusion, and production with a W or a Z; Section T4.6), each simulated in its own file with its own σ and W. Processes that do not interfere quantum mechanically add incoherently, so their expected counts add bin by bin, $s_i = \sum_{\mathrm{samples}} s_i^{(k)}$, and the normalised shape is the mixture $f(x) = \sum_k \nu_k f_k(x)/\sum_k \nu_k$ [29, (20)–(21)], [Cowan §6.9]. Processes that *do* interfere must be simulated together as one sample. The 2012 analysis is an example: gluon fusion to a Higgs boson and gluon fusion directly to two photons lead to the same final state, their amplitudes interfere destructively, and ATLAS lowered the gluon-fusion rate by 2–5% to account for it [118, Sec. 5.4].

T4.4.3 The background from the data: sidebands

What is in the background? Three kinds of event give two photon candidates. Real photon pairs from quark–antiquark annihilation or gluon fusion through a quark loop (the *irreducible* background, $\gamma\gamma$). A photon and a jet in which one hadron, typically a neutral pion decaying to two nearly collinear photons, looks like a single photon (γj). Two such jets (jj). The last two are *reducible*: better identification and isolation reduce them. ATLAS measured the composition of its selected 2012 sample with data-driven techniques: 74% $\gamma\gamma$, 22% γj, 3% jj and 1% Drell–Yan electron pairs mistaken for photons [118, Sec. 5.2]. That decomposition was used only to study the background shape; the search itself fitted the background directly to the data.

The method. In every event category ATLAS fitted the $m_{\gamma\gamma}$ spectrum between 100 and 160 GeV with a smooth function whose parameters were free: a fourth-order Bernstein polynomial, the exponential of a second-order polynomial, or a single exponential, depending on the category

[1] The text of [125, Sec. 2.2.1] writes the effective count as $(\delta_b/\nu_b^{\mathrm{MC}})^2$, the inverse; the next sentence there, $\tau_b = (\nu_b^{\mathrm{MC}}/\delta_b)^2$, has it the right way round. For unit weights, $(s_i/\delta s_i)^2 = N_{\mathrm{sel,i}}$, as it should be.

[118, Sec. 5.5]. The function plays the role of the background template b_i, with its normalisation and shape parameters measured by the data themselves. Cranmer calls this the "effective model narrative": no physical model of the background, only a flexible enough curve [29, Secs. 4.2–4.3].

What could go wrong: the spurious signal. A smooth function that is not flexible enough can bend away from the true background shape near 125 GeV and leave a deficit or excess there that the fit then attributes to signal. ATLAS measured this directly. They fitted signal plus background models to very large simulated background samples, which contain no signal at all, and recorded the largest fitted "signal" over the mass range [118, Sec. 5.5]. A function was accepted only if this **spurious signal** was below 20% of the statistical error of the fitted signal; the remaining spurious signal, 0.2–4.6 events at 7 TeV and 0.3–6.8 events at 8 TeV depending on the category, was kept as a systematic uncertainty. The capstone repeats this test on the open data (Section 13.22.2).

Example: an exponential through the sidebands of the open data. As an illustration, the script `02_open_data.py` of Section T4.7 histograms the 151 767 selected diphoton masses of the 2016 open data in 1 GeV bins and, without a proper fit or uncertainties, draws a straight line through the logarithm of the counts in the bins outside 120–130 GeV. The slope is −0.0231 per GeV: the background falls by 2.3% for every GeV of mass. Extrapolated into the window, the exponential predicts $b = 26\,254$ background events between 120 and 130 GeV, where 26 457 events were recorded. The simulated Standard Model signal from Section T4.4.2, histogrammed with (T4.29), puts $s = 371.2$ expected events in the same window. Panel (a) of Figure T4.10 shows the three side by side.

The numbers give the problem its size. The signal is $s/b = 1.41\%$ of the background in the window, far too small to see by eye; the typical Poisson fluctuation of the background count is $\sqrt{b} = 162$ events; and the ratio $s/\sqrt{b} = 2.29$ is the rough number of such fluctuations the expected signal amounts to. Whether that is enough, and how much better the full histogram does than one window, is decided by the statistic q_0 and its Asimov approximation in Sections 4.16 and 4.16.3. The excess of recorded over predicted events, $26\,457 - 26\,254 = 203$, is about one such fluctuation.

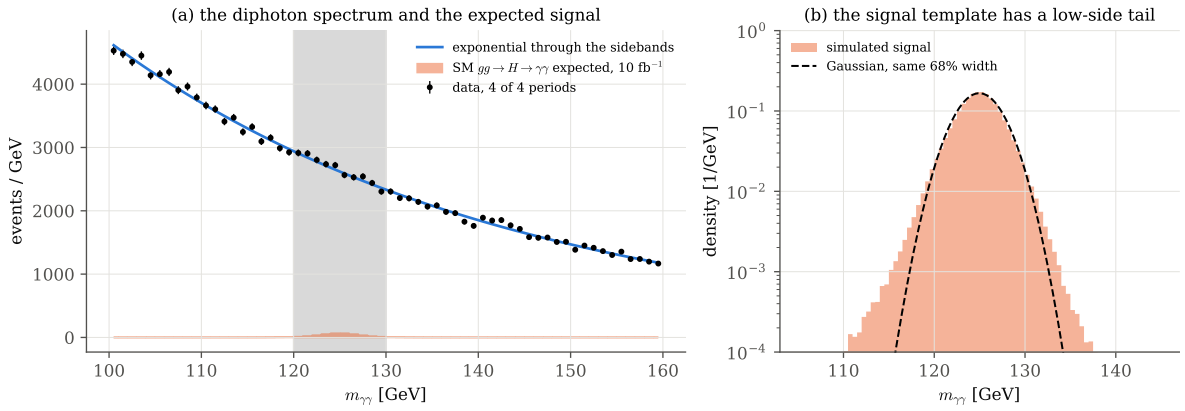


Figure T4.10: (a) The diphoton mass spectrum of the 2016 open data after the selection of Section T4.2.3, with an exponential drawn through the sidebands outside the grey window, and the expected Standard Model gluon-fusion signal (filled) at the bottom of the plot: a few dozen events per GeV under a background of thousands. (b) The simulated signal shape on a logarithmic scale against a Gaussian of the same central 68% width. The core is Gaussian; the low-mass side has the tail of photons that lost energy, the shape the Crystal Ball of (T4.26) was built for.

Example: the signal template's width and tail. Panel (b) of Figure T4.10 is the signal template of the open data. Its median is 125.0 GeV and half of its central 68% range is 2.40 GeV. Of the weighted signal, 93.1% lies within two such widths of the median (a Gaussian would have 95.4%), 0.83% lies more than three widths below and 0.54% more than three widths above (a Gaussian would have 0.13% on each side). This half-width is not the same quantity as the Crystal Ball core width $\sigma_{\mathrm{CB}} = 1.6$ GeV of the 2012 analysis: the central 68% range includes part of the tails, which a core width ignores. The asymmetry of the tails is the energy loss described in Section T4.3.3; Section 13.22.1 fits the shape properly.

T4.4.4 The signal strength

The question. The search must be able to say "there is no signal", "there is the predicted signal", and anything in between or beyond. *Which single number expresses that?*

For a fixed hypothesised mass m_H, the Standard Model fixes the cross section of every production mode, every branching fraction, and so the size and shape of the signal template. The searches of 2012 therefore did not fit the Higgs couplings themselves. They introduced one factor that scales the predicted signal [29, Sec. 4.1.2.1], [118, Sec. 7]:

$$\mu = \frac{\sigma\, B}{(\sigma\, B)_{\mathrm{SM}}}, \qquad \nu_i(\mu) = \mu\, s_i + b_i, \qquad s_i = (\sigma B)_{\mathrm{SM}}\, L\, \varepsilon A\, f_i. \tag{T4.31}$$

Here s_i is the Standard Model signal template in bin i (the expected count (T4.29), with the fraction f_i of the selected signal falling in that bin), b_i the background template, and μ the **signal strength**. The value $\mu = 0$ is the background-only hypothesis, the null hypothesis of the discovery test; $\mu = 1$ is the Standard Model Higgs boson; any other value is a signal of the predicted shape but a different size. The second relation is the template model (4.88) of Section 4.15, where its likelihood is written down.

Working definition T4.5 (signal strength)

The **signal strength** μ is the ratio of the true signal rate to the Standard Model prediction for the same mass hypothesis, $\mu = \sigma B/(\sigma B)_{\mathrm{SM}}$ [cross section times branching fraction into the final state that is searched]. *Operationally:* multiply the simulated signal template by μ, add the background template, and compare with the data bin by bin, $\nu_i = \mu s_i + b_i$.

What μ can and cannot tell apart. A diphoton search counts events of production times decay, so it measures only the product σB. A Higgs boson produced twice as often and decaying to photons half as often gives the same $\mu = 1$. Nor can it tell μ from anything else that multiplies the signal: a luminosity or a photon efficiency 10% higher than assumed gives exactly the same histogram as a μ 10% higher. Only an independent calibration separates the two, which is why every such factor enters the search as a nuisance parameter with its own calibration measurement (Section T4.5). Finally, one μ for all production modes is a convenience; a theory beyond the Standard Model would change the modes differently, and once a signal was seen ATLAS also fitted separate strengths $\mu_i = \sigma_i/\sigma_{i,\mathrm{SM}}$ for groups of modes [118, Sec. 9.3].

Example: what $\mu = 1.4$ would mean in the open data. ATLAS measured $\hat{\mu} = 1.4 \pm 0.3$ in 2012 (Section T4.6). If the true signal strength were 1.4, the 120–130 GeV window of the open data would hold $1.4 \times 371.2 = 520$ signal events on top of the background instead of 371, a difference of 150 events: less than one Poisson fluctuation of the background count (162). A single window of the open data cannot distinguish $\mu = 1$ from $\mu = 1.4$; the full histogram and the precise statistics of Section 4.15 are needed even to see the signal.

Key idea: Signal from simulation, background from the data

The expected count in bin i is $\nu_i = \mu s_i + b_i$. The signal template s_i is a weighted histogram of simulated events, normalised by $\sigma L w_e / W$ per event; the background template b_i is a smooth function whose parameters are measured by the sidebands of the data. The signal strength $\mu = \sigma B/(\sigma B)_{\text{SM}}$ scales the predicted signal: $\mu = 0$ is no signal, $\mu = 1$ the Standard Model.

Where this is used. The template model $\nu_i = \mu s_i + b_i$ and its Poisson likelihood are Sections 4.15 and 4.15.1; the scan of a peak template over the mass axis is Section 4.14.3. The capstone builds the signal template from the four simulated production modes with (T4.29) and fits the background shape to the sidebands (Sections 13.22 and 13.22.1). The response of a detector to a true value, which the simulation chain of Figure T4.9 encodes, is the response matrix of the unfolding problem in Section 6.14 [Cowan §11.1].

T4.5 Systematic uncertainties: nuisance parameters and their calibrations

Every ingredient of the expected count $\nu_i = \mu s_i + b_i$ of Section T4.4 was measured or computed by someone, and none exactly. The luminosity L came from a calibration of the beams, the photon efficiency from a study of electrons, the energy scale that places the peak from the Z boson, the cross section from a perturbative calculation truncated at some order. *How does each of these imperfections enter the prediction, and what keeps it under control?*

The smallest example: one bin, one calibration. Take a single counting window with expected count $\nu = \mu\,\kappa\, s + b$. Here s is the nominal Standard Model signal in the window, b the background, and κ the ratio of the true value of luminosity times efficiency to the value assumed when s was computed; $\kappa = 1$ if the calibrations were perfect. The two dials μ and κ appear only as their product. Any data in this window are explained equally well by $(\mu, \kappa) = (1.4, 1)$ and by $(1, 1.4)$.

The naive choice. Fix $\kappa = 1$ and ignore it. If the true luminosity times efficiency is in fact 11% lower, the fitted μ comes out 11% low, and no amount of data reveals it: more events only make the wrong answer more precise.

The better choice, and its price. Measure κ separately, in a calibration that does not involve the Higgs signal at all, and carry its result and its error into the search. Then $\mu\kappa$ is measured by the window and κ by the calibration, and μ follows. The price is a floor on the precision. With infinite data in the window the product $\mu\kappa$ is known exactly, and the relative error on μ is still the relative error of the calibration. For the diphoton search of 2012, the photon efficiency alone was known to 11% at 8 TeV [118, Sec. 5.6], so μ could not have been measured better than about 11% however many collisions were recorded.

The lesson: a parameter the data cannot separate from the signal must be pinned down by its own measurement, and that measurement's precision becomes part of the result. Each such parameter is a **nuisance parameter** [a parameter the prediction depends on and that we must fit, but do not want to report], and the measurement that pins it down is an **auxiliary** or **calibration measurement** [29, Sec. 2.2], [119, §40.2.2.2]. A diphoton search has about a dozen such parameters, each with a size, an effect on the templates and a calibration behind it.

T4.5.1 Sources and their effects

Cranmer makes one distinction the bookkeeping depends on: the *source* of an uncertainty (for example, the calorimeter's energy calibration) is not the same thing as its *effect* on a particular template (the signal peak moves; the background shape hardly changes) [29, Sec. 4.1.4]. One source gets one nuisance parameter, even if it affects many templates, channels and data sets. ATLAS treated the luminosity as fully correlated across the diphoton, four-lepton and WW channels, and the photon and electron energy scales as five shared parameters [118, Sec. 8]. The effects come in three kinds, which HistFactory calls by the names in brackets [125, Table 1]:

- *Normalisation* (`OverallSys`): the whole template is multiplied by a factor near one. Luminosity, efficiencies and cross sections act this way.
- *Shape* (`HistoSys`): the fractions f_i themselves change. The energy scale moves the peak; the resolution widens or narrows it.
- *Bin by bin* (`StatError`, `ShapeSys`): each bin gets its own factor, for example for the finite simulated sample of (T4.30).

Figure T4.11 draws the sources of the 2012 diphoton search with the effect each one has.

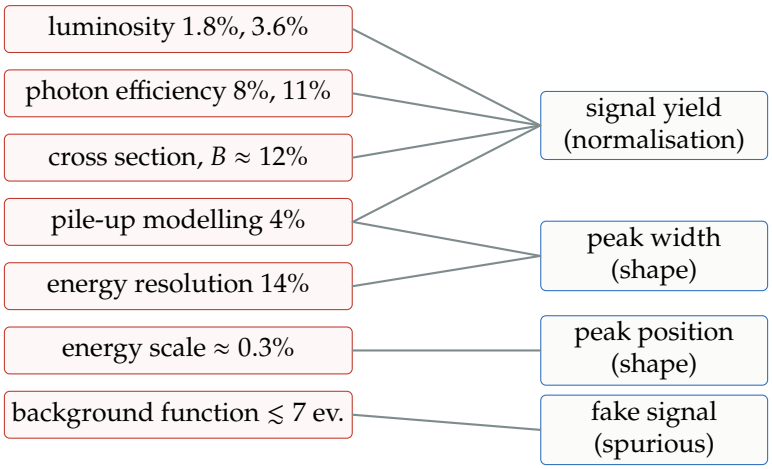


Figure T4.11: The main sources of systematic uncertainty in the 2012 diphoton search (left, with their sizes; two numbers are for the 7 and 8 TeV data) and the effect each has on the templates (right). Most sources only rescale the expected signal. The energy scale moves the peak, the resolution changes its width, and pile-up does both. The background function's uncertainty is a number of events of spurious signal per category.

Example: why the photon efficiency counts twice. Every selected event has two photons, and each passes the identification with probability ε_γ, so the yield is proportional to $\varepsilon_{\gamma 1}\varepsilon_{\gamma 2}$. A calibration error that moves both efficiencies in the same direction by $\delta\varepsilon_\gamma/\varepsilon_\gamma$ changes the yield by

$$\frac{\delta N}{N} \overset{(1)}{=} \frac{\delta\varepsilon_{\gamma 1}}{\varepsilon_{\gamma 1}} + \frac{\delta\varepsilon_{\gamma 2}}{\varepsilon_{\gamma 2}} \overset{(2)}{=} 2\,\frac{\delta\varepsilon_\gamma}{\varepsilon_\gamma}. \tag{T4.32}$$

(1) The logarithm of a product is the sum of the logarithms; differentiate.
(2) The same calibration applies to both photons, so the two errors are equal and add linearly, not in quadrature.

A per-photon uncertainty of 5.5% thus becomes the 11% on the yield quoted for the 8 TeV data.

Example: adding up the yield uncertainties. For the 8 TeV diphoton signal the experimental normalisation uncertainties are the photon efficiency (11%), pile-up (4%), luminosity (3.6%), trigger (1%) and isolation (0.5%) [118, Sec. 5.6]; the theory uncertainties on the gluon-fusion prediction are the QCD scale (+7/−8%, about 7.5%), the parton densities (8%) and the branching fraction (5%) [118, Sec. 8]. The sources are independent, so their variances add (Section 3.6):

$$\left(\frac{\delta\nu}{\nu}\right)_{\text{exp}} \overset{(1)}{=} \sqrt{11^2 + 4^2 + 3.6^2 + 1^2 + 0.5^2}\,\% = 12.3\%,$$
$$\left(\frac{\delta\nu}{\nu}\right)_{\text{th}} \overset{(1)}{=} \sqrt{7.5^2 + 8^2 + 5^2}\,\% = 12.1\%, \qquad \left(\frac{\delta\nu}{\nu}\right)_{\text{tot}} \overset{(2)}{=} \sqrt{12.3^2 + 12.1^2}\,\% = 17\%. \tag{T4.33}$$

(1) Independent relative errors on factors of a product add in quadrature.
(2) The experimental and theoretical groups are independent of each other as well.

The one large experimental term dominates its group: removing the luminosity entirely would lower 12.3% only to 11.8%. This is typical of quadrature sums and tells an analysis where improvement pays.

Example: the peak moves with the energy scale. Suppose every measured photon energy is too large by the same fraction δ: $E_j \to (1+\delta)E_j$. The angle between the photons is unchanged, so by (T4.18)

$$m'^2_{\gamma\gamma} \overset{(1)}{=} 2\,(1+\delta)E_1\,(1+\delta)E_2\,(1-\cos\theta_{12}) \overset{(2)}{=} (1+\delta)^2\, m^2_{\gamma\gamma}, \qquad \text{so} \qquad \frac{\delta m}{m} = \delta. \tag{T4.34}$$

(1) Insert the scaled energies into $m^2 = 2E_1E_2(1-\cos\theta_{12})$.
(2) Factor out $(1+\delta)^2$; the square root gives $m' = (1+\delta)m$.

A common scale error moves the whole peak by the same fraction and leaves its relative width unchanged. ATLAS quoted a systematic error of 0.4 GeV on the mass 126.0 GeV, dominated by the photon and electron energy scales [118, Sec. 9.3]: a scale known to about 0.3%. That is $0.3\% \times 125 = 0.375$ GeV of peak shift, a quarter of the peak's width.

What the three kinds of effect look like. The script `01_mass_lineshape.py` takes a Crystal Ball signal template of core width 1.6 GeV, the shape of (T4.26) with the 2012 width, and applies a ±0.3% energy-scale shift, a ±14% resolution change (the total ATLAS uncertainty on the mass resolution, itself the quadrature sum $\sqrt{12^2 + 6^2 + 4^2} = 14$ of calorimeter resolution, detector material and pile-up [118, Sec. 5.6]) and a ±11% yield change (Figure T4.12). The fraction of the signal inside a ±2 GeV window around 125 GeV is 75.5% nominally, 69.6% and 81.8% for the wider and narrower peak, and 74.3% after the scale shift. A shape uncertainty is therefore also a normalisation uncertainty for any analysis that counts in a fixed window, and the two must not be counted twice.

T4.5.2 From up and down variations to a continuous parameter

The question. A calibration study delivers, for each source, the template at its nominal value and at "plus one" and "minus one" standard deviation. The likelihood of Chapter 4 needs the template at *every* value of the nuisance parameter. *How is the gap filled?*

The convention is to define a parameter α for each source (not the tail parameter of the Crystal Ball (T4.26)) with $\alpha = 0$ the nominal value and $\alpha = \pm 1$ the one-standard-deviation variations, and to interpolate. The simplest rule, used for a normalisation factor $r(\alpha)$ with variations $r^{\pm}$ (HistFactory calls it η, which here would be confused with the pseudorapidity), is

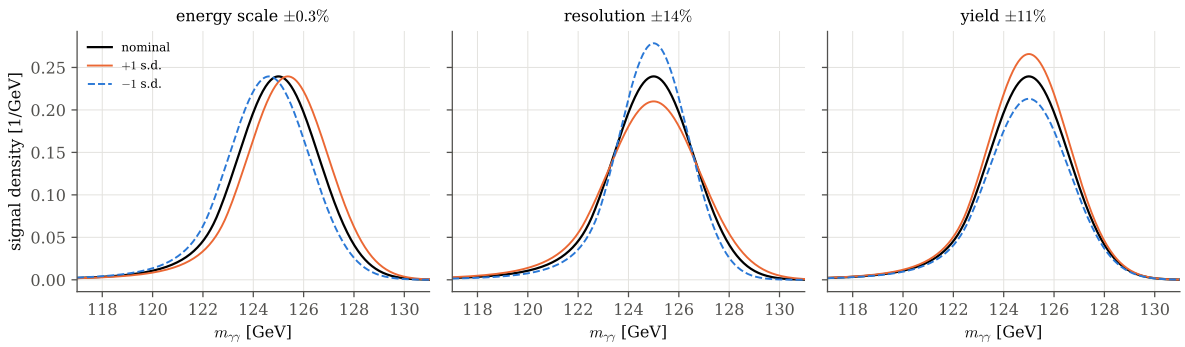


Figure T4.12: The three effects of a systematic uncertainty on a signal template, with "up" and "down" variations by one standard deviation. Left: an energy-scale shift of 0.3% slides the peak sideways by 0.4 GeV. Middle: a 14% change of the resolution makes the peak wider and lower, or narrower and taller, with the same area. Right: an 11% change of the yield rescales the peak without changing its shape.

piecewise linear [29, (23)–(25)]:

$$r(\alpha) = \begin{cases} 1 + \alpha\,(r^{+} - 1), & \alpha \geq 0, \\ 1 + \alpha\,(1 - r^{-}), & \alpha < 0, \end{cases} \tag{T4.35}$$

and the same rule, bin by bin, interpolates the up and down templates of a shape uncertainty. For the 8 TeV luminosity, $r^{\pm} = 1 \pm 0.036$; half a standard deviation up, $\alpha = 0.5$, scales the signal by 1.018. HistFactory assembles all of this into one expression for the expected count of sample s in bin b [125, (6)],

$$\nu_{sb} = \lambda\ \gamma_b\ \phi_s\ r_s(\boldsymbol{\alpha})\ \sigma_{sb}(\boldsymbol{\alpha}), \tag{T4.36}$$

which reads, in our notation: the luminosity $\lambda = L$; a per-bin factor γ_b for the Monte Carlo statistics; free factors ϕ_s, of which the signal strength μ is one; the interpolated normalisation (T4.35) (HistFactory's η_s); and the interpolated template σ_{sb}, our s_i or b_i. Every factor except ϕ_s comes with a constraint from its calibration.

Watch out: An up/down pair is not a measurement of the shape

Interpolating between three templates assumes the template changes smoothly and, in each bin separately, roughly linearly with α. A peak that *moves* violates this: halfway between a template shifted left and one shifted right, bin-by-bin interpolation gives a broadened or even two-humped peak, not a peak shifted halfway. Cranmer points to "horizontal" interpolation for such cases [29, Sec. 4.1.5]; the diphoton searches sidestep it because their signal model is an analytic function of the mass, whose mean can simply be shifted, $f(m \mid \bar{m}(1+\delta), \sigma_m)$.

T4.5.3 Calibrations: how a nuisance parameter is measured

A nuisance parameter is only as good as the measurement that constrains it. Two calibrations matter most for a diphoton search: the luminosity, which sets every normalisation, and the photon energy scale, which places the peak.

The luminosity: a van der Meer scan. Section T4.2.2 derived the luminosity from the overlap of two bunches, $L_{\text{inst}} \propto N_1 N_2 \int \rho_1 \rho_2 \, dx\, dy$. The bunch populations N_1, N_2 are measured by current monitors on the beams. The overlap integral is the hard part: the transverse sizes of the

beams at the collision point, a few tens of micrometres, cannot be measured directly with the needed precision. Van der Meer's idea is to measure the overlap by moving the beams. *What happens to the rate of collisions when one beam is displaced sideways by Δ?*

Let the x profiles of the two bunches be Gaussians of widths s_1 and s_2. The x part of the overlap at separation Δ is $\int g(x;s_1)\,g(x-\Delta;s_2)\,\mathrm{d}x$. This is the convolution integral (1.39): it is the probability density, at the value Δ, of the difference $X_1 - X_2$ of two independent Gaussian positions. A difference of independent Gaussians is Gaussian with the variances added, so

$$\int g(x;s_1)\,g(x-\Delta;s_2)\,\mathrm{d}x = \frac{1}{\sqrt{2\pi}\,\Sigma_x}\,e^{-\Delta^2/2\Sigma_x^2}, \qquad \Sigma_x = \sqrt{s_1^2+s_2^2}. \tag{T4.37}$$

For equal beams, $\Sigma_x = \sqrt{2}\,\sigma_x$ and this is (T4.10). The rate R of any process seen by a luminosity detector is proportional to the overlap, so scanning Δ and recording $R(\Delta)$ traces a Gaussian whose width is Σ_x. A second scan in y gives Σ_y. With them the head-on luminosity of one pair of bunches is known in absolute terms, $L_0 = f N_1 N_2/(2\pi\Sigma_x\Sigma_y)$, and the peak rate of the scan defines the detector's **visible cross section** $\sigma_{\mathrm{vis}} = R_0/L_0$. From then on, in ordinary running, every measured rate converts to a luminosity as $L_{\mathrm{inst}} = R/\sigma_{\mathrm{vis}}$.

The toy scan. The script `03_luminosity.py` simulates such a scan for one pair of LHC design bunches ($\sigma_x = 16.7\,\mu$m, so $\Sigma_x = 23.62\,\mu$m) and a luminosity detector with $\sigma_{\mathrm{vis}} = 60\,$mb. It counts Poisson events for 0.2 s at each of 25 separations between $\pm 5\Sigma_x$ and fits a Gaussian to the rates (Figure T4.13). The fit returns $\Sigma_x = 23.59 \pm 0.03\,\mu$m, a precision of 0.13%, and $\sigma_{\mathrm{vis}} = 60.04\,$mb against the true 60. Panel (b) checks the overlap formula itself: the integral done numerically on a grid agrees with $e^{-\Delta^2/4\sigma^2}$ to a relative 3.531×10^{-14}.

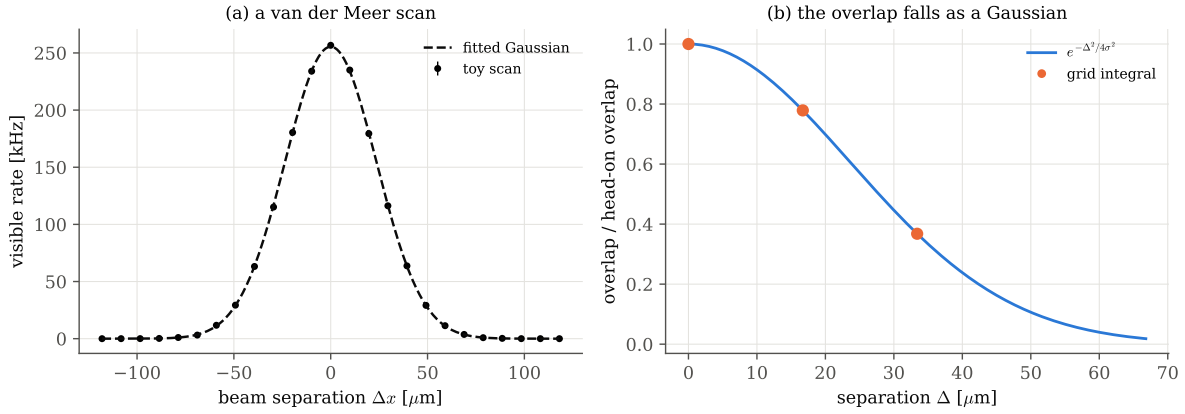


Figure T4.13: (a) A toy van der Meer scan: the collision rate seen by a luminosity detector as one beam is moved sideways across the other. The rate traces a Gaussian whose width is the combined size Σ_x of the two beams, which is what the luminosity formula needs. (b) The overlap of two equal Gaussian bunches falls with their separation as $e^{-\Delta^2/4\sigma^2}$; the points are the overlap integral computed numerically.

The toy's statistical precision, a tenth of a per cent, is far better than the luminosity uncertainties of 1.8% and 3.6% that ATLAS quoted for 2011 and 2012 [118, Sec. 4.3]. The real uncertainty comes from what the toy leaves out: bunch profiles that are not exactly Gaussian or do not factorise into an x part times a y part, the calibration of the separation Δ itself, and the stability of the luminosity detector between the scan, done in special conditions, and months of ordinary running. The toy shows the principle; the systematics of the real calibration set the number.

The energy scale: the Z boson as a standard candle. The Z boson's mass is known to 2 MeV, and $Z \to e^+e^-$ decays are plentiful. Electrons and photons make nearly the same showers in the calorimeter. So the electron pairs measure the energy scale: if the reconstructed m_{ee} peak sits at $(1+\delta)\,m_Z$, then by (T4.34) the electron energies are too large by the fraction δ, and they are corrected by $1/(1+\delta)$. ATLAS derived such η-dependent corrections, of order ±1%, from $Z \to e^+e^-$ events, and calibrated the photon identification and isolation efficiencies on electrons from Z decays and on photons radiated in $Z \to \ell^+\ell^-\gamma$ decays [118, Secs. 5.1, 5.6]. What remains uncertain is the transfer from electrons to photons (the material in front of the calorimeter affects them differently), which is why the material description appears among the scale and resolution uncertainties.

How a calibration enters the search. A calibration is a measurement with a result a_k and an error σ_k. Its probability, read as a function of the true parameter θ_k, is a likelihood factor that multiplies the likelihood of the search histogram [119, (40.17)], [29, Sec. 2.2]. Section 4.15.2 writes this full likelihood (4.89) with Gaussian constraint terms (4.90), and Section 4.15.3 removes the nuisance parameters by profiling.

Key idea: One source, one parameter, one calibration

Each source of systematic uncertainty is a nuisance parameter α_k, with $\alpha_k = 0$ nominal and $\alpha_k = \pm 1$ its one-standard-deviation variations. It can rescale templates (luminosity, efficiencies, cross sections), move or widen them (energy scale, resolution), or fake a signal (background function). Its value is constrained by an independent calibration measurement, whose likelihood multiplies that of the search.

Where this is used. The nuisance parameters $\boldsymbol{\theta} = (\beta, \kappa, \delta)$ of Section 4.15.2, a background normalisation, a signal yield and an energy-scale shift, are three of the sources above in their simplest form; their constraint terms are the calibrations of this section. In the capstone (Chapter 13), the selection script `code/ch13/31_hyy_select.py` reads the release's `photon_pt_syst` column (Section T4.7) to size the energy-scale variation, and the spurious-signal test of Section 13.22.2 sizes the uncertainty of the background function.

T4.6 The Higgs boson search of 2012

On 4 July 2012 the ATLAS and CMS collaborations reported, each with its own detector and its own analysis, an excess of events near 125 GeV that the background could hardly produce, and that looked like the decays of the Higgs boson [118, 124]. Those two papers use every piece assembled so far: cross sections and luminosities, invariant masses of photon and lepton pairs, simulated signal templates, background functions fitted to sidebands, a signal strength, nuisance parameters with calibrations, and finally a local and a global significance. The diphoton part of the search is also the template for the capstone of Chapter 13.

T4.6.1 Their question and their setup

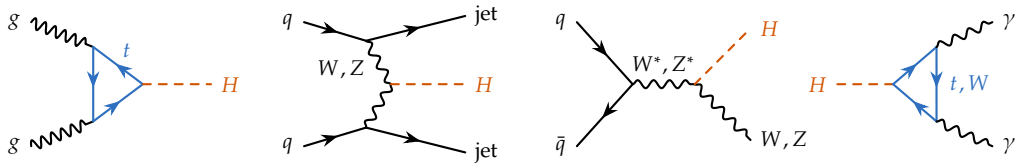
The question. Is there a particle with the properties of the Standard Model Higgs boson, and if so, at what mass? By 2011 the earlier experiments and the first LHC data had excluded most masses below 600 GeV except a window between 114 and 127 GeV, and both LHC experiments had seen excesses near 125 GeV of about three standard deviations [118, Sec. 1]. Three standard deviations had come and gone before. The 2012 data were to settle it.

What the theory fixes. For a given Higgs mass m_H, the Standard Model predicts everything

else: how often the boson is made, and how often it decays into each final state. There are four main ways to make it at the LHC [118, Sec. 3], drawn in Figure T4.14.

- *Gluon fusion* (ggF): two gluons make a Higgs boson through a loop of top quarks. About 88% of the diphoton signal at $m_H = 126.5\,\text{GeV}$.
- *Vector-boson fusion* (VBF): two quarks each radiate a W or Z, which fuse; the two quarks fly off as jets close to the beams. About 7%.
- *Associated production* with a W or Z (WH, ZH): a quark and an antiquark make a heavy vector boson that radiates the Higgs boson. About 3% and 2%.
- *Production with a top-quark pair* ($t\bar{t}H$): about 0.5%.

These are the fractions ATLAS quotes for its diphoton signal [118, Sec. 5.4]. The total cross section at $m_H = 125\,\text{GeV}$ was predicted to be 17.5 pb at $\sqrt{s} = 7\,\text{TeV}$ and 22.3 pb at 8 TeV [118, Sec. 3].



(a) ggF, 88% (b) VBF, 7% (c) WH, ZH, 5% (d) $H \to \gamma\gamma$

Figure T4.14: How a Higgs boson is made at the LHC (a–c, with the share of each mode in the 2012 diphoton signal; $t\bar{t}H$, 0.5%, is not drawn) and how it decays to two photons (d). The Higgs boson couples to mass, so it couples neither to the massless gluons nor to the photons directly; both (a) and (d) go through a loop of heavy charged particles, which is why the diphoton branching fraction is only about two per mille.

The channels. A Higgs boson of 125 GeV decays mostly to $b\bar{b}$, but the LHC makes b-quark pairs by the billion. Two final states combine a small branching fraction with a clean signature and an excellent mass resolution [118, Secs. 4–5], [124, Sec. 1]:

- $H \to \gamma\gamma$: two isolated photons, $\sigma B = 50\,\text{fb}$ at 8 TeV, mass resolution 1.6 GeV (Section T4.3.3);
- $H \to ZZ^{(*)} \to 4\ell$: two pairs of electrons or muons, one pair with the mass of a Z, $\sigma B = 2.8\,\text{fb}$ at 8 TeV, resolution 1.7–2.3 GeV.

The superscript $(*)$ says that one of the two Z bosons is off its mass shell: 125 GeV is not enough to make two real Z bosons of 91 GeV each. ATLAS added $H \to WW^{(*)} \to e\nu\mu\nu$, which has many more events but, with two neutrinos escaping, no peak in any mass. CMS also used $\tau\tau$ and $b\bar{b}$. The data sets were small by later standards; Table T4.1 collects them.

	ATLAS [118]	CMS [124]
data, 7 TeV (2011)	4.6–4.8 fb^{-1}	4.9–5.1 fb^{-1}
data, 8 TeV (2012)	5.8–5.9 fb^{-1}	5.0–5.3 fb^{-1}
channels	$\gamma\gamma$, 4ℓ, WW (+ $b\bar{b}$, $\tau\tau$ at 7 TeV)	$\gamma\gamma$, 4ℓ, WW, $\tau\tau$, $b\bar{b}$
pile-up per crossing	about 10 (7 TeV), 20 (8 TeV)	—
$\gamma\gamma$ events, 100–160 GeV	23 788 (7 TeV), 35 251 (8 TeV)	—

Table T4.1: The data of the two discovery papers. The luminosity ranges reflect the different data-quality requirements of the channels.

T4.6.2 Their expected counts, recomputed

The key equation. The expected number of selected signal events in each channel is (T4.15), $\nu_S = \sigma B\, L\, \varepsilon A$, with σB the predicted cross section times branching fraction, L the integrated luminosity and εA the fraction of produced events that survive the selection (Section T4.2.3). Both papers quote all four numbers or enough of them to recompute the others. Doing so shows what a discovery with 10 fb^{-1} looked like in events.

Example: the diphoton channel at 8 TeV. ATLAS expected $N_S = 110.5$ signal events in all categories together between 100 and 160 GeV, from $\sigma B = 50$ fb and $L = 5.9$ fb^{-1} [118, Table 4]. Then

$$\sigma B\, L = 50\,\text{fb} \times 5.9\,\text{fb}^{-1} = 295 \text{ produced}, \qquad \varepsilon A \overset{(1)}{=} \frac{N_S}{\sigma B L} = \frac{110.5}{295} = 0.375. \tag{T4.38}$$

(1) Solve $N_S = \sigma B L\, \varepsilon A$ for the selection efficiency.

ATLAS quotes an overall selection efficiency of "about 40%" [118, Sec. 5.4]. The recomputed 37.5% is a little lower because N_S also includes the 2–5% reduction of the gluon-fusion rate from interference (Section T4.4.2), and because "about" is about. The same selection kept $N_D = 35\,251$ data events in the same 60 GeV range. The signal was 0.3% of the selected events, concentrated in a few GeV.

Example: the four-lepton channel. The four-lepton signal is tiny: $\sigma B = 2.2$ fb at 7 TeV and 2.8 fb at 8 TeV [118, Sec. 4], so with 4.8 and 5.8 fb^{-1}

$$\sigma B L = 2.2 \times 4.8 + 2.8 \times 5.8 = 10.6 + 16.2 = 26.8 \text{ produced four-lepton decays.} \tag{T4.39}$$

In a window of ±5 GeV around 125 GeV ATLAS expected 5.3 signal events (2.09 + 2.29 + 0.90 in the 4μ, mixed and $4e$ sub-channels), an efficiency of 5.3/26.8 = 20%, and 4.9 background events, mostly non-resonant $ZZ^{(*)}$ production (2.4) and Z+jets and $t\bar{t}$ (2.5). It observed 13 [118, Table 3].

The two channels sit at opposite ends of the counting problem. In the diphoton channel thousands of background events per GeV surround a signal of a few dozen per GeV; the background is measured by the sidebands to high precision, and the question is whether a bump of a few per cent is real. In the four-lepton channel the signal and background are comparable, 5.3 against 4.9, and the question is whether 13 is too many for a Poisson count of mean 4.9, the small-count p-value of Section 4.1.4. Both questions have the same structure, $\nu = \mu s + b$ (Section T4.4.4); only the regime differs.

T4.6.3 What could have produced the excess?

What was seen. Both experiments saw more events than the background predicted, near 126 GeV in ATLAS and 125 GeV in CMS, in both the diphoton and the four-lepton channel (Figure T4.10 shows what such a spectrum looks like). *What could have produced it?* Before any number, the candidates.

A fluctuation of the background at this mass. Background events arrive as Poisson counts (Section T4.2.4), and a count can be high by chance. The probability that the background alone gives an excess at least as large as the one observed, at the mass where it was observed, is the local p-value p_0. For the ATLAS combination it was 1.7×10^{-9} at $m_H = 126.5$ GeV, a local significance of 5.9 standard deviations once the uncertainties of the energy scales and resolutions were included (6.0 without them); the Standard Model Higgs boson was expected

to give 4.9 on average [118, Sec. 9.2]. CMS found 5.0 standard deviations at 125.5 GeV, where 5.8 were expected [124, Sec. 7.1].

A fluctuation anywhere in the search range. The excess was not looked for at 126 GeV only; it could have appeared at any mass in the range searched, and a fluctuation somewhere is far more likely than a fluctuation at one fixed place. The global significance, discussed below, answers this version of the question: 5.1 standard deviations for ATLAS over 110–600 GeV.

A background shape that bends. If the smooth background function were wrong near 126 GeV, the fit would mistake the difference for signal. That possibility is the spurious signal of Section T4.4.3; it was measured in simulation and included in the likelihood as an extra uncertainty [118, Sec. 5.5].

An artefact of one detector. A calibration fault in the calorimeter could fake a diphoton bump. It could not also fake a four-lepton bump at the same mass, whose muons are measured by the tracker and the muon spectrometer. Nor could one detector's fault appear in the other experiment. The masses agree: $126.0 \pm 0.4\,(\text{stat}) \pm 0.4\,(\text{sys})$ GeV from ATLAS [118, Sec. 9.3] and $125.3 \pm 0.4 \pm 0.5$ GeV from CMS [124, Sec. 7], 0.7 GeV apart with a combined uncertainty of about 0.9 GeV.

A new particle. If the excess is a Higgs boson, its size should match the prediction, $\mu = 1$ in the notation of (T4.31). ATLAS measured $\hat{\mu} = 1.4 \pm 0.3$ at 126 GeV and CMS $\hat{\mu} = 0.87 \pm 0.23$, both compatible with one [118, Sec. 9.3], [124, Sec. 7]. Both papers add that a particle decaying to two photons cannot have spin one, and that decays to pairs of neutral vector bosons make it a neutral boson [118, Sec. 10], [124, Sec. 7].

Taken one at a time, each candidate except the last was either made improbable by a number or ruled out by a cross-check. Both collaborations therefore wrote "observation of a new particle". The first two candidates are the statistical part of the argument; we look at them more closely.

T4.6.4 Local and global significance

From p to Z. Particle physicists quote a p-value as the number Z of standard deviations a one-sided Gaussian tail would need to have the same probability, $p = 1 - \Phi(Z)$, with Φ the standard normal cumulative distribution (Section 4.9.1). Five standard deviations, $p = 2.9 \times 10^{-7}$, is the conventional threshold for "observation" (Section 4.9.2). The table below gives the values used in this section.

Z	3.28	4.5	4.6	5.0	5.1	5.3	5.9	6.0
$p = 1 - \Phi(Z)$	5.2×10^{-4}	3.4×10^{-6}	2.1×10^{-6}	2.9×10^{-7}	1.7×10^{-7}	5.8×10^{-8}	1.8×10^{-9}	9.9×10^{-10}

Local and global. The **local significance** answers: if there is only background, how probable is an excess at least this large *at this mass*? The **global significance** answers: how probable is an excess at least this large *anywhere in the range that was searched*? The second probability is larger, because the search effectively performs many tests, one for each place a peak could sit; this is the **look-elsewhere effect** [119, §40.3.2.2], [124, Sec. 4]. Their ratio, the trials factor, is roughly the number of independent places in the range. ATLAS states that it grows with the width of the searched mass range, with the significance of the excess, and with the sharpness of the peaks (a better mass resolution means narrower peaks and more independent places) [118, Sec. 7].

Example: the trials factors of 2012. Using the table above, the ratios of global to local p-values are:

- ATLAS, local 5.9: global 5.1 over 110–600 GeV, a trials factor $1.7 \times 10^{-7} / 1.8 \times 10^{-9} = 94$; global 5.3 over 110–150 GeV, a trials factor $5.8 \times 10^{-8} / 1.8 \times 10^{-9} = 32$ [118, Sec. 9.2].
- CMS, local 5.0: global 4.6 over 115–130 GeV, a factor 7; global 4.5 over 110–145 GeV, a factor 12 [124, Sec. 7.1].

The pattern follows the ranges: for CMS, widening the window from 15 to 35 GeV raised the trials factor from 7 to 12, and for ATLAS, widening it from 40 to 490 GeV raised it from 32 to 94. The factors of the two experiments cannot be compared directly, because the trials factor also grows with the significance itself (5.9 for ATLAS, 5.0 for CMS) and depends on each experiment's resolution. The range up to 600 GeV adds fewer places than its length suggests, because at high masses the peaks are wide (the 4ℓ resolution and the natural width grow with mass, Section T4.3.2) and the WW channel has no peak at all.

Recall: the toy scan of Chapter 4. Section 4.14.3 ran exactly this computation on simulated background-only spectra: sixty 1 GeV bins from 100 to 160 GeV, a Gaussian peak of width 1.5 GeV scanned over 105–155 GeV. There the biggest bump had a local significance of 3.28, which the global p-value of 0.0159 turned into 2.15 standard deviations, a trials factor of 31.3. The ATLAS search over 110–150 GeV is the closest real counterpart: a range of 40 against 50 GeV, peaks of comparable width, and a trials factor of 32 against 31. What is the same is the mechanism, many places for a fluctuation; what differs is the size of the excess. The toy bump was a fluctuation that the global correction rightly demoted; the 2012 excess survived it with more than five standard deviations, and the global significance, not the local one, is the number that justified the word "observation". How the global p-value is computed without millions of toy experiments is Section 4.14.4.

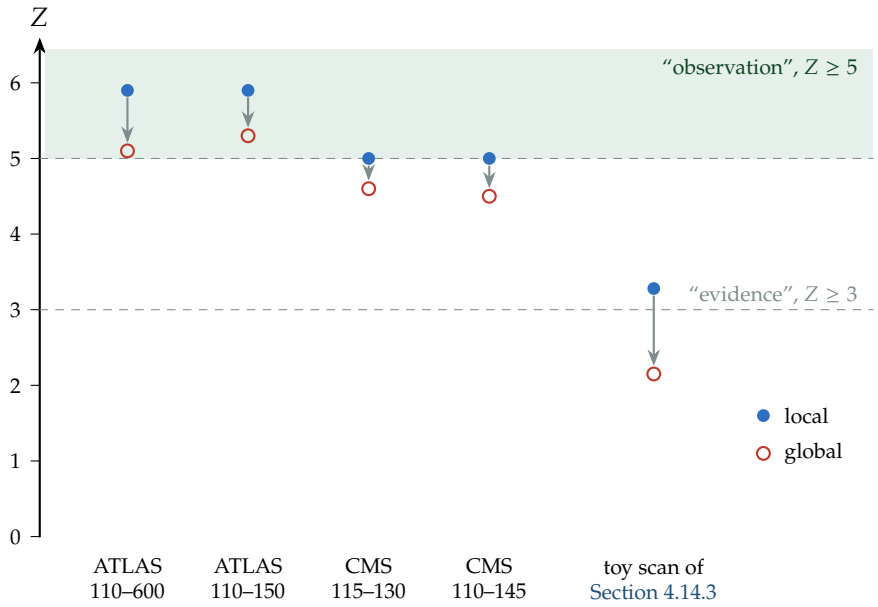


Figure T4.15: Local (filled) and global (open) significance of the largest excess, for the two 2012 searches with two search ranges each (in GeV), and for the simulated background-only scan of Chapter 4. Looking in a wider range always lowers the significance. The 2012 excesses stayed near or above five standard deviations after the correction; the toy bump, three standard deviations locally, dropped to about two.

T4.6.5 From 2012 to the open data

The capstone (Chapter 13) asks the 2012 diphoton question of the 2016 open data, with one channel instead of several. What carries over unchanged is the chain from collisions to expected counts: $\nu = \sigma B L \varepsilon A$ for the signal, a simulated template for its shape, a smooth function fitted to the sidebands for the background, μ for the size of the signal, nuisance parameters for the calibrations. What changes is the collision energy and so the cross section: the gluon-fusion $\sigma B(\gamma\gamma)$ stored in the open-data files is 0.102 pb = 102 fb at 13 TeV, twice the 50 fb of *all* production modes together at 8 TeV. With 10 fb^{-1} the open data therefore contain more Higgs bosons decaying to two photons than the whole 2012 diphoton sample of ATLAS. They also contain more background, and the open data were not processed with the full analysis machinery of the collaboration. Whether the bump appears, and how strongly, is the subject of Chapter 13.

The lesson in one line. A discovery is a chain of physical predictions, each with its calibration, ending in one comparison of counts; the statistics can only be as good as the expected counts ν_i it is given.

T4.7 The open data: what is in the files

In 2020 ATLAS released to the public a part of the proton collisions it recorded at $\sqrt{s} = 13$ TeV in 2016, together with simulated events for the main Standard Model processes [121, 123]. The capstone of Chapter 13 analyses its diphoton part. The files are not the raw detector output: they are tables with one row per recorded event and one column per reconstructed quantity, the reconstructed objects of Section T4.1 written out as numbers. Everything this chapter has introduced has a column, or a combination of columns, in these tables. Reading the columns correctly is the last physical step before the statistics.

The questions. What exactly is in a row? Which column is which quantity of this chapter, and in what units? And what does a real event look like when we compute its invariant mass by hand?

T4.7.1 The release

The recorded events come from the first four data-taking periods of 2016, about 270 million collisions written to disk. Only events taken while all relevant parts of the detector were working are kept, and after these quality requirements the data correspond to an integrated luminosity of $L = 10\,\text{fb}^{-1}$ [123, Sec. 2]. The format is a ROOT tree, a column-oriented table, with more than 80 columns (**branches**) [123, Sec. 2]. The overview of the release lists the photon requirements as reconstruction in the inner detector and electromagnetic calorimeter, $E_T > 25$ GeV, $|\eta| < 2.37$, tight identification and loose isolation [123, Table 1]. The diphoton files nevertheless keep candidates that fail the tight identification and record the result in a flag: requiring the flag on both photons keeps only 979 059 of the 7 798 424 data events (Figure T4.5). The flag must be checked.

The release is split into collections by final state. The diphoton collection, `GamGam`, holds four data files, one per period (A to D), and simulated Higgs bosons with $m_H = 125$ GeV decaying to two photons, one file per production mode (gluon fusion, vector-boson fusion, WH, ZH). It contains only events that passed a two-photon trigger and have two photon candidates. In all four data files the script `02_open_data.py` finds 7 798 424 events, every one of them with the photon trigger flag set and exactly two photon candidates stored (Figure T4.17d). The first two lines of the cut flow of Figure T4.5 therefore remove nothing: the release had already applied them.

Recall: the luminosity of the release. Section T4.2 used $L = 10\,\text{fb}^{-1}$ to predict 1020 produced and 395 selected gluon-fusion Higgs decays. The capstone uses $L = 10.06\,\text{fb}^{-1}$ (Section 13.21.1), the value the release's own analysis code multiplies by, of which the published "$10\,\text{fb}^{-1}$" is the rounded form. The two differ by 0.6%, a sixth of the 3.6% luminosity uncertainty of the 2012 data (Section T4.5.3); every expected count in this chapter would be 0.6% larger with the capstone's value.

T4.7.2 The columns

A row of the diphoton tree has three kinds of column (Figure T4.16). *Event-level numbers* describe the whole crossing: which run and event it is, which trigger fired, and for simulated events the weights. *Per-object vectors* hold one entry per reconstructed photon (or lepton, or jet): `photon_pt` is a list of transverse momenta, the first entry for the leading photon, the second for the next. *Normalisation constants* of the simulation, σB and the total weight W of (T4.29), are stored in every row of a simulated file. Table T4.2 lists the columns a diphoton analysis uses; the full list of the tree is written by the script to `results/chT4/02_branches.txt`.

event-level	one entry per photon	γ_1	γ_2	simulation only
runNumber	photon_pt	176.1	39.63	mcWeight
eventNumber	photon_eta	−0.834	0.095	scaleFactor_*
trigP = true	photon_phi	−0.294	0.567	XSection
photon_n = 2	photon_E	240.9	39.81	SumWeights
	photon_isTightID	true	true	photon_pt_syst

Figure T4.16: The anatomy of one row of the diphoton tree. Left: numbers that describe the whole event. Middle: vectors with one entry per photon candidate, here filled with the first selected data event (momenta and energies in GeV after dividing the stored MeV by 1000). Right: columns that are meaningful only for simulated events: the weights and the normalisation of (T4.29), and the energy-scale variation of each photon.

Four columns need a word beyond the table.

- *Isolation.* `photon_etcone20` adds up the transverse energy deposited in the calorimeter in a cone of radius $\Delta R = 0.2$ around the photon, excluding the photon's own shower. A photon from a Higgs decay is alone, so this sum is small; a "photon" that is really a neutral pion inside a jet is surrounded by the jet's other hadrons. Requiring less than 4 GeV, as ATLAS did in 2012 [118, Sec. 5.1], keeps 97.7% of simulated signal photons and only 56.0% of the photon candidates in the data (Figure T4.17c).
- *Conversion.* A photon crossing the inner detector may convert into an electron–positron pair before it reaches the calorimeter; its energy is then measured less well and its shower starts earlier. Of the selected photons in the data, 34.7% are converted. The 2012 analysis split events into categories by conversion status because the resolution differs (Section T4.6.2); the capstone uses the unconverted central category the same way.
- *Scale factors.* The simulation reproduces efficiencies only approximately. Each factor is the ratio of an efficiency measured in data (for example with photons from $Z \to \ell^+\ell^-\gamma$, Section T4.5.3) to the same efficiency in simulation, and multiplies the event's weight. The product $w_e = \texttt{mcWeight} \times \prod \texttt{scaleFactor}$ is the weight of (T4.29).
- *The energy-scale variation.* `photon_pt_syst` gives, for each simulated photon, how far its p_T moves when the energy scale is varied by its uncertainty. Its median relative size in the

column	meaning	symbol, section
`runNumber`, `eventNumber`	which data-taking run and which crossing	—
`trigP`	the photon trigger fired	Section T4.1
`photon_n`	number of photon candidates stored	—
`photon_pt`	transverse momentum, MeV	p_T, (T4.1)
`photon_eta`	pseudorapidity $-\ln\tan(\theta/2)$	η, (T4.4)
`photon_phi`	azimuth around the beam, rad	ϕ, Section T4.1.2
`photon_E`	energy, MeV	$E = p_T\cosh\eta$
`photon_isTightID`	shower shape passes the tight photon identification	ε, Section T4.2.3
`photon_etcone20`	calorimeter transverse energy in a cone $\Delta R < 0.2$ around the photon, MeV	isolation
`photon_ptcone30`	sum of track p_T in a cone $\Delta R < 0.3$, MeV	isolation
`photon_convType`	0 if unconverted, otherwise how it converted to e^+e^-	categories
`photon_pt_syst`	energy-scale variation of the photon's p_T, MeV (simulation)	δ, (T4.34)
`mcWeight`	generator weight of the event	w_e, (T4.29)
`scaleFactor_*`	data/simulation corrections of the photon efficiency (`_PHOTON`), the trigger efficiency (`_PhotonTRIGGER`) and the pile-up (`_PILEUP`)	factors of w_e
`XSection`	cross section times branching fraction, pb	σB
`SumWeights`	sum of the weights of all generated events	W, (T4.29)

Table T4.2: The columns a diphoton analysis reads, their meaning, and the symbol each carries. Momenta and energies are stored in MeV. The cone size $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$ is a distance in the (η, ϕ) plane.

selected signal is 0.71%. By (T4.34) this is directly the relative shift of the diphoton peak, the energy-scale nuisance parameter of Section T4.5.1. In the data files the column carries no information, as it must: a recorded energy has no "varied" version.

Watch out: The files store MeV

Momenta and energies in the 2016 release are in MeV, not GeV. An invariant mass computed without dividing by 1000 comes out near 125 000, and a cut "$p_T > 40$" applied to the raw column keeps every photon. The script checks the median p_T and divides by 1000 when it is above a thousand.

T4.7.3 One event by hand

The question. Take the first event of period A that passes the selection of Section T4.2.3. What is its diphoton mass, and do the columns agree with each other?

The two photons have, after conversion to GeV, $p_{T1} = 176.1$ and $p_{T2} = 39.63$ GeV, $\eta_1 = -0.8335$ and $\eta_2 = 0.0947$, $\phi_1 = -0.2942$ and $\phi_2 = 0.5673$ rad, and energies $E_1 = 240.9$ and $E_2 = 39.81$ GeV.

Check one: are the energies those of massless particles? For a photon, $E = |\boldsymbol{p}| = p_T\cosh\eta$ (T4.5). With $\cosh(0.8335) = 1.368$ and $\cosh(0.0947) = 1.0045$, $176.1 \times 1.368 = 240.9$ and $39.63 \times 1.0045 = 39.81$ GeV: both stored energies are reproduced to the last digit. The columns are consistent, and the photons are massless as they should be.

The mass from the detector variables. With (T4.19), $\Delta\eta = \eta_1 - \eta_2 = -0.9282$ and $\Delta\phi = \phi_1 - \phi_2 = -0.8615$,

$$m_{\gamma\gamma}^2 \overset{(1)}{=} 2\,p_{T1}p_{T2}\big(\cosh\Delta\eta - \cos\Delta\phi\big) \overset{(2)}{=} 2 \times 176.1 \times 39.63 \times (1.4626 - 0.6513)$$

$$\overset{(3)}{=} 13\,958 \times 0.8113 = 11\,324\,\text{GeV}^2, \qquad m_{\gamma\gamma} = 106.4\,\text{GeV}. \tag{T4.40}$$

(1) The diphoton mass in detector variables, (T4.19).
(2) $\cosh(-0.9282) = \cosh(0.9282) = 1.4626$ (cosh is even) and $\cos(-0.8615) = 0.6513$.
(3) Multiply out.

The mass from four-vectors. Build $p_x = p_T \cos\phi$, $p_y = p_T \sin\phi$, $p_z = p_T \sinh\eta$ for each photon and add: the pair has $\boldsymbol{P} = (202.0, -29.8, -160.6)\,\text{GeV}$ and $E = 280.7\,\text{GeV}$, so $m^2 = E^2 - |\boldsymbol{P}|^2$ by (T4.17) gives $m = 106.4\,\text{GeV}$ again. The script finds 106.4 and 106.4 GeV for the two routes.

This event is not a Higgs boson candidate in any useful sense: its mass is 19 GeV below the peak, and in any case a single event is neither signal nor background (Section 13.21.1). It is a typical background event. The pair is strongly boosted along the beam, $P_z = -161\,\text{GeV}$, which is the unknown parton motion of Section T4.1.2; the mass does not care.

T4.7.4 What the columns look like

Figure T4.17 histograms four columns for the data and for the simulated gluon-fusion signal, after the trigger and two-photon requirement and before any other cut.

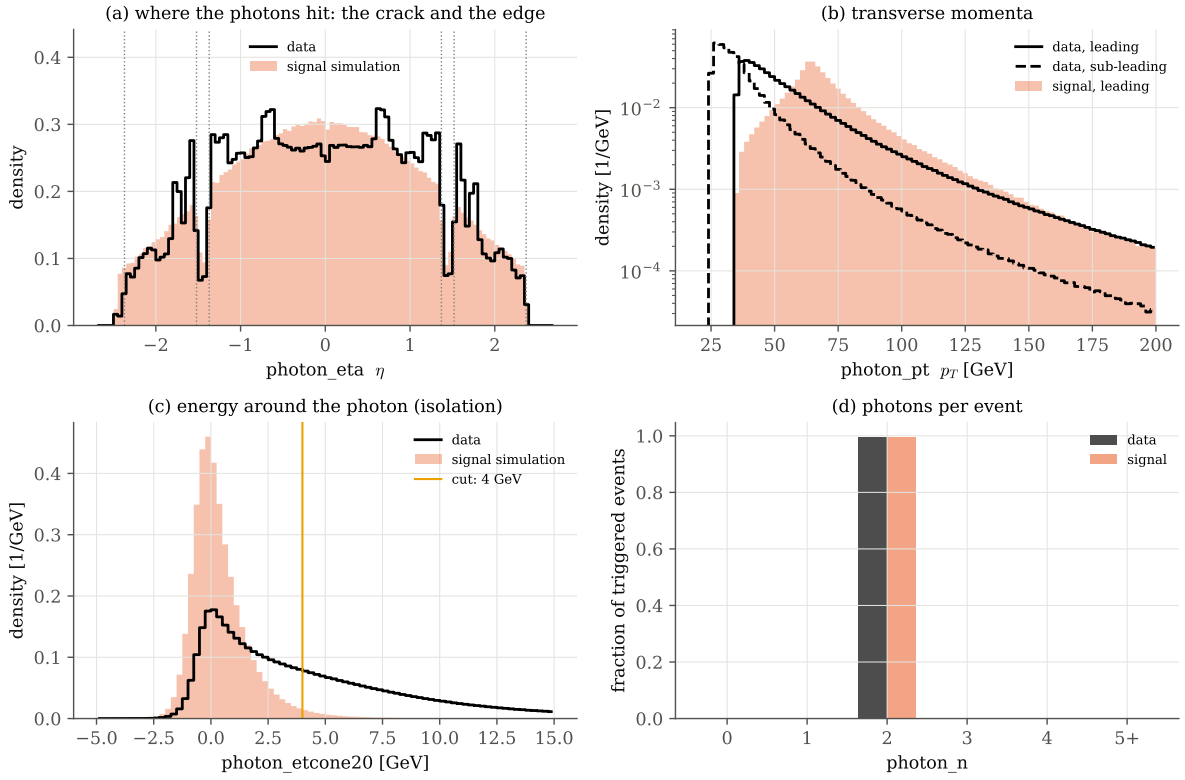


Figure T4.17: Four columns of the diphoton files, for recorded events (black) and simulated Higgs decays (filled). (a) Photon pseudorapidity: no photon beyond the calorimeter edge at $|\eta| = 2.37$, and a dip in the data at the barrel/end-cap gap $1.37 < |\eta| < 1.52$ (dotted lines). (b) Transverse momenta: the data fall steeply from the trigger thresholds, while the signal photons peak near half the Higgs mass. (c) Isolation energy: signal photons are alone, with a narrow peak near zero; the data have a long tail from jets faking photons, which the 4 GeV cut removes. (d) Every event in the files has exactly two photon candidates.

Example: reading the pseudorapidity panel. In panel (a) the signal photons spread smoothly over the calorimeter, more of them at small $|\eta|$, because the decay products of a Higgs boson

with modest longitudinal motion come out at all angles. The data have structure the signal lacks: the density of data candidates changes abruptly at the boundaries of the calorimeter regions, while the simulated signal, made of real photons, changes smoothly. The data are mostly fake photons from jets, and how often a jet passes as a photon depends on the calorimeter region. Of the data photon candidates, 3.2% sit inside the gap $1.37 < |\eta| < 1.52$, where the energy measurement is poor; the selection removes them. This is the acceptance cut of Section T4.2.3 seen in the raw columns.

Example: reading the momentum panel. In panel (b) the data start at the trigger and preselection thresholds, about 35 GeV for the leading and 25 GeV for the second photon, and fall over three orders of magnitude by 200 GeV. The signal's leading photon peaks near $m_H/2 = 62.5$ GeV, the energy of a photon from a Higgs boson at rest (Section T4.3.1), with a tail to higher values from the boost. Cuts on p_T therefore remove much of the low-mass background while keeping most of the signal; the capstone uses cuts relative to $m_{\gamma\gamma}$ instead, for the reason given in Section 13.21.1.

Recall: this selection and the capstone's. Section T4.2.3 followed the gluon-fusion signal through a selection modelled on the 2012 analysis: tight identification, $|\eta| < 2.37$ outside the gap, $p_T > 40$ and 30 GeV, calorimeter isolation below 4 GeV, and $100 < m_{\gamma\gamma} < 160$ GeV. It kept 151 767 data events and 395 expected signal events. The capstone (Section 13.21.1) uses the selection of the release's own example analysis: exactly two photons each passing tight identification with $p_T > 25$ GeV, both a calorimeter and a track isolation below 6.5% of the photon's p_T, the relative cuts $p_{T1}/m_{\gamma\gamma} > 0.35$ and $p_{T2}/m_{\gamma\gamma} > 0.25$, and $105 < m_{\gamma\gamma} < 160$ GeV. It keeps 76 394 data events and 357.2 gluon-fusion events (397.9 from all four production modes). The capstone's data sample is half as large as ours and its gluon-fusion signal 10% smaller: the extra track isolation and the relative isolation cut remove many more jets than signal photons, and the narrower mass range removes 5 GeV of the steepest part of the background. If the background under the peak falls in the same proportion as the whole sample, $s/\sqrt{b}$ in a fixed window grows by $0.9/\sqrt{0.5} \approx 1.3$, which is the purpose of the tighter selection; Section 13.21.1 shows its cut flow.

Where this is used. The capstone reads exactly the columns of Table T4.2: the photon momenta for $m_{\gamma\gamma}$ via (T4.19), the identification, isolation and conversion flags for the selection and the categories, and `mcWeight`, the scale factors, `XSection` and `SumWeights` for the signal template (T4.29).

T4.8 The symbols of a collider search

The statistics of a search treats each quantity of a collider measurement as a given number or function. Table T4.3 lists each one with its meaning, the relation that defines or computes it, a typical size for the diphoton search, and the places where it is used. Two symbols are easily confused and are kept apart throughout the book: μ is always the signal strength, never a mean; the mean number of events is always ν. The integrated luminosity is L and the likelihood is $\mathcal{L}$; HistFactory and Cranmer write λ for the luminosity for the same reason.

Table T4.3: The symbols of a collider search: meaning, working relation, typical size in the diphoton search, and where each is used.

symbol	meaning	working relation	typical size	used in
$\sqrt{s}$	collision energy of the two protons	—	7, 8, 13 TeV	Section T4.6 and Chapter 13
p_T	momentum transverse to the beam	$\lvert\boldsymbol{p}\rvert \sin\theta$, (T4.1)	25–200 GeV	Section T4.7 and Chapter 13
y	rapidity; shifts by a constant under a boost along the beam	$\frac{1}{2}\ln\frac{E+p_z}{E-p_z}$, (T4.2)	—	Section T4.1.2
η	pseudorapidity, a measure of the polar angle	$-\ln\tan\frac{\theta}{2}$, (T4.4)	$\lvert\eta\rvert < 2.37$	Section T4.7 and Chapter 13
ϕ	azimuth around the beam	—	$(-\pi, \pi]$	Chapter 13
σ	cross section of a process	$N = \sigma L$	$\sigma_{\text{inel}} = 78$ mb; $\sigma B = 0.102$ pb	Section T4.2 and Chapter 13
L_{inst}	instantaneous luminosity	$\frac{n_b f N_1 N_2}{4\pi\sigma_x\sigma_y}F$, (T4.11)	$10^{34}\,\text{cm}^{-2}\text{s}^{-1}$	Section T4.2.2
L	integrated luminosity	$\int L_{\text{inst}}\,\mathrm{d}t$	$10\,\text{fb}^{-1}$	Section 4.15 and Chapter 13
$\bar{n}_{\text{pu}}$	collisions per bunch crossing (pile-up)	$\sigma_{\text{inel}}L_{\text{inst}}/n_b f$, (T4.13)	20–50	Section T4.2.2
A, ε	acceptance, efficiency	fraction of produced events kept	$\varepsilon A = 0.39$	Section T4.2.3 and Chapter 13
ν	expected number of selected events	$\sigma L \varepsilon A$, (T4.15); Poisson mean	395 signal	Section 4.15.1
$m_{\gamma\gamma}$	invariant mass of the photon pair	$m^2 = 2p_{T1}p_{T2}\times$ $(\cosh\Delta\eta - \cos\Delta\phi)$, (T4.19)	100–160 GeV	Section 4.14.3 and Chapter 13
Γ, τ	natural width, lifetime	$\Gamma = \hbar/\tau$; Breit–Wigner (T4.23)	$\Gamma_H = 4$ MeV, $\Gamma_Z = 2.5$ GeV	Section T4.3.2
σ_E/E	calorimeter energy resolution	$a/\sqrt{E}\oplus c$, (T4.24)	1.4% at 62.5 GeV	Section T4.3.3
σ_m	mass resolution	$\frac{1}{2}\sqrt{\sum(\sigma_{E_j}/E_j)^2}\,m$, (T4.25)	1.2–1.6 GeV	Section 4.14.3 and Chapter 13
α, n	Crystal Ball tail start and power	(T4.26), (T4.27)	1.5, 5	Section 13.22.1
s_i, b_i	signal and background templates in bin i	$s_i = \sum_{e\in i}\sigma L w_e/W$, (T4.29); b_i from sidebands	$\sum s_i \approx 400$	Sections 4.15 and 13.22
w_e, W	weight of a simulated event, total weight	`mcWeight`×scale factors; `SumWeights`	$W = 5.6\times10^7$	Section T4.4.2
μ	signal strength	$\sigma B/(\sigma B)_{\text{SM}}$; $\nu_i = \mu s_i + b_i$, (T4.31)	1.4 ± 0.3 (ATLAS 2012)	Sections 4.15 and 4.16
α_k	nuisance parameter of source k	0 nominal, ±1 one s.d.; (T4.35)	—	Section 4.15.2
δ	energy-scale shift	$\delta m/m = \delta$, (T4.34)	0.3%	Section 4.15.2
Σ_x	combined width of two beams	$\sqrt{s_1^2 + s_2^2}$, (T4.37)	24 μm	Section T4.5.3
σ_{vis}	visible cross section of a luminosity detector	R_0/L_0	tens of mb	Section T4.5.3

symbol	meaning	working relation	typical size	used in
Z, p_0	significance, background-only p-value	$p_0 = 1 - \Phi(Z)$	5.9, 1.7×10^{-9}	Sections 4.9.1 and 4.14

Recap

- An event is a list of reconstructed particles, each with p_T, η, ϕ; the transverse plane and rapidity differences are used because the motion of the colliding partons along the beam is unknown.
- With the integrated luminosity fixed, the number of selected events of a process is Poisson with mean $\nu = \sigma L \varepsilon A$.
- The invariant mass of a decay's products equals the parent's mass in every event; the peak's width combines the natural width (Breit–Wigner) and the detector resolution (Gaussian or Crystal Ball core), and is dominated by the larger.
- The expected count in a bin is $\nu_i = \mu s_i + b_i$: signal templates from weighted simulation, background from the sidebands of the data, $\mu = \sigma B / (\sigma B)_{\mathrm{SM}}$.
- Every uncertain input is a nuisance parameter with its own calibration: van der Meer scans for the luminosity, Z decays for energy scale and efficiencies, simulated background samples for the spurious signal.
- In 2012 ATLAS saw a local 5.9σ and a global 5.1σ excess at 126 GeV, CMS a local 5.0σ at 125.5 GeV, with $\hat{\mu} = 1.4 \pm 0.3$ and 0.87 ± 0.23.
- The 2016 open data store, per event, the photon four-momenta in MeV, identification, isolation and conversion flags, and for simulation the weights, scale factors, σB and the total weight.

Chapter summary

concept	operational meaning	used later in
event, reconstructed object	one bunch crossing kept by the trigger, written as a list of photons, leptons, jets with (p_T, η, ϕ, E)	Chapter 13 (the data rows)
transverse plane, pseudorapidity	coordinates insensitive to the unknown parton motion along the beam; $\eta = -\ln\tan(\theta/2)$	Chapter 13 (acceptance cuts)
cross section, luminosity	$N = \sigma L$: the process's effective area times the beams' collision opportunities per area	Section 4.15, Chapter 13
acceptance, efficiency	fraction of produced events inside the detector and passing the selection; $\nu = \sigma L \varepsilon A$	Section 4.15.2, Chapter 13
Poisson counts	with L fixed, independent crossings and independent selection make every count Poisson	Section 4.15.1
invariant mass	$m^2 = (\sum E)^2 - \lvert\sum \boldsymbol{p}\rvert^2$, the parent's mass in every decay	Section 4.14.3, Chapter 13
natural width, Breit–Wigner	a finite lifetime spreads the mass as a Cauchy density of full width $\Gamma = \hbar/\tau$	Section 2.10.5 (recalled)
resolution, Crystal Ball	measurement error spreads the peak; Gaussian core plus a low-side power-law tail	Section 13.22.1, Section 6.14
template, Monte Carlo weight	expected count per bin from a weighted simulated sample, $\sigma L w_e / W$ per event	Section 4.15, Section 13.22

concept	operational meaning	used later in
sideband background, spurious signal	a smooth function fitted away from the peak; its possible bias measured on simulated background	Section 13.22.2
signal strength μ	$\sigma B/(\sigma B)_{\mathrm{SM}}$; $\nu_i = \mu s_i + b_i$; $\mu = 0$ is the null hypothesis	Section 4.16, Chapter 13
nuisance parameter, calibration	every uncertain input gets a parameter and the likelihood of its calibration measurement	Sections 4.15.2 and 4.15.3
van der Meer scan	the collision rate versus beam separation is a Gaussian of width $\Sigma = \sqrt{s_1^2 + s_2^2}$; it calibrates L	Section T4.5.3
local and global significance	p-value of an excess at its mass, or anywhere in the searched range; their ratio is the trials factor	Sections 4.14 and 4.14.4
open-data columns	`photon_pt`, `photon_eta`, `photon_phi`, `photon_E` (MeV), flags, weights, `XSection`, `SumWeights`	Section 13.21.1

Reading the literature. The kinematics of collider events (four-vectors, rapidity, pseudorapidity, invariant masses), the cross-section formulae and the luminosity of colliding beams are collected in the reviews of the Particle Data Group [119], which are the standard reference for every relation used here; its statistics review summarises the treatment of nuisance parameters and the look-elsewhere effect. Cranmer's lectures [29] explain how a collider analysis turns physics into a statistical model, in the three narratives (simulation, data-driven, effective model) used in Section T4.4, and the HistFactory note [125] writes that model down in the form LHC software uses. Cowan's textbook [Cowan §§6.9, 11.1] gives the extended likelihood and the response-matrix picture of a detector at an introductory level. The two discovery papers [118, 124] are short and readable once the vocabulary of this chapter is in hand; read the ATLAS diphoton section and the description of the statistical procedure first. The design of the LHC and of the ATLAS detector is documented in [117, 122], the inelastic cross section in [120], and the open-data release in [121, 123] together with the documentation on the CERN open-data portal.

Chapter T5

Choosing an Estimator

Where this chapter is going

The estimators we have met so far use different kinds of information. A weighted average accounts for differences in measurement precision, a likelihood fit uses a model of how the data were generated, and a posterior mean also incorporates prior information. Moment matching and least squares offer further choices. Faced with a new data set, how do we decide which assumptions we can defend and which estimator makes use of them?

Consider first what we know about the noise. Can we specify its full distribution, estimate only its mean and variance, or say only that most readings are reliable and a few may be outliers? Do we have prior information worth using? Does the parameter shift the data (it sits in their mean) or does it make them scatter more (it sits in their covariance)? Are many parameters estimated at once? Do the data arrive all together or one reading at a time? And does the best recipe need the very value it is trying to find? Figure T5.1 draws the questions and where each one leads.

These choices lead to five families of estimators. For *linear* estimators we derive the conditions under which a weighted average has the smallest variance in its class. Assessing *Bayes* estimators through their errors in repeated experiments leads us to the Wiener filter. When several parameters are estimated together, *shrinkage* can reduce their combined error, while *robust* estimators limit the influence of outlying readings. Parameters that enter through the covariance call for *quadratic* estimators, for which noise contributes to the mean as well as the scatter. The second half of the chapter adds the estimators that iterate, that correct their own bias, and that update as data stream in, and closes with a field guide that turns the questions into a decision procedure.

In the book's pipeline (Chapter 0) all of this sits on one arrow: from the data to the summary S. Choosing the estimator is choosing S, and every later arrow (its sampling distribution, the inference) inherits that choice.

T5.1 The questions that choose an estimator

What we already have. Chapter 3 gave us a toolbox. For readings of different quality we have the weighted mean, whose weights are the inverse variances (Section 3.7.1). For a model with a known noise law we have maximum likelihood (Section 3.9), and for Gaussian noise it becomes least squares (Section 3.10). When the likelihood is awkward we can match moments (Section 3.9.2). With a prior we can report the posterior mean, median or mode, each the best answer for a different loss (Section 3.16.4). The Cramér–Rao bound tells us how well any unbiased recipe can do (Section 3.13.3). Each tool came with its own setting, and in that setting

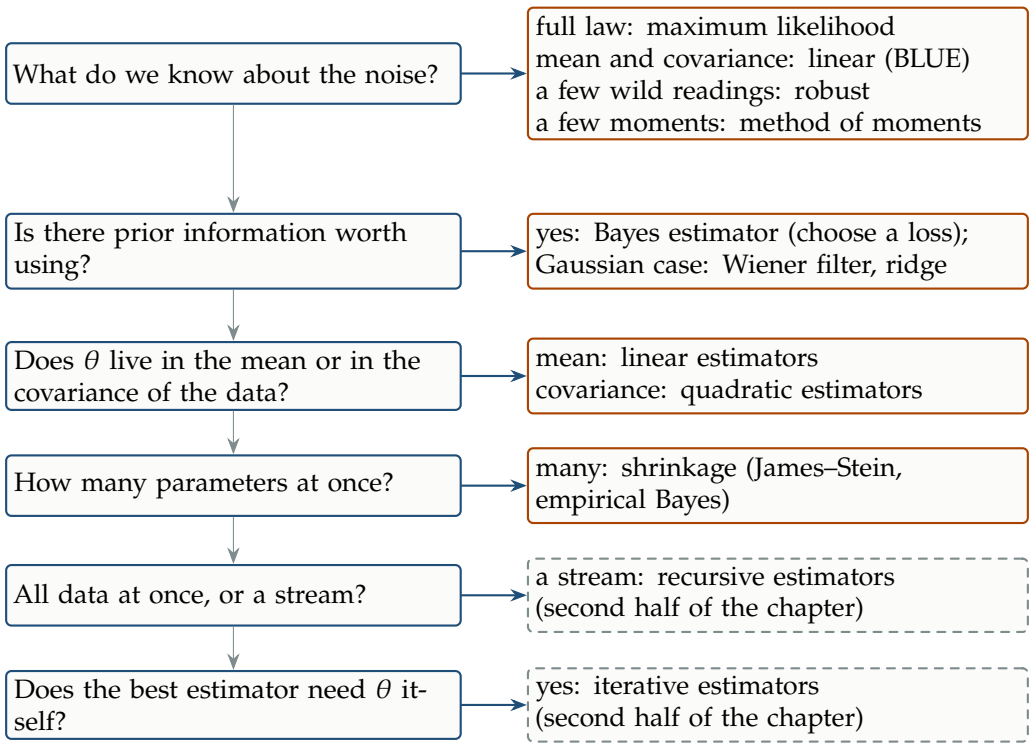


Figure T5.1: The questions that choose an estimator (left, blue) and the family each answer points to (right). The questions are asked in order, and a real problem usually needs several answers at once: a power spectrum from a masked map, for example, is a covariance parameter (third row), estimated for many multipoles at once (fourth row), with a recipe that needs a guess of the spectrum it is estimating (last row). The dashed boxes are built in the second half of the chapter.

it is excellent.

When that is not enough. A real data set does not arrive with a label saying which setting it belongs to. Two experiments can produce readings with the same average and the same spread, and yet the best way to combine them can differ by a large factor. What decides it is what we know about how the readings go wrong.

A story with the whole trick in it. Suppose we want to know the time and we have no clock, only the watches of twenty-five people in a room. Most watches are a minute or two fast or slow, in no particular direction. Averaging the twenty-five readings is then the natural thing to do: the small errors partly cancel. But suppose a few of the watches stopped hours ago, or were never reset after a flight. One of those can drag the average by an hour, while the middle reading, the one with twelve watches earlier and twelve later, does not care how far off the stopped watches are. Now suppose we also know that the church clock across the square is rarely more than a minute wrong: we should lean towards it, more so when the watches disagree among themselves. Finally, suppose the question changes. We no longer want the time; we want to know *how badly people set their watches*, the spread of their errors. Then our own squinting at each dial adds a spread of its own to what we see, and unless we subtract it we will blame the watch owners for our eyesight. Four situations, four different best answers, from the same twenty-five dials.

The laboratory is the same. A CCD pixel read many times has small thermal noise, plus the occasional cosmic-ray hit that adds a huge spike. A source position can be measured by flashes whose scatter has no finite variance at all, as for the lighthouse of [Sivia §2.4, pp. 31–36], whose

flashes land on the shore with the Cauchy law of Section 3.3.4. A CMB experiment wants the *variance* of the sky fluctuations, the power spectrum, and the detector noise adds a variance of its own. The idea of this chapter is that the choice of estimator is not a matter of taste: it follows from a few questions about the problem, and the answer to each question changes the best recipe in a way we can compute.

The story in symbols: three laboratories, one number. Three laboratories measure the same quantity μ, each taking n readings $x_i = \mu + e_i$. In all three the noise e_i has mean zero and variance one, so a summary that sees only the mean and the variance cannot tell the laboratories apart. They differ in the *law* of the noise:

- Laboratory A has Gaussian noise, $e \sim \mathcal{N}(0,1)$: many small independent kicks, as the central limit theorem predicts.
- Laboratory B has Laplace (double-exponential) noise, $f_B(e) = \frac{1}{2b}e^{-|e|/b}$, whose tails fall exponentially rather than like a Gaussian. Its variance is $2b^2$, so $b = 1/\sqrt{2}$ gives variance one.
- Laboratory C has a Gaussian core of width s, but a fraction $\epsilon = 0.05$ of the readings comes from a Gaussian ten times wider: the cosmic-ray hits. Its variance is $(1-\epsilon)s^2 + \epsilon(10s)^2 = s^2(1+99\epsilon)$, so $s = 1/\sqrt{1+99\epsilon} = 0.410$ gives variance one.

The left panel of Figure T5.2 draws the three laws on a logarithmic scale. Laboratory C has the narrowest core and the fattest tails, even though all three have the same variance.

The Laplace variance quoted above is a one-line integral, and it shows the kind of step we will do often:

$$\operatorname{Var} e \overset{(1)}{=} \int_{-\infty}^{\infty} e^2 \,\frac{1}{2b} e^{-|e|/b}\,\mathrm{d}e \overset{(2)}{=} \frac{1}{b}\int_0^{\infty} e^2 e^{-e/b}\,\mathrm{d}e \overset{(3)}{=} \frac{1}{b}\,\Gamma(3)\,b^3 = 2b^2. \tag{T5.1}$$

(1) The mean is zero by symmetry, so the variance is the second moment.
(2) The integrand is even: twice the integral over $e > 0$, where $|e| = e$.
(3) The gamma integral $\int_0^{\infty} x^{a-1}e^{-x/b}\mathrm{d}x = \Gamma(a)b^a$ with $a = 3$, and $\Gamma(3) = 2$.

The obvious choice: the sample mean. Its variance is $\operatorname{Var}\bar{x} = \sigma^2/n$ (Eq. (3.20)), and that formula used nothing but the variance of a single reading. So in all three laboratories $n \operatorname{Var}\bar{x} = 1$, exactly. The mean is blind to the shape of the noise.

A second choice: the sample median. For large n its variance is $1/(4nf(0)^2)$, where $f(0)$ is the density of the noise at its centre (Section 3.13.4, [CasellaBerger Ex. 10.2.3, p. 390]). The median cares only about how many readings fall just above and just below the centre, so what matters is how much probability sits there. For the three laboratories:

$$\begin{aligned}
&\text{A:}\quad n\operatorname{Var}(\mathrm{med}) \overset{(1)}{=} \frac{1}{4\,(1/\sqrt{2\pi})^2} = \frac{\pi}{2} = 1.571,\\
&\text{B:}\quad n\operatorname{Var}(\mathrm{med}) \overset{(2)}{=} \frac{1}{4\,(1/2b)^2} = b^2 = 0.500,\\
&\text{C:}\quad n\operatorname{Var}(\mathrm{med}) \overset{(3)}{=} \frac{1}{4f_C(0)^2} = 0.289, \qquad \text{with}\quad f_C(0) = \frac{1}{\sqrt{2\pi}}\left[\frac{1-\epsilon}{s} + \frac{\epsilon}{10s}\right] = 0.9293.
\end{aligned} \tag{T5.2}$$

(1) The standard Gaussian density at zero is $1/\sqrt{2\pi}$.
(2) The Laplace density at zero is $1/(2b)$, and $b^2 = \frac{1}{2}$.
(3) Each Gaussian component contributes its weight times its peak height $1/(\sqrt{2\pi}\,\text{width})$. The narrow core, of width $s = 0.410$, makes the peak tall.

The median loses to the mean in laboratory A (by the factor $\pi/2$ already met in Section 3.13.4), wins by a factor of two in laboratory B, and wins by a factor of about 3.5 in laboratory C.

The best one can do. The Cramér–Rao bound (Section 3.13.3) says that no unbiased estimator has n Var below $1/I$, where $I = \mathbb{E}[(\partial_\mu \ln f(x-\mu))^2]$ is the Fisher information of one reading. For laboratory A, $I = 1$. For laboratory B, $\partial_\mu \ln f_B(x-\mu) = \operatorname{sgn}(x-\mu)/b$, whose square is $1/b^2$ for every reading, so $I = 1/b^2 = 2$ and the bound is 0.500. For laboratory C the integral has to be done numerically; it gives $I = 5.386$ and the bound 0.186. So in laboratory C the plain mean is more than five times worse than what the data allow.

Example: in laboratory B the best estimator is the median. Maximum likelihood attains the Cramér–Rao bound for large n (Section 3.13.3), so let us find it for Laplace noise. The log likelihood of the n readings is

$$\ln \mathcal{L}(\mu) \overset{(1)}{=} \sum_{i=1}^{n} \ln\Big[\frac{1}{2b} e^{-|x_i-\mu|/b}\Big] \overset{(2)}{=} -n\ln(2b) - \frac{1}{b}\sum_{i=1}^{n}|x_i-\mu|,$$

$$\frac{\mathrm{d}\ln\mathcal{L}}{\mathrm{d}\mu} \overset{(3)}{=} \frac{1}{b}\sum_{i=1}^{n}\operatorname{sgn}(x_i-\mu) \overset{(4)}{=} \frac{1}{b}\big[\#\{x_i>\mu\} - \#\{x_i<\mu\}\big]. \tag{T5.3}$$

(1) Independent readings: the likelihood is a product, its log a sum (Eq. (3.78)).
(2) The log of each factor is $-\ln(2b) - |x_i-\mu|/b$.
(3) The slope of $|x_i-\mu|$ with respect to μ is -1 when $\mu < x_i$ and $+1$ when $\mu > x_i$, so $-|x_i-\mu|$ has slope $\operatorname{sgn}(x_i-\mu)$.
(4) Each reading above μ contributes $+1$, each reading below contributes -1.

The slope is positive while more readings lie above μ than below, and negative once more lie below. So the likelihood peaks where the counts balance: at the sample median (for odd n, exactly the middle reading). Maximising a Laplace likelihood is minimising $\sum_i |x_i-\mu|$, just as maximising a Gaussian likelihood is minimising $\sum_i (x_i-\mu)^2$. The two noise laws give two different "best averages", and the difference is entirely in how hard each reading is allowed to pull.

The numbers. The script `code/chT5/01_three_labs.py` runs 20000 experiments of $n = 25$ readings in each laboratory, with true $\mu = 0$, and computes the mean, the median and the maximum-likelihood estimate (found by a grid search around the median). It reports n times the mean squared error:

	mean	median	maximum likelihood	Cramér–Rao $1/I$
A: Gaussian	1.014	1.544	1.014	1.000
B: Laplace	1.008	0.673	0.672	0.500
C: 5% wild	1.005	0.291	0.191	0.186

The mean column is one in every row, as predicted. In laboratory A the maximum-likelihood estimate *is* the mean, and it sits on the bound. In laboratory B maximum likelihood is the median, as just derived; at $n = 25$ both still sit above the asymptotic value 0.500, and Section T5.2 shows that this approach to the limit is slow for Laplace noise. In laboratory C the median already gains a factor 3.46 over the mean, and maximum likelihood, which knows the exact mixture, comes within a few per cent of the bound. The right panel of Figure T5.2 shows the three sampling distributions in laboratory C: the mean's is wide, because now and then a wild reading drags it, while the median and maximum likelihood ignore the wild readings almost completely.

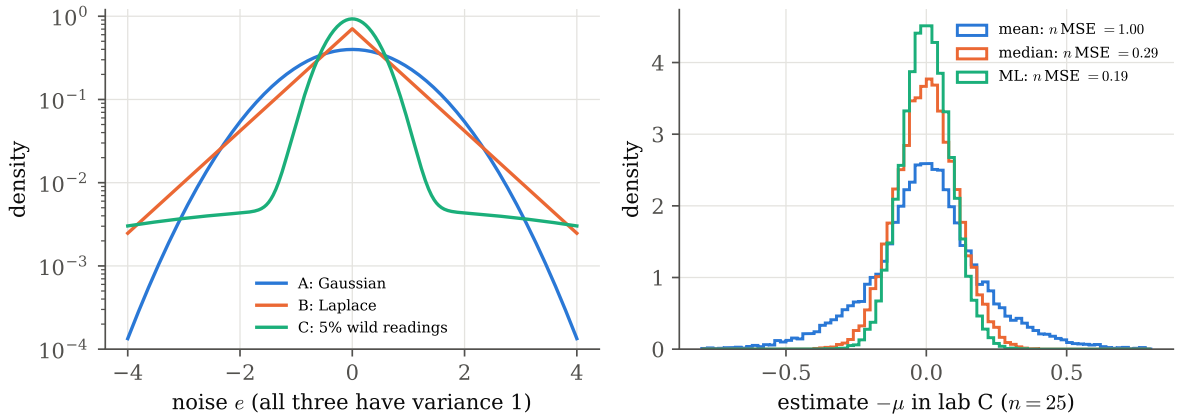


Figure T5.2: Left: the noise laws of the three laboratories, all of variance one, on a logarithmic scale. The Gaussian (A) has the thinnest tails; the Laplace law (B) falls in straight lines; laboratory C has a narrow core and a broad skirt made by the 5% of wild readings. Right: in laboratory C, the estimates of μ from 20000 experiments of 25 readings. The mean (blue) is about twice as wide as the median (orange) and the maximum-likelihood estimate (green).

A picture: how hard each reading pulls. Every estimator in this example can be written as the solution of one equation. Maximum likelihood sets the slope of the log likelihood to zero, $\sum_i \psi(x_i - \mu) = 0$ with $\psi(e) = -\mathrm{d}\ln f(e)/\mathrm{d}e$. Read $\psi(e)$ as the *pull* that a reading at distance e from the current guess exerts on the guess; the estimate is where the pulls from above and below balance. For Gaussian noise $-\ln f = \frac{1}{2}e^2 + \text{const}$, so $\psi(e) = e$: a reading pulls in proportion to its distance, like a spring. A reading ten units away pulls ten times as hard as one at one unit, and this is why one stopped watch can move the average by an hour. For Laplace noise $\psi(e) = \mathrm{sgn}(e)/b$: every reading pulls equally hard, whatever its distance, so only the counts above and below matter. For laboratory C the pull grows like a spring near the centre, then *falls back* for readings far out, because the mixture law "knows" that far readings are probably cosmic rays and carry little information about μ. Figure T5.3 draws the three pulls. Choosing an estimator is choosing how hard each reading may pull, and the best choice is the one that matches the law of the noise.

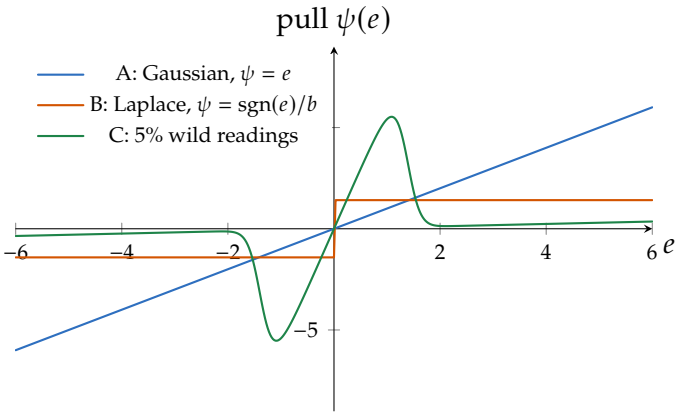


Figure T5.3: The pull $\psi(e) = -\mathrm{d}\ln f/\mathrm{d}e$ that a reading at distance e exerts on the estimate, for the three noise laws (all of variance one). Gaussian noise pulls like a spring, without limit. Laplace noise pulls with the same strength at every distance. The law of laboratory C pulls hard near the centre and then lets go: a reading more than about 1.3 units away is treated as a probable cosmic ray.

One idea, many names. Choosing a rule that turns data into an answer, by asking how much each possible error would cost us and which rule keeps the expected cost low, is called **statistical decision theory** [Wasserman Ch. 12]. The cost is the *loss*, its average over repeated experiments the *risk*, and the rule the estimator. The "balance of pulls" equation $\sum_i \psi(x_i - \mu) = 0$ is called an *estimating equation*, and estimators defined by one are *M-estimators* ([CasellaBerger §10.2.2, p. 391]): maximum likelihood is the case $\psi = -(\ln f)'$, least squares the case $\psi(e) = e$, the median the case $\psi = \text{sgn}$. The same idea gives the matched filter of a gravitational-wave search (the weights match the noise spectrum, Chapter 10) and the weighted mean (the weights match each reading's variance, Section 3.7.1). In every case the recipe is tuned to what we know about how the data go wrong.

What follows turns the three laboratories into a short list of questions, then places every estimator of the book on one table. The rest of the chapter builds the families that the earlier chapters did not.

T5.1.1 Six questions

Each question below is asked about the problem, not about the data already in hand, and each answer points to a family of estimators. Figure T5.1 at the start of the chapter draws them in order.

(i) What do we know about the noise? This is the question the three laboratories answered. If we know its full law, the likelihood is available and maximum likelihood is efficient for large samples (Section 3.13.3). If we know only the mean and the covariance of the noise, as is common when an instrument comes with a noise covariance matrix but no promise that the noise is Gaussian, the natural recipe is the best *linear* unbiased estimator of Section T5.2, which needs nothing else. If we know that most readings are well behaved but some may be wild, we need a *robust* estimator (Section T5.5). If we know only a few moments of the data as functions of the parameter, the method of moments matches them (Section 3.9.2, [Cowan Ch. 8, pp. 114–117]). In laboratory A all of these coincide with the mean; in laboratory C they differ by a factor of five.

(ii) Do we have prior information worth using? If earlier experiments or physics already tell us roughly where μ lies, a Bayes estimator uses it, and its quality depends on whether the prior is right. Near the prior guess it beats every unbiased estimator; far from it, it is worse. Section T5.3 computes both sides of this trade, and the linear–Gaussian case becomes the Wiener filter.

(iii) Where does the parameter live? In the three laboratories μ shifts every reading: it lives in the *mean* of the data, and an estimator linear in the data can find it. Sometimes the parameter changes how much the data *scatter*: the amplitude of the sky fluctuations, the level of a noise, a power spectrum. Then the mean of the data is zero whatever the parameter, so no linear combination of the readings can see it, and the estimator must be *quadratic* in the data (Section T5.6).

(iv) How many parameters at once? One number is estimated well by its own data. When many similar numbers are estimated together, the calibrations of a hundred detectors, the band powers of a spectrum, each one can borrow strength from the others, and shrinking all estimates towards a common value lowers the total error (Section T5.4).

(v) Do the data arrive all at once or in a stream? If readings keep coming, we would rather update the estimate than recompute it from scratch. Recursive estimators do this; the second half of the chapter builds them.

(vi) Does the best estimator depend on what it estimates? The best weights for a power spectrum depend on the spectrum; the best weights for a robust fit depend on which readings are outliers, which depends on the fit. Such estimators are found by iterating, also in the second half of the chapter.

Example: three questions for one problem. A CCD camera takes n dark frames to measure the dark current of one pixel, the rate μ at which it collects charge with the shutter closed. (i) The noise of each frame is read-out noise plus Poisson noise of the dark charge, nearly Gaussian, but cosmic rays hit a few per cent of the frames: we are in laboratory C, and a robust estimator is called for. (ii) The manufacturer quotes a typical dark current, but it varies from pixel to pixel by more than our measurement error, so it is a weak prior at best. (iv) We are not measuring one pixel but four million at once, and they come from the same manufacturing process: shrinking each pixel's estimate towards the chip-wide average lowers the total error. The "best" estimator for this pixel is therefore a robust estimate, shrunk towards the chip average. No single textbook recipe gives it, but each ingredient answers one question.

T5.1.2 Every estimator in one picture

The estimators met so far, and those of this chapter, differ in three things: what they *assume* about the data, what they *optimise*, and therefore *when* they are the right tool. Section T5.1.2 lists them side by side. Reading down the "assumes" column shows a ladder: the more we know about the noise, the more efficient the estimator that knowledge allows, and the more it costs when the knowledge is wrong.

Two features of the table are worth saying in words. First, almost every row "optimises" something different, so there is no single best estimator: there is a best estimator *for a stated criterion under stated assumptions*. Section 3.2.3 already showed that even for one Gaussian mean no estimator has the lowest mean squared error at every true value. Second, the rows are not rivals. In laboratory A the weighted mean, least squares, the BLUE, maximum likelihood and the minimum-variance unbiased estimator are *one and the same estimator*, the sample mean. They come apart only when the noise is not Gaussian, when a prior is used, or when the parameter moves from the mean into the covariance. Those departures are what the rest of the chapter is about.

Watch out: Inefficient is not wrong

Under the heavy-tailed noise of laboratory C the sample mean is still unbiased, and its error bar $\sigma/\sqrt{n}$ is still correct. What it loses is *efficiency*: it needs about five times as many readings as maximum likelihood to reach the same precision. A badly chosen estimator rarely gives a wrong answer; it gives an honest but needlessly large error bar. The exception is a lopsided contamination, cosmic rays that only ever add charge, which biases the mean as well (Section T5.5).

T5.2 Linear estimators and the Gauss–Markov theorem

What we already have. For readings of one quantity with known variances, Section 3.7.1 found the weights that make a weighted average scatter least: each reading is weighted by

Taxonomy: Which estimator assumes what, and optimises what

estimator	assumes	optimises / guarantees	where
method of moments	a few moments as functions of θ	consistency; simple, often a starting point	Section 3.9.2
maximum likelihood	the full law of the data	efficient for large n (reaches Cramér–Rao)	Section 3.9
least squares	Gaussian noise, or only mean and covariance	ML for Gaussian; BLUE for any noise	Section 3.10
weighted mean	independent readings, known variances	minimum variance among weighted averages	Section 3.7.1
linear (BLUE)	$\boldsymbol{d} = \mathbf{A}\boldsymbol{\theta} + \boldsymbol{n}$, known Cov $\boldsymbol{n}$	minimum variance among linear unbiased	Section T5.2
minimum-variance unbiased	full law, a complete sufficient statistic	minimum variance among all unbiased	Section 3.16.5
Bayes (mean, median, mode)	a prior and a loss	lowest average risk over the prior	Sections 3.16.4 and T5.3
Wiener filter, ridge	Gaussian signal prior and noise	minimum mean squared error	Section T5.3.3
shrinkage (James–Stein)	many similar means, Gaussian noise	total mean squared error below ML for $k \geq 3$	Section T5.4
robust (M, Huber, median)	most readings well behaved	bounded influence; efficiency near ML	Section T5.5
quadratic (QML)	θ in the covariance, Gaussian data	minimum variance unbiased quadratic	Section T5.6
iterative	best estimator needs θ	converges to the optimum	second half
recursive	data in a stream	same answer, updated reading by reading	second half

its inverse variance. Section 3.7.4 let the readings be correlated, and Section 3.10.2 fitted several parameters at once by least squares and found the covariance of the result, $(\mathbf{A}^\mathsf{T}\mathbf{C}^{-1}\mathbf{A})^{-1}$ (Eq. (3.88)). Along the way Section 3.7.4 mentioned that these answers are instances of a theorem named after Gauss and Markov, without proving it.

When that is not enough. Two questions were left open. First, least squares was derived as maximum likelihood for *Gaussian* noise. Is it still a good recipe when the noise is not Gaussian, or when we do not know its law? Second, with several parameters and correlated noise, could some other combination of the readings, one we did not think of, do better? The theorem answers both, and its proof is short enough to do in full.

A story: three weighings of two letters. A kitchen scale has a random error of about a gram. We want the weights of two letters, so we weigh the first, then the second, and then, almost by accident, both together. What is our best estimate of the first letter? The obvious answer is the first weighing. But the third weighing also knows about the first letter: subtracting the second weighing from it gives a second, independent estimate. Combining all three readings, with the right signs and weights, beats the obvious answer. Nothing about the scale's error needs to be Gaussian for this to work; it only needs to average to zero and to have a known spread.

The same situation is everywhere in physics. A CMB telescope sweeps across the sky and each sample of the time stream sees one pixel: the pixel values are the unknowns, the time stream is the data, and every pixel is seen many times along different scans, with noise

correlated in time. Turning the time stream into a map is the weighing problem with millions of letters. The same form appears in a straight-line fit, in calibrating detector gains from single and coincident hits, and in fitting foreground templates. In each case the expected readings are *linear* in the unknowns. To find the minimum-variance linear unbiased estimator, we need to know the noise covariance.

The story in symbols. Call the two weights m_1, m_2 and the three readings $d_1 = m_1 + n_1$, $d_2 = m_2 + n_2$, $d_3 = m_1 + m_2 + n_3$, with independent errors of mean zero and variance σ^2. In matrix form

$$\boldsymbol{d} = \mathbf{A}\boldsymbol{\theta} + \boldsymbol{n}, \qquad \boldsymbol{d} = \begin{pmatrix} d_1 \\ d_2 \\ d_3 \end{pmatrix}, \quad \mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}, \quad \boldsymbol{\theta} = \begin{pmatrix} m_1 \\ m_2 \end{pmatrix}, \quad \operatorname{Cov} \boldsymbol{n} = \sigma^2 \mathbb{1}. \tag{T5.4}$$

A linear estimate of m_1 is $\hat{m}_1 = w_1 d_1 + w_2 d_2 + w_3 d_3$, with three dials. Its mean is $w_1 m_1 + w_2 m_2 + w_3(m_1 + m_2) = (w_1 + w_3)m_1 + (w_2 + w_3)m_2$, so it is unbiased for every pair of weights exactly when $w_1 + w_3 = 1$ and $w_2 + w_3 = 0$. That leaves one free dial, say $w_3 = t$, giving $w_1 = 1 - t$, $w_2 = -t$, and variance

$$\operatorname{Var} \hat{m}_1 \overset{(1)}{=} \sigma^2\left[(1-t)^2 + t^2 + t^2\right] \overset{(2)}{=} \sigma^2\left[1 - 2t + 3t^2\right] \overset{(3)}{=} \sigma^2\left[3\left(t - \tfrac{1}{3}\right)^2 + \tfrac{2}{3}\right]. \tag{T5.5}$$

(1) Independent errors: the variance of a weighted sum is $\sigma^2 \sum_i w_i^2$ (Eq. (3.19)).
(2) Expand $(1-t)^2 = 1 - 2t + t^2$ and add.
(3) Complete the square: $3t^2 - 2t + 1 = 3(t^2 - \frac{2}{3}t + \frac{1}{9}) + 1 - \frac{1}{3}$.

The naive choice $t = 0$ (use d_1 alone) gives σ^2. The choice $t = 1$ (use $d_3 - d_2$) gives $2\sigma^2$. The minimum is at $t = \frac{1}{3}$:

$$\hat{m}_1 = \tfrac{1}{3}(2d_1 - d_2 + d_3), \qquad \operatorname{Var} \hat{m}_1 = \tfrac{2}{3}\sigma^2, \tag{T5.6}$$

and by symmetry $\hat{m}_2 = \frac{1}{3}(-d_1 + 2d_2 + d_3)$. The price of the gain is a correlation: both estimates use d_3, and $\operatorname{Cov}(\hat{m}_1, \hat{m}_2) = \frac{\sigma^2}{9}[2 \cdot (-1) + (-1) \cdot 2 + 1 \cdot 1] = -\frac{\sigma^2}{3}$, a correlation coefficient of $-\frac{1}{2}$. Lesson: a reading that involves several unknowns informs all of them, and the best linear recipe uses every reading, with weights that can be negative.

A picture: the best weights touch the smallest ellipse. Strip the problem to two readings of one quantity μ, with variances 1 and 4. The weights (w_1, w_2) form a plane. Unbiasedness, $w_1 + w_2 = 1$, is a straight line in that plane. The variance $w_1^2 + 4w_2^2$ is constant on ellipses centred at the origin, twice as long in w_1 as in w_2. Moving along the line, we cross ellipses of different sizes; the best weights are where the line just touches the smallest ellipse it can reach (Figure T5.4). At that point the line is tangent to the ellipse, so the ellipse's gradient $(2w_1, 8w_2)$ is perpendicular to the line, that is parallel to $(1, 1)$: $2w_1 = 8w_2$, so $w_1 = 4w_2$, and with $w_1 + w_2 = 1$ the weights are $(0.8, 0.2)$, the inverse-variance weights of Eq. (3.61). The Gauss–Markov theorem is this picture in any number of dimensions: a flat set of allowed weights, a family of nested ellipsoids of equal variance, and one point of contact.

One idea, many names. The best linear recipe goes by many names. Statisticians call it **generalised least squares** or the **best linear unbiased estimator** (BLUE); in regression it is the least-squares solution $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\boldsymbol{Y}$ [Wasserman Thm 13.13, p. 217]; CMB experiments call it the map-making equation; the internal linear combination that cleans the CMB of Galactic dust (Section 3.7.4) is the same formula with one unknown; and the Kalman filter of the second half of the chapter applies it one reading at a time. Linear algebra is the natural language because

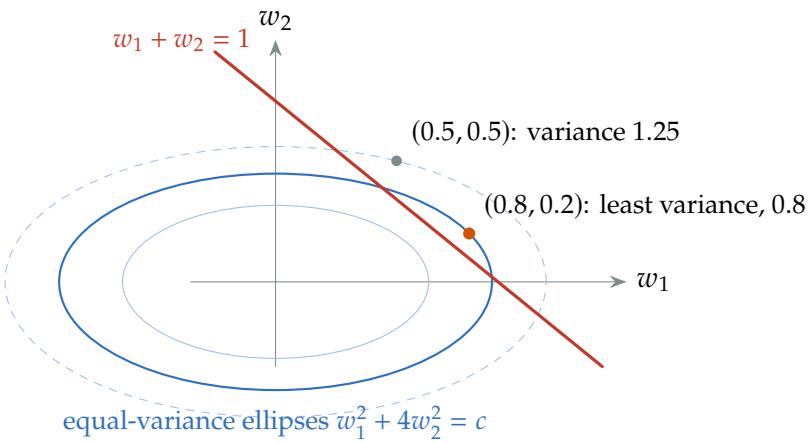


Figure T5.4: Two readings of one quantity, with variances 1 and 4. Every pair of weights on the red line gives an unbiased average. The blue ellipses are sets of weights with equal variance ($c = 0.4, 0.8$ and 1.25, from the inside out). The smallest ellipse the line reaches touches it at the inverse-variance weights $(0.8, 0.2)$, with variance 0.8. Equal weights (grey dot) lie on a larger ellipse, with variance 1.25.

the data, the unknowns and the weights are all lists of numbers, and "unbiased" and "least variance" turn into a linear constraint and a quadratic form.

We first set up the general linear model and prove that one recipe beats every other linear unbiased one, then ask what it needs from the noise, and finally what it cannot do when the noise has heavy tails.

T5.2.1 The linear model and its best linear unbiased estimator

Working definition T5.1 (linear model, BLUE)

A **linear model** writes N readings as $\boldsymbol{d} = \mathbf{A}\boldsymbol{\theta} + \boldsymbol{n}$, where $\boldsymbol{\theta}$ holds m unknown parameters, $\mathbf{A}$ is a known $N \times m$ matrix [column j says how strongly parameter j enters each reading] with independent columns, and the noise $\boldsymbol{n}$ has mean zero and a known, positive-definite covariance $\mathbf{C}$. A **linear estimator** is $\hat{\boldsymbol{\theta}} = \mathbf{W}\boldsymbol{d}$ for a fixed $m \times N$ matrix $\mathbf{W}$. It is unbiased if $\mathbb{E}\,\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$ for every $\boldsymbol{\theta}$. The **best linear unbiased estimator** is the unbiased $\mathbf{W}$ for which every combination $\boldsymbol{u}^{\mathsf{T}}\hat{\boldsymbol{\theta}}$ has the smallest possible variance.

The noise here needs only two properties, a mean of zero and a known covariance. No distribution has been assumed.

What does unbiased require? The mean of a linear estimator is

$$\mathbb{E}\,\hat{\boldsymbol{\theta}} \overset{(1)}{=} \mathbf{W}\,\mathbb{E}\,\boldsymbol{d} \overset{(2)}{=} \mathbf{W}\mathbf{A}\boldsymbol{\theta}, \tag{T5.7}$$

(1) Expectation is linear and $\mathbf{W}$ is a fixed matrix.
(2) $\mathbb{E}\,\boldsymbol{d} = \mathbf{A}\boldsymbol{\theta} + \mathbb{E}\,\boldsymbol{n}$ and $\mathbb{E}\,\boldsymbol{n} = \mathbf{0}$.

and $\mathbf{W}\mathbf{A}\boldsymbol{\theta} = \boldsymbol{\theta}$ for *every* $\boldsymbol{\theta}$ means $\mathbf{W}\mathbf{A} = \mathbb{1}_m$: taking $\boldsymbol{\theta}$ to be each unit vector in turn reads off the columns of $\mathbf{W}\mathbf{A}$ one at a time. This is the matrix version of $w_1 + w_3 = 1$, $w_2 + w_3 = 0$ in the weighing story.

How much does it scatter? The covariance of a linear map of the data is the map applied on both sides (Eq. (3.47)): $\operatorname{Cov}\hat{\boldsymbol{\theta}} = \mathbf{W}\mathbf{C}\mathbf{W}^{\mathsf{T}}$.

A candidate. Least squares with weight $\mathbf{C}^{-1}$ gave $\mathbf{W}_0 = (\mathbf{A}^\mathsf{T}\mathbf{C}^{-1}\mathbf{A})^{-1}\mathbf{A}^\mathsf{T}\mathbf{C}^{-1}$ (Eq. (3.87)). It is unbiased, since $\mathbf{W}_0\mathbf{A} = (\mathbf{A}^\mathsf{T}\mathbf{C}^{-1}\mathbf{A})^{-1}(\mathbf{A}^\mathsf{T}\mathbf{C}^{-1}\mathbf{A}) = \mathbb{1}$, and its covariance is

$$\mathbf{W}_0\mathbf{C}\mathbf{W}_0^\mathsf{T} \overset{(1)}{=} (\mathbf{A}^\mathsf{T}\mathbf{C}^{-1}\mathbf{A})^{-1}\mathbf{A}^\mathsf{T}\mathbf{C}^{-1}\,\mathbf{C}\,\mathbf{C}^{-1}\mathbf{A}\,(\mathbf{A}^\mathsf{T}\mathbf{C}^{-1}\mathbf{A})^{-1} \overset{(2)}{=} (\mathbf{A}^\mathsf{T}\mathbf{C}^{-1}\mathbf{A})^{-1}. \tag{T5.8}$$

(1) Insert $\mathbf{W}_0$; its transpose is $\mathbf{C}^{-1}\mathbf{A}(\mathbf{A}^\mathsf{T}\mathbf{C}^{-1}\mathbf{A})^{-1}$ because $\mathbf{C}$ and $\mathbf{A}^\mathsf{T}\mathbf{C}^{-1}\mathbf{A}$ are symmetric.
(2) $\mathbf{C}^{-1}\mathbf{C}\mathbf{C}^{-1} = \mathbf{C}^{-1}$, then one factor $\mathbf{A}^\mathsf{T}\mathbf{C}^{-1}\mathbf{A}$ cancels against its inverse.

Any other unbiased recipe. Take any $\mathbf{W}$ with $\mathbf{W}\mathbf{A} = \mathbb{1}$ and write it as the candidate plus a difference, $\mathbf{W} = \mathbf{W}_0 + \mathbf{D}$. Both satisfy the constraint, so $\mathbf{D}\mathbf{A} = \mathbf{W}\mathbf{A} - \mathbf{W}_0\mathbf{A} = \mathbf{0}$. Then

$$\begin{aligned}\mathbf{W}\mathbf{C}\mathbf{W}^\mathsf{T} &\overset{(1)}{=} \mathbf{W}_0\mathbf{C}\mathbf{W}_0^\mathsf{T} + \mathbf{W}_0\mathbf{C}\mathbf{D}^\mathsf{T} + \mathbf{D}\mathbf{C}\mathbf{W}_0^\mathsf{T} + \mathbf{D}\mathbf{C}\mathbf{D}^\mathsf{T} \\ &\overset{(2)}{=} \mathbf{W}_0\mathbf{C}\mathbf{W}_0^\mathsf{T} + (\mathbf{A}^\mathsf{T}\mathbf{C}^{-1}\mathbf{A})^{-1}(\mathbf{D}\mathbf{A})^\mathsf{T} + (\mathbf{D}\mathbf{A})(\mathbf{A}^\mathsf{T}\mathbf{C}^{-1}\mathbf{A})^{-1} + \mathbf{D}\mathbf{C}\mathbf{D}^\mathsf{T} \\ &\overset{(3)}{=} (\mathbf{A}^\mathsf{T}\mathbf{C}^{-1}\mathbf{A})^{-1} + \mathbf{D}\mathbf{C}\mathbf{D}^\mathsf{T}.\end{aligned} \tag{T5.9}$$

(1) Multiply out $(\mathbf{W}_0 + \mathbf{D})\mathbf{C}(\mathbf{W}_0 + \mathbf{D})^\mathsf{T}$.
(2) In the cross term, $\mathbf{W}_0\mathbf{C} = (\mathbf{A}^\mathsf{T}\mathbf{C}^{-1}\mathbf{A})^{-1}\mathbf{A}^\mathsf{T}\mathbf{C}^{-1}\mathbf{C} = (\mathbf{A}^\mathsf{T}\mathbf{C}^{-1}\mathbf{A})^{-1}\mathbf{A}^\mathsf{T}$, and $\mathbf{A}^\mathsf{T}\mathbf{D}^\mathsf{T} = (\mathbf{D}\mathbf{A})^\mathsf{T}$. The other cross term is its transpose.
(3) $\mathbf{D}\mathbf{A} = \mathbf{0}$ kills both cross terms; the first term is (T5.8).

The extra piece $\mathbf{D}\mathbf{C}\mathbf{D}^\mathsf{T}$ can only add variance. For any combination $\boldsymbol{u}^\mathsf{T}\hat{\boldsymbol{\theta}}$ of the estimates, its contribution is $\boldsymbol{u}^\mathsf{T}\mathbf{D}\mathbf{C}\mathbf{D}^\mathsf{T}\boldsymbol{u} = \boldsymbol{v}^\mathsf{T}\mathbf{C}\boldsymbol{v}$ with $\boldsymbol{v} = \mathbf{D}^\mathsf{T}\boldsymbol{u}$, and this is non-negative because $\mathbf{C}$ is a covariance (the variance of the number $\boldsymbol{v}^\mathsf{T}\boldsymbol{n}$). It is zero for every $\boldsymbol{u}$ only if $\mathbf{D} = \mathbf{0}$, since $\mathbf{C}$ is positive definite. So no linear unbiased recipe can beat $\mathbf{W}_0$, and only $\mathbf{W}_0$ itself can match it.

Key idea: Gauss–Markov theorem

In the linear model $\boldsymbol{d} = \mathbf{A}\boldsymbol{\theta} + \boldsymbol{n}$ with $\mathbb{E}\,\boldsymbol{n} = \mathbf{0}$ and $\operatorname{Cov}\boldsymbol{n} = \mathbf{C}$, the estimator $\hat{\boldsymbol{\theta}} = (\mathbf{A}^\mathsf{T}\mathbf{C}^{-1}\mathbf{A})^{-1}\mathbf{A}^\mathsf{T}\mathbf{C}^{-1}\boldsymbol{d}$ is unbiased, has covariance $(\mathbf{A}^\mathsf{T}\mathbf{C}^{-1}\mathbf{A})^{-1}$, and every other linear unbiased estimator has covariance larger by a positive semidefinite matrix. It is the unique best linear unbiased estimator, whatever the law of the noise.

The proof used nothing but $\mathbb{E}\,\boldsymbol{n} = \mathbf{0}$ and $\operatorname{Cov}\boldsymbol{n} = \mathbf{C}$. That is the content of the remark in [Cowan §7.2, p. 97] that the least-squares estimators have zero bias and minimum variance "regardless of" the distributions of the measurements; [CasellaBerger §11.3.2, pp. 440–443] proves the one-regressor case and names the theorem.

Example: the weighing design, through the theorem. For (T5.4), $\mathbf{A}^\mathsf{T}\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$, whose inverse is $\frac{1}{3}\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$, and with $\mathbf{C} = \sigma^2\mathbb{1}$ the factors of σ^2 cancel in $\mathbf{W}_0$:

$$\mathbf{W}_0 = (\mathbf{A}^\mathsf{T}\mathbf{A})^{-1}\mathbf{A}^\mathsf{T} \overset{(1)}{=} \frac{1}{3}\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \overset{(2)}{=} \frac{1}{3}\begin{pmatrix} 2 & -1 & 1 \\ -1 & 2 & 1 \end{pmatrix}, \qquad \operatorname{Cov}\hat{\boldsymbol{\theta}} \overset{(3)}{=} \frac{\sigma^2}{3}\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}.$$

(1) $\mathbf{A}^\mathsf{T}$ is the transpose of (T5.4).
(2) Row times column: e.g. the top-right entry is $2\cdot 1 + (-1)\cdot 1 = 1$.
(3) (T5.8) with $\mathbf{C} = \sigma^2\mathbb{1}$: $\sigma^2(\mathbf{A}^\mathsf{T}\mathbf{A})^{-1}$.

The first row is (T5.6), now obtained without any search. The decomposition (T5.9) also accounts exactly for the worse recipes. "Use d_1 alone" is the row $(1,0,0)$, so its $\mathbf{D}$-row is $(1,0,0) - \frac{1}{3}(2,-1,1) = \frac{1}{3}(1,1,-1)$; it satisfies $\mathbf{D}\mathbf{A} = \mathbf{0}$ (check: $\frac{1}{3}(1+0-1) = 0$ and $\frac{1}{3}(0+1-1) = 0$),

and adds $\sigma^2 \cdot \frac{1+1+1}{9} = \frac{\sigma^2}{3}$ to the variance, giving $\frac{2}{3}\sigma^2 + \frac{1}{3}\sigma^2 = \sigma^2$. "Use $d_3 - d_2$" is $(0, -1, 1)$, its $\mathbf{D}$-row is $\frac{1}{3}(-2, -2, 2)$, which adds $\sigma^2 \cdot \frac{12}{9} = \frac{4}{3}\sigma^2$, giving $2\sigma^2$. Both agree with (T5.5) at $t = 0$ and $t = 1$. The script `code/chT5/02_blue_heavy_tails.py` simulates 200000 weighings with $\sigma = 0.1$ and finds variances $0.6716\,\sigma^2$, $1.0024\,\sigma^2$ and $2.0145\,\sigma^2$ for the three recipes, and a correlation -0.5007 between $\hat{m}_1$ and $\hat{m}_2$.

Example: one quantity, the weighted mean again. If all N readings measure the same μ, then $\mathbf{A}$ is a column of ones, $\mathbf{1}$, and $\boldsymbol{\theta} = \mu$. With independent readings, $\mathbf{C} = \mathrm{diag}(\sigma_1^2, \dots, \sigma_N^2)$, the theorem gives

$$\hat{\mu} \overset{(1)}{=} \frac{\mathbf{1}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{d}}{\mathbf{1}^{\mathsf{T}}\mathbf{C}^{-1}\mathbf{1}} \overset{(2)}{=} \frac{\sum_k d_k/\sigma_k^2}{\sum_k 1/\sigma_k^2}, \qquad \mathrm{Var}\,\hat{\mu} \overset{(3)}{=} \left(\sum_k \frac{1}{\sigma_k^2}\right)^{-1}, \tag{T5.10}$$

(1) $\mathbf{A}^{\mathsf{T}}\mathbf{C}^{-1}\mathbf{A} = \mathbf{1}^{\mathsf{T}}\mathbf{C}^{-1}\mathbf{1}$ is a number, so its inverse is a division.
(2) $\mathbf{C}^{-1}$ is diagonal with entries $1/\sigma_k^2$.
(3) (T5.8).

which is the inverse-variance weighted mean of Eq. (3.61), and with a full $\mathbf{C}$ the correlated version Eq. (3.68). The Lagrange-multiplier argument of Section 3.7.1 was the case $m = 1$ of the proof above.

T5.2.2 What the theorem needs, and what it gives

For Gaussian noise it is the best estimator of all. If $\boldsymbol{n}$ is Gaussian, the BLUE is also the maximum-likelihood estimator (Section 3.10.1), and the Fisher matrix of the model is $\mathbf{F} = \mathbf{A}^{\mathsf{T}}\mathbf{C}^{-1}\mathbf{A}$: Eq. (10.20) with a mean $\mathbf{A}\boldsymbol{\theta}$ whose derivative with respect to θ_j is column j of $\mathbf{A}$, and a covariance that does not depend on $\boldsymbol{\theta}$. So the BLUE covariance (T5.8) equals $\mathbf{F}^{-1}$, the Cramér–Rao bound (Eq. (10.10)). No unbiased estimator at all, linear or not, does better: for Gaussian noise the BLUE is the minimum-variance unbiased estimator.

For other noise it is the best only among linear recipes. The restriction to linear estimators is a real restriction. The median is not a linear function of the readings (the reading it returns depends on the order of all of them), so Gauss–Markov says nothing about it, and the three laboratories of Section T5.1 showed the median beating the mean, the BLUE of a common level, by a factor of two or more. Gauss–Markov guarantees that the BLUE is never *bad* when the covariance is right; it does not guarantee that it is good.

It needs the covariance to exist. For Cauchy noise, such as the lighthouse flashes of [Sivia §2.4], the variance of a single reading is infinite, so $\mathbf{C}$ does not exist and neither does the BLUE. The mean of n Cauchy readings has the same Cauchy law as one reading (Section 3.3.4), and averaging gains nothing.

Example: the Fisher information of Cauchy noise. To see how much better than the median one can do for Cauchy noise, we need the Cramér–Rao bound. With $f(e) = \frac{1}{\pi(1+e^2)}$ and $e = x - \mu$,

$$I \overset{(1)}{=} \int_{-\infty}^{\infty} \left(\frac{2e}{1+e^2}\right)^2 \frac{\mathrm{d}e}{\pi(1+e^2)} \overset{(2)}{=} \frac{4}{\pi}\int_{-\pi/2}^{\pi/2} \frac{\tan^2 t}{\sec^6 t}\,\sec^2 t\,\mathrm{d}t \overset{(3)}{=} \frac{4}{\pi}\int_{-\pi/2}^{\pi/2} \sin^2 t\cos^2 t\,\mathrm{d}t$$
$$\overset{(4)}{=} \frac{4}{\pi}\int_{-\pi/2}^{\pi/2} \frac{\sin^2 2t}{4}\,\mathrm{d}t \overset{(5)}{=} \frac{4}{\pi}\cdot\frac{\pi}{8} = \frac{1}{2}. \tag{T5.11}$$

(1) $I = \mathbb{E}[(\partial_\mu \ln f)^2]$, and $\partial_\mu \ln f(x-\mu) = \partial_\mu[-\ln(1+(x-\mu)^2)] = 2e/(1+e^2)$.

(2) Substitute $e = \tan t$: $1 + e^2 = \sec^2 t$ and $\mathrm{d}e = \sec^2 t\,\mathrm{d}t$; e from $-\infty$ to ∞ is t from $-\pi/2$ to $\pi/2$.
(3) $\tan^2 t / \sec^4 t = \sin^2 t \cos^2 t$.
(4) $\sin t \cos t = \frac{1}{2} \sin 2t$.
(5) $\sin^2 2t$ averages to $\frac{1}{2}$ over the interval of length π, so the integral of $\sin^2 2t$ is $\pi/2$, and divided by four, $\pi/8$.

So no unbiased estimator of a Cauchy centre has n Var below $1/I = 2$. The median reaches $\pi^2/4 = 2.467$ (Section 3.13.4), so it wastes about a fifth of the information; maximum likelihood reaches 2 for large n.

T5.2.3 When linear is not enough: heavy tails

How quickly do these large-n statements take hold? We repeat the comparison of the three laboratories for n from 5 to 101, with 10000 experiments at each n, for Laplace noise, Cauchy noise and the contaminated Gaussian of laboratory C. This is the second part of the script `02_blue_heavy_tails.py`. The maximum-likelihood estimate is found by evaluating the log likelihood on a grid of candidate values around the median and refining the best grid point with a parabola, a direct search with no iteration.

Listing T5.1: Mean, median and maximum likelihood for one noise law and one n (`code/chT5/02_blue_heavy_tails.py`)

```
66      rows = []
67      for n in ns:
68          x = L["draw"]((Mb, n))
69          half = 6 / np.sqrt(n) + (2.0 if name == "Cauchy" else 0.3)
70          ml = ml_location(x, L["logf"], half=half, ngrid=481)
71          rows.append([n * np.mean(x.mean(1)**2), n * np.mean(np.median(x, 1)**2), n *
    np.mean(ml**2)])
72      out[name] = np.array(rows)
```

Line 68 draws 10000 data sets of n readings around $\mu = 0$. Line 69 sets the half-width of the grid: wide for small n, and wider for the Cauchy law, whose likelihood can have several peaks. Line 70 finds the maximum-likelihood estimates, and line 71 records n times the mean squared error of the three estimators. Figure T5.5 shows the result.

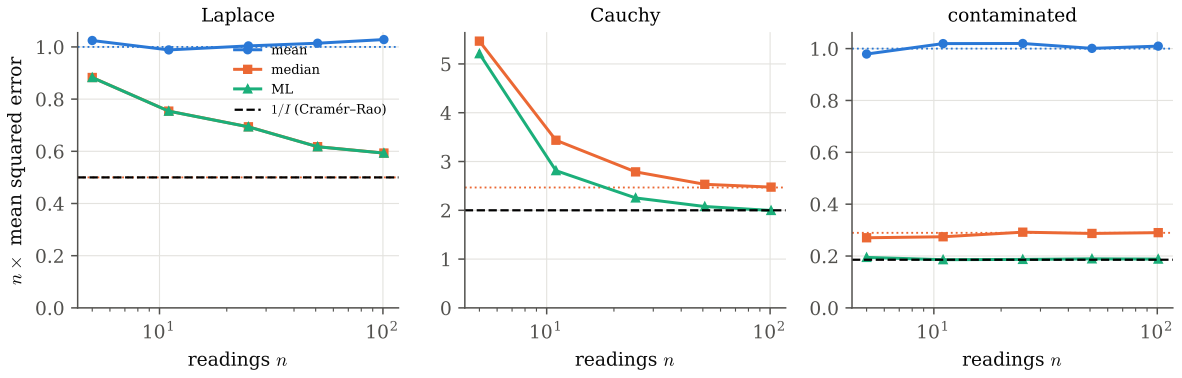


Figure T5.5: n times the mean squared error of three estimators of a centre, against the number of readings n, for three heavy-tailed noise laws of unit scale. Dotted lines: the large-n values of the mean (blue, 1) and the median (orange, $1/4f(0)^2$). Dashed black: the Cramér–Rao value $1/I$. The mean is absent for Cauchy noise, where its error is infinite. Maximum likelihood (green) approaches the bound in every case; for Laplace noise it is the median, and both approach their limit slowly.

Three things stand out. For the contaminated Gaussian, every curve is flat from $n = 5$ on: the large-n formulas already hold, and maximum likelihood sits at 0.188 against the bound 0.186, five times better than the mean (1.010). For Cauchy noise the small-sample penalty is large: at $n = 5$ the median and maximum likelihood have n MSE = 5.467 and 5.210, and only at $n = 101$ do they reach 2.475 and 1.997, close to $\pi^2/4$ and 2. For Laplace noise the median (which is maximum likelihood) is still at 0.593 for $n = 101$, above the asymptotic 0.5, and the excess only roughly halves each time n grows fourfold: the large-n formula is a statement about a limit, and how soon it applies depends on the noise law. In all three cases the mean, the BLUE of a common level, is the worst of the three whenever it exists.

Watch out: The BLUE needs the right covariance, not just any covariance

Gauss–Markov ranks recipes by their variance under the *true* covariance $\mathbf{C}$. Weights built from a wrong $\mathbf{C}$, for example correlated noise treated as independent, still give an unbiased estimate, but its variance is larger than (T5.8), and the error bar computed from the wrong $\mathbf{C}$ is wrong as well (Section 3.10.5 shows both effects).

Where this leads. Three ways out of the linear world follow from the three laboratories. If the full law of the noise is known, maximum likelihood uses it (Section 3.9). If only "most readings are well behaved" is known, a robust estimator limits the pull of any one reading (Section T5.5). And if we are willing to give up unbiasedness itself, a prior or a shrinkage rule can lower the mean squared error below anything Gauss–Markov allows, which is where the next two sections go.

T5.3 Bayes estimators under repeated experiments

What we already have. A posterior distribution is a whole curve, and to report one number we must say what kind of mistake we fear. Section 3.16.4 showed that the posterior mean minimises the expected squared error, the posterior median the expected absolute error, and the posterior mode a zero–one loss (Eq. (3.114), [Wasserman Thm 12.8, p. 198]). For Gaussian readings with a Gaussian prior the posterior mean is a weighted average of the data's mean and the prior's mean (Eq. (5.32)). Those statements average over the *parameter*, with the data held fixed at what we observed.

When that is not enough. A colleague who does not share our prior will ask a different question: "Suppose the true value is θ. If the experiment were repeated many times, how far would your recipe land from θ, on average?" That average is over the *data*, with the parameter held fixed. It is the question we asked of every estimator in Chapter 3, and a Bayes estimator can be asked it too. The answer is not "always better" or "always worse"; it depends on where the truth lies relative to the prior, and knowing how is what lets us decide whether to use a prior at all.

A story with the whole trick in it. Two weather forecasters predict tomorrow's noon temperature. The first reads only today's sky and instruments, which are noisy. The second does the same and then pulls the forecast part of the way towards the long-term average for that day of the year. On ordinary days, which are most days, the second forecaster is more accurate: the pull cancels part of the noise in the instruments. On the day of a freak heatwave, the second forecaster is systematically too cold, because the pull drags the forecast towards a normal day that tomorrow is not. Whether the pull is worth it depends on how often the weather is unusual, and by how much.

Physics data behave the same way. A noisy map of the microwave sky, filtered so that each feature is pulled towards the zero that a prior expects on small scales, looks cleaner and is closer to the true sky on average, but every hot spot in it is fainter than it really is. An analysis of a rare decay that pulls a branching ratio towards the Standard Model value will be more accurate if the Standard Model is right, and will hide new physics if it is not.

The story in symbols. Let the data reduce to an average $\bar{x} \sim \mathcal{N}(\theta, s^2)$, where s is its standard error, and let the prior be $\theta \sim \mathcal{N}(\mu_0, \tau^2)$. The posterior mean is (Eq. (5.32), [Wasserman Ex. 12.9, p. 198])

$$\hat{\theta}_B = w\,\bar{x} + (1-w)\,\mu_0, \qquad w = \frac{\tau^2}{\tau^2 + s^2}, \tag{T5.12}$$

a pull of the data towards μ_0 that is strong when the prior is narrow ($\tau \ll s$) and weak when it is broad. The dial is τ, and the truth θ is the unknown that decides whether the dial helps. For a fixed θ, $\bar{x}$ varies from experiment to experiment and μ_0 does not, so

$$\begin{aligned} \mathbb{E}_\theta\,\hat{\theta}_B - \theta &\overset{(1)}{=} w\theta + (1-w)\mu_0 - \theta = (1-w)(\mu_0 - \theta), \\ \mathrm{Var}_\theta\,\hat{\theta}_B &\overset{(2)}{=} w^2 s^2, \\ R(\theta) \equiv \mathbb{E}_\theta(\hat{\theta}_B - \theta)^2 &\overset{(3)}{=} w^2 s^2 + (1-w)^2(\theta - \mu_0)^2. \end{aligned} \tag{T5.13}$$

(1) $\mathbb{E}_\theta\,\bar{x} = \theta$, and the expectation is linear.
(2) Only $\bar{x}$ varies, with variance s^2, and it is multiplied by w.
(3) Mean squared error is variance plus squared bias (Eq. (3.13)).

The plain average $\bar{x}$ has $R = s^2$ for every θ. The Bayes estimator does better exactly when

$$\begin{aligned} w^2 s^2 + (1-w)^2(\theta-\mu_0)^2 < s^2 &\iff (1-w)^2(\theta-\mu_0)^2 < (1-w^2)s^2 \\ &\iff (\theta-\mu_0)^2 < \frac{1+w}{1-w}\,s^2 \iff (\theta-\mu_0)^2 < 2\tau^2 + s^2, \end{aligned} \tag{T5.14}$$

since $1 - w^2 = (1-w)(1+w)$ and, from (T5.12), $(1+w)/(1-w) = (2\tau^2 + s^2)/s^2$. So the pull helps whenever the truth lies within $\sqrt{2\tau^2 + s^2}$ of the prior's centre. For $\tau = s$, $w = \frac{1}{2}$: at $\theta = \mu_0$ the error is $s^2/4$, four times smaller than the plain average's, and the advantage lasts out to $|\theta - \mu_0| = \sqrt{3}\,s = 1.732\,s$. Beyond that the squared bias grows without limit. Lesson: a prior buys accuracy where it is right and costs accuracy where it is wrong, and (T5.14) says exactly where the boundary is.

A picture: the risk curve. Plotting the error $R(\theta)$ against the truth θ gives the **risk curve** of an estimator, its report card for every possible state of nature. The left panel of Figure T5.6 shows the flat line of $\bar{x}$ and the parabolas (T5.13) for three prior widths. A narrow prior digs a deep valley at μ_0 and climbs steeply away from it; a broad prior makes a shallow valley and stays close to the flat line. No curve lies below the flat line everywhere. Choosing an estimator means choosing which part of the θ axis to be good at.

One idea, many names. Pulling an estimate towards a preferred value to lower its scatter, at the price of a bias, is called *regularisation* in inverse problems (Section 6.14.3 used it to unfold a detector response), *ridge regression* in statistics and machine learning, the *Wiener filter* in signal processing and cosmology, and *shrinkage* when many estimates are pulled together (Section T5.4). Asking how a Bayes estimator behaves over repeated data is the subject of **statistical decision theory** [Wasserman Ch. 12], which gives the vocabulary below.

We first define the averages that make "better" precise, then work the binomial example where one estimator is best in the worst case, and then derive the Wiener filter, the Bayes estimator of a whole signal.

T5.3.1 Two averages and the Bayes risk

Working definition T5.2 (loss, risk, Bayes risk, minimax, admissible)

A **loss** $L(\theta, \hat{\theta})$ is the cost of reporting $\hat{\theta}$ when the truth is θ, e.g. $(\hat{\theta} - \theta)^2$. The **risk** is its average over repeated data at fixed truth, $R(\theta, \hat{\theta}) = \mathbb{E}_\theta\, L(\theta, \hat{\theta}(X))$; for squared loss it is the mean squared error [Wasserman Def. 12.1, p. 194]. The **Bayes risk** averages the risk over a prior, $r(\pi, \hat{\theta}) = \int R(\theta, \hat{\theta})\pi(\theta)\,\mathrm{d}\theta$. An estimator is **minimax** if its largest risk, $\sup_\theta R(\theta, \hat{\theta})$, is the smallest possible, and **admissible** if no other estimator has a risk at least as low for every θ and strictly lower for some [nobody beats it everywhere] [Wasserman Def. 12.4, 12.17].

The Bayes risk averages twice: over the data at fixed θ, then over θ. Doing the two integrals in the other order explains why the posterior mean is the estimator with the smallest Bayes risk [Wasserman Thm 12.7, p. 197]:

$$r(\pi, \hat{\theta}) \overset{(1)}{=} \int\!\!\int L(\theta, \hat{\theta}(x))\, p(x \mid \theta)\, \pi(\theta)\,\mathrm{d}x\,\mathrm{d}\theta \overset{(2)}{=} \int \Big[\int L(\theta, \hat{\theta}(x))\, p(\theta \mid x)\,\mathrm{d}\theta\Big] p(x)\,\mathrm{d}x. \quad \text{(T5.15)}$$

(1) The risk is the integral of the loss against the sampling density $p(x \mid \theta)$; then average over $\pi(\theta)$.
(2) Bayes' theorem $p(x \mid \theta)\pi(\theta) = p(\theta \mid x)p(x)$ (Eq. (3.81)), and swap the order of integration.

The bracket is the posterior expected loss at the observed x, and $p(x) \geq 0$. Choosing $\hat{\theta}(x)$ to make the bracket as small as possible for every x makes the whole integral as small as possible; for squared loss that choice is the posterior mean (Eq. (3.114)). So the two viewpoints agree on average: averaged over the prior, nothing beats the posterior mean. They disagree point by point, as the risk curves show.

Example: the Bayes risk of the Gaussian pull. Average (T5.13) over the prior, under which $\mathbb{E}(\theta - \mu_0)^2 = \tau^2$:

$$r \overset{(1)}{=} w^2 s^2 + (1-w)^2\tau^2 \overset{(2)}{=} \frac{\tau^4 s^2 + s^4\tau^2}{(\tau^2 + s^2)^2} \overset{(3)}{=} \frac{s^2\tau^2}{\tau^2 + s^2} = w\, s^2. \quad \text{(T5.16)}$$

(1) Linearity: only the bias term depends on θ.
(2) $w = \tau^2/(\tau^2 + s^2)$ and $1 - w = s^2/(\tau^2 + s^2)$.
(3) Factor $s^2\tau^2(\tau^2 + s^2)$ out of the numerator.

Averaged over the prior, the Bayes estimator's error is $w\, s^2$, smaller than the s^2 of $\bar{x}$ by the factor $w < 1$. The gain is large when the prior is narrow compared with the data's error, and disappears as $\tau \to \infty$. It is also exactly the posterior variance of Eq. (5.32), $1/(1/s^2 + 1/\tau^2)$: when the prior is the true distribution of θ, the average error of the posterior mean is the posterior's own width.

T5.3.2 Example: the best worst case for a success probability

A detector registers y hits in n trials, each a hit with probability p. The maximum-likelihood estimate is y/n, unbiased with risk $R = p(1-p)/n$ (Section 3.1.3). A Beta(α, β) prior gives the

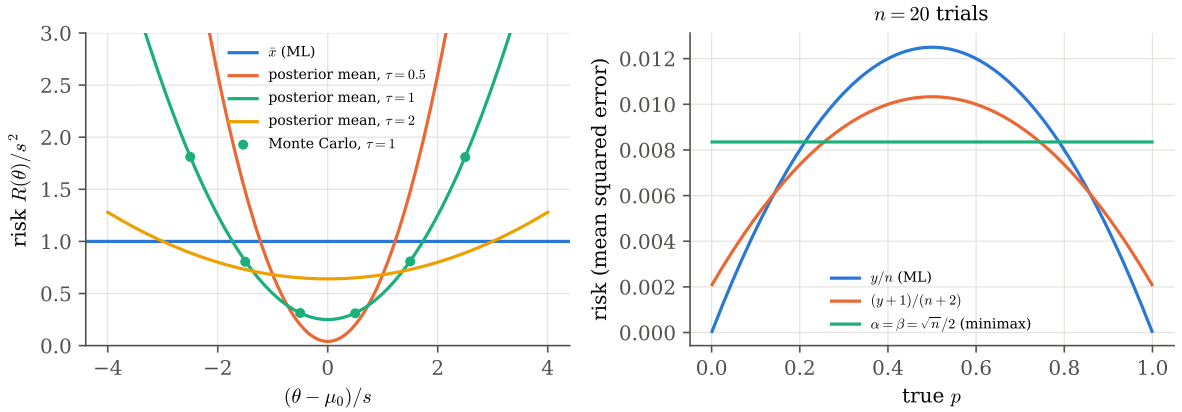


Figure T5.6: Risk curves. Left: an average $\bar{x} \sim \mathcal{N}(\theta, s^2)$ (flat blue line) against the posterior mean with a Gaussian prior centred at μ_0, for three prior widths τ; dots are Monte Carlo estimates (40000 experiments each) for $\tau = s$. Each curve dips below the flat line only within $\sqrt{2\tau^2 + s^2}$ of the prior's centre. Right: three estimators of a probability p from $n = 20$ trials. The maximum-likelihood y/n is best near $p = 0$ and $p = 1$, the minimax estimator (flat green line) is best near $p = \frac{1}{2}$, and the uniform-prior posterior mean lies in between.

posterior mean $\hat{p} = (y+\alpha)/(n+\alpha+\beta)$ (Eq. (5.30)). Its risk, using $\mathbb{E}\, y = np$ and $\operatorname{Var} y = np(1-p)$, is [Wasserman Ex. 12.3, pp. 194–195]

$$R(p) \overset{(1)}{=} \frac{\operatorname{Var} y}{(n+\alpha+\beta)^2} + \left(\frac{np+\alpha}{n+\alpha+\beta} - p\right)^2 \overset{(2)}{=} \frac{np(1-p) + \big(\alpha - (\alpha+\beta)p\big)^2}{(n+\alpha+\beta)^2}. \tag{T5.17}$$

(1) Variance plus squared bias; the shift by α does not change the variance.
(2) Put the bias over the common denominator: $np + \alpha - (n+\alpha+\beta)p = \alpha - (\alpha+\beta)p$.

Can the prior be chosen so that the risk does not depend on p at all? Take $\alpha = \beta = a$ and expand the numerator: $np - np^2 + a^2(1-2p)^2 = np - np^2 + a^2 - 4a^2p + 4a^2p^2$. The terms in p and p^2 cancel when $4a^2 = n$, that is $a = \sqrt{n}/2$, and then the numerator is $n/4$ for every p:

$$R(p) = \frac{n/4}{(n+\sqrt{n})^2} = \frac{1}{4(1+\sqrt{n})^2} \qquad \text{for all } p. \tag{T5.18}$$

For $n = 20$ this constant is 0.00835, while the risk of y/n reaches $1/(4n) = 0.01250$ at $p = \frac{1}{2}$. A Bayes estimator with constant risk is minimax [Wasserman Thm 12.11, p. 199], and the reason is one line: its Bayes risk is the constant itself, so another estimator with a lower maximum risk would also have a lower Bayes risk, which the Bayes estimator does not allow. The right panel of Figure T5.6 shows the price: near $p = 0$ or $p = 1$, where hits are rare or near-certain, the minimax estimator is much worse than y/n, because the prior Beta(2.236, 2.236) drags it towards $\frac{1}{2}$. A physicist measuring a trigger efficiency of 99% would not want it. The minimax criterion protects the worst case, and here the worst case is not where we expect to be.

For the Gaussian mean the same logic gives the opposite verdict about the plain average: $\bar{x}$ has constant risk s^2, and it is the limit of the posterior means (T5.12) as $\tau \to \infty$; it is minimax and, for a single mean, admissible [Wasserman Thm 12.14, 12.20, pp. 200–203]. No estimator of one Gaussian mean beats $\bar{x}$ everywhere. Section T5.4 shows that this stops being true as soon as three or more means are estimated together.

T5.3.3 The linear–Gaussian case: the Wiener filter and ridge regression

Now let the unknown be a whole signal: a vector $\boldsymbol{s}$ of pixel values, Fourier modes or regression coefficients. The data are linear in it, $\boldsymbol{d} = \mathbf{A}\boldsymbol{s} + \boldsymbol{n}$, as in the linear model of Section T5.2, with Gaussian noise of covariance $\mathbf{N}$. The prior says the signal is itself a Gaussian random vector of mean zero and covariance $\mathbf{S}$, as a Gaussian random field is (Chapter 7). The question is the one of (T5.12), for many numbers at once: what is the posterior mean of $\boldsymbol{s}$, and how does it behave over repeated data?

The posterior. Up to factors that do not depend on $\boldsymbol{s}$, the posterior is likelihood times prior,

$$\begin{aligned} -2\ln p(\boldsymbol{s}\mid\boldsymbol{d}) &\overset{(1)}{=} (\boldsymbol{d}-\mathbf{A}\boldsymbol{s})^{\mathsf{T}}\mathbf{N}^{-1}(\boldsymbol{d}-\mathbf{A}\boldsymbol{s}) + \boldsymbol{s}^{\mathsf{T}}\mathbf{S}^{-1}\boldsymbol{s} + \text{const} \\ &\overset{(2)}{=} \boldsymbol{s}^{\mathsf{T}}(\mathbf{S}^{-1}+\mathbf{A}^{\mathsf{T}}\mathbf{N}^{-1}\mathbf{A})\boldsymbol{s} - 2\boldsymbol{s}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}}\mathbf{N}^{-1}\boldsymbol{d} + \text{const} \\ &\overset{(3)}{=} (\boldsymbol{s}-\hat{\boldsymbol{s}})^{\mathsf{T}}\mathbf{P}(\boldsymbol{s}-\hat{\boldsymbol{s}}) + \text{const}, \qquad \mathbf{P} = \mathbf{S}^{-1}+\mathbf{A}^{\mathsf{T}}\mathbf{N}^{-1}\mathbf{A}, \quad \mathbf{P}\hat{\boldsymbol{s}} = \mathbf{A}^{\mathsf{T}}\mathbf{N}^{-1}\boldsymbol{d}. \end{aligned} \tag{T5.19}$$

(1) Gaussian likelihood (Eq. (3.92)) and Gaussian prior; normalisations are constants.
(2) Expand the first quadratic form; $\boldsymbol{d}^{\mathsf{T}}\mathbf{N}^{-1}\boldsymbol{d}$ is a constant, and the two cross terms are equal numbers.
(3) Complete the square: expanding the last line gives $\boldsymbol{s}^{\mathsf{T}}\mathbf{P}\boldsymbol{s} - 2\boldsymbol{s}^{\mathsf{T}}\mathbf{P}\hat{\boldsymbol{s}} + \text{const}$, which matches line two when $\mathbf{P}\hat{\boldsymbol{s}} = \mathbf{A}^{\mathsf{T}}\mathbf{N}^{-1}\boldsymbol{d}$.

So the posterior is Gaussian with mean $\hat{\boldsymbol{s}}$ and covariance $\mathbf{P}^{-1}$, and the posterior mean, also its mode and median, is

$$\hat{\boldsymbol{s}} = (\mathbf{S}^{-1}+\mathbf{A}^{\mathsf{T}}\mathbf{N}^{-1}\mathbf{A})^{-1}\mathbf{A}^{\mathsf{T}}\mathbf{N}^{-1}\boldsymbol{d} = \mathbf{S}\mathbf{A}^{\mathsf{T}}(\mathbf{A}\mathbf{S}\mathbf{A}^{\mathsf{T}}+\mathbf{N})^{-1}\boldsymbol{d}. \tag{T5.20}$$

The second form follows from a "push-through" identity, checked by multiplying out: $\mathbf{A}^{\mathsf{T}}\mathbf{N}^{-1}(\mathbf{A}\mathbf{S}\mathbf{A}^{\mathsf{T}}+\mathbf{N}) = \mathbf{A}^{\mathsf{T}}\mathbf{N}^{-1}\mathbf{A}\mathbf{S}\mathbf{A}^{\mathsf{T}} + \mathbf{A}^{\mathsf{T}} = (\mathbf{A}^{\mathsf{T}}\mathbf{N}^{-1}\mathbf{A}+\mathbf{S}^{-1})\mathbf{S}\mathbf{A}^{\mathsf{T}}$; multiply on the left by $\mathbf{P}^{-1}$ and on the right by $(\mathbf{A}\mathbf{S}\mathbf{A}^{\mathsf{T}}+\mathbf{N})^{-1}$. The first form inverts a matrix the size of the signal, the second one the size of the data, and in practice one uses whichever is smaller. This is the **Wiener filter** [HobsonBMC (6.23), p. 141; (10.20), p. 236].

Compare it with the BLUE. Without the prior ($\mathbf{S}^{-1} \to \mathbf{0}$, an infinitely broad prior) the first form of (T5.20) becomes $(\mathbf{A}^{\mathsf{T}}\mathbf{N}^{-1}\mathbf{A})^{-1}\mathbf{A}^{\mathsf{T}}\mathbf{N}^{-1}\boldsymbol{d}$, the Gauss–Markov estimator. The prior adds $\mathbf{S}^{-1}$ to the information matrix, just as a Gaussian prior adds its precision to a Fisher matrix (Eq. (10.26)). The price is a bias: for fixed $\boldsymbol{s}$, $\mathbb{E}\,\hat{\boldsymbol{s}} = \mathbf{P}^{-1}\mathbf{A}^{\mathsf{T}}\mathbf{N}^{-1}\mathbf{A}\,\boldsymbol{s} \neq \boldsymbol{s}$.

Example: denoising a stationary signal, one Fourier mode at a time. Take $\mathbf{A} = \mathbb{1}$, a signal and a noise that are both stationary on a periodic grid. Both covariances are then diagonal in Fourier space, with the power spectra $P_s(k)$ and $P_n(k)$ on the diagonal (the Wiener–Khinchin theorem, Eq. (7.12)), so (T5.20) acts on each Fourier mode separately, as the scalar case (T5.12) with $\mu_0 = 0$:

$$\hat{s}_k = w_k\, d_k, \qquad w_k = \frac{P_s(k)}{P_s(k)+P_n(k)}. \tag{T5.21}$$

Modes where the signal dominates pass almost untouched ($w_k \to 1$); modes where the noise dominates are suppressed ($w_k \to 0$). The error and the bias per mode follow from (T5.13) with $s^2 \to P_n$ and $\tau^2 \to P_s$:

$$\begin{aligned} \mathbb{E}\,|\hat{s}_k - s_k|^2 &\overset{(1)}{=} (1-w_k)^2 P_s + w_k^2 P_n \overset{(2)}{=} \frac{P_n^2 P_s + P_s^2 P_n}{(P_s+P_n)^2} = \frac{P_s P_n}{P_s+P_n}, \\ \mathbb{E}\,|\hat{s}_k|^2 &\overset{(3)}{=} w_k^2 (P_s+P_n) = \frac{P_s^2}{P_s+P_n} < P_s. \end{aligned} \tag{T5.22}$$

(1) $\hat{s}_k - s_k = (w_k - 1)s_k + w_k n_k$ with independent s_k and n_k; average over both, as in the Bayes risk (T5.16).
(2) $1 - w_k = P_n/(P_s + P_n)$; then factor $P_s P_n (P_s + P_n)$ out of the numerator.
(3) The data mode $d_k = s_k + n_k$ has power $P_s + P_n$, multiplied by w_k^2.

The error per mode is smaller than both P_n (the raw data) and P_s (the prior guess $\hat{s} = 0$), so the Wiener filter beats both. But the filtered signal has *less* power than the true one: every feature is fainter, the "oversmoothing" of [HobsonBMC p. 237].

The script `code/chT5/03_bayes_risk.py` draws 4000 periodic signals of 512 pixels with a smooth spectrum that equals the white noise level at $k \approx 25.1$, adds the noise, and filters with (T5.21). The mean squared error per pixel drops from 0.9987 in the raw data to 0.1063 after filtering (prediction 0.1061). The variance of the filtered map is 0.931 (prediction 0.935), against 1.038 for the true signal: the filtered map is ten times closer to the truth and visibly fainter (Figure T5.7). The script also checks that the two matrix forms in (T5.20) and the Fourier form agree on a small grid.

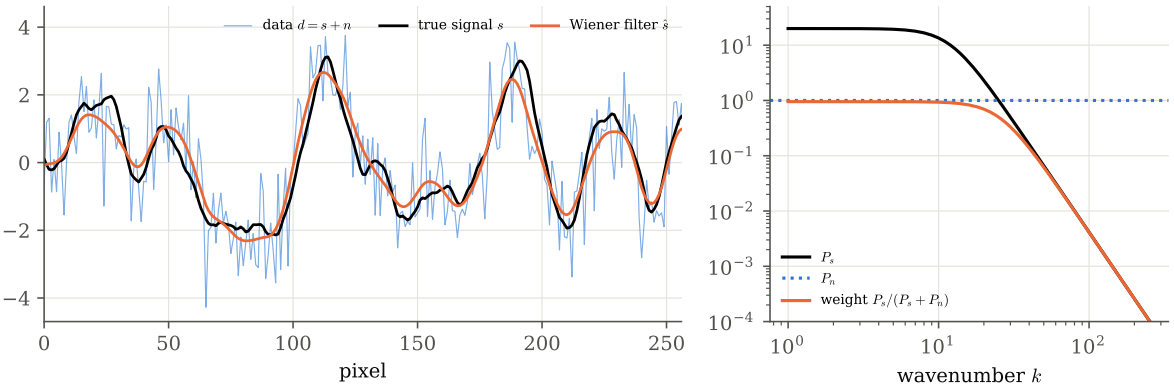


Figure T5.7: Left: one realisation of a smooth periodic signal (black), the signal plus white noise (light blue), and the Wiener-filtered estimate (orange). The filter removes most of the noise but also flattens the peaks. Right: the signal and noise power spectra and the filter weight $P_s/(P_s + P_n)$, which is close to one where the signal dominates and falls like P_s/P_n where the noise does.

Example: ridge regression is a Wiener filter. Fit m coefficients $\boldsymbol{\theta}$ to N readings with independent noise of variance σ^2, $\boldsymbol{d} = \mathbf{A}\boldsymbol{\theta} + \boldsymbol{n}$, and put an independent Gaussian prior of width τ on each coefficient: $\mathbf{N} = \sigma^2 \mathbb{1}$, $\mathbf{S} = \tau^2 \mathbb{1}$. The first form of (T5.20) becomes

$$\hat{\boldsymbol{\theta}} \overset{(1)}{=} \left(\frac{1}{\tau^2} + \frac{\mathbf{A}^{\mathbb{T}}\mathbf{A}}{\sigma^2} \right)^{-1} \frac{\mathbf{A}^{\mathsf{T}}\boldsymbol{d}}{\sigma^2} \overset{(2)}{=} (\mathbf{A}^{\mathsf{T}}\mathbf{A} + \lambda \mathbb{1})^{-1} \mathbf{A}^{\mathsf{T}}\boldsymbol{d}, \qquad \lambda = \frac{\sigma^2}{\tau^2}, \tag{T5.23}$$

(1) Insert $\mathbf{S}^{-1} = \mathbb{1}/\tau^2$ and $\mathbf{N}^{-1} = \mathbb{1}/\sigma^2$.
(2) Multiply the bracket by σ^2 and its inverse by $1/\sigma^2$; the factors cancel against the $1/\sigma^2$ on the right.

the **ridge regression** estimator. Without the $\lambda \mathbb{1}$ it is ordinary least squares; adding λ to the diagonal of $\mathbf{A}^{\mathsf{T}}\mathbf{A}$ keeps the inverse well behaved when the columns of $\mathbf{A}$ are nearly dependent (a polynomial of high degree, an unfolding matrix), which is exactly where least squares has enormous variance. The Tikhonov unfolding of Eq. (6.72) is the same estimator with a curvature penalty in place of $\lambda \mathbb{1}$, and its MSE-optimal filter factor Eq. (6.75) is (T5.21) once more. In the map-to-map comparison of Eq. (12.70) the weight $C/(C + N)$ that maximises the correlation with the true map is the same filter again.

Watch out: A Wiener-filtered map has the wrong power spectrum

By (T5.22), the power of a Wiener-filtered map is $P_s^2/(P_s + P_n)$, not P_s. Measuring a power spectrum, or counting hot spots, or computing any statistic of the field on the filtered map gives a biased answer unless the filter is undone in the statistic. The filtered map is the best map in the mean-squared sense; it is not a typical realisation of the sky.

When to use a Bayes estimator. Three situations favour it. When the prior is genuinely known, as for a population of objects whose distribution we have measured before, or a Gaussian random field whose spectrum is known, the Bayes estimator has the smallest average error, and nothing else should be used. When the data alone are nearly degenerate, as in an unfolding or a fit with many coefficients, a mild prior turns an unusable estimate into a usable one, at a bias we can compute. When the goal is a map or a reconstruction to look at, the Wiener filter is the natural choice. In the opposite situation, a single parameter where an unexpected value is the discovery we hope for, a prior that pulls towards the expected value is the wrong tool: an unbiased estimate with an honest error bar, and a likelihood that others can combine with their own priors, serves better.

T5.4 Many parameters at once: shrinkage and Stein's paradox

What we already have. For one Gaussian mean, Section T5.3.2 recalled that the plain average cannot be beaten everywhere: any estimator better than it for some true values is worse for others. Section 3.2.3 showed the simplest such trade, $c\bar{x}$ with $c < 1$, which wins near zero and loses far from it, and Section T5.3 showed that a prior turns this pull into the posterior mean. In every case the pull had to be paid for with a known prior or a lucky truth.

When that is not enough. Often we estimate not one number but many of the same kind: the dark currents of the four million pixels of a CCD (the example at the end of Section T5.1.1), the calibration gains of a hundred detectors, the band powers of a spectrum, the luminosities of a sample of galaxies. Each has its own data, and estimating each by its own average looks like the obvious and even the only fair thing to do. Surprisingly, when there are three or more of them, it is not the best thing to do, even with no prior at all.

A story with the whole trick in it. Early in a baseball season, a few players have batted spectacularly well and a few spectacularly badly, over only a few dozen attempts each. Who will have the best average at the end of the season? Experience says: the stars of April will still be good, but not as good as in April, and the April failures will recover somewhat. Part of an extreme early record is luck, and luck does not repeat. So the best forecast for every player pulls his early average part of the way towards the average of all players. The pull is not a prior opinion about any player: its size is learned from the spread of all the early records together. Efron and Morris used exactly this example to make the idea famous [126, 127].

The physics is the same. Among a hundred detector gains measured with equal noise, the ones that came out highest are, on average, the ones whose noise happened to be positive. Pulling all the estimates towards their common average undoes part of that selection and lowers the total error. The idea of the section is that estimating many similar numbers together, and letting them borrow strength from each other, beats estimating each on its own.

The story in symbols. Take k readings $X_i \sim \mathcal{N}(\theta_i, 1)$, $i = 1, \dots, k$, independent, one per unknown mean, and judge an estimator by its total squared error $\sum_i (\hat{\theta}_i - \theta_i)^2$. Write

$\boldsymbol{X} = (X_1, \dots, X_k)$ and $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$. Maximum likelihood takes $\hat{\boldsymbol{\theta}} = \boldsymbol{X}$, with total risk $\sum_i \operatorname{Var} X_i = k$ for every $\boldsymbol{\theta}$. The **James–Stein estimator** [128] pulls the whole vector towards the origin by a factor that depends on its length:

$$\hat{\boldsymbol{\theta}}_{\mathrm{JS}} = \left(1 - \frac{c}{|\boldsymbol{X}|^2}\right)\boldsymbol{X}, \qquad c = k - 2, \qquad |\boldsymbol{X}|^2 = \sum_i X_i^2. \tag{T5.24}$$

The dial is c; $c = 0$ is maximum likelihood. The next subsection derives the total risk for any c,

$$R(\boldsymbol{\theta}) = k - \left[2c(k-2) - c^2\right] \mathbb{E}_{\boldsymbol{\theta}} \frac{1}{|\boldsymbol{X}|^2}, \tag{T5.25}$$

which is below k for every $\boldsymbol{\theta}$ whenever $0 < c < 2(k-2)$, and lowest for $c = k - 2$. The size of the gain: with $k = 10$ means that are all zero, the James–Stein total risk is 2 instead of 10. The price: the estimates are biased towards the origin, and one mean that differs a lot from the others can be estimated worse than by its own reading. Lesson: for three or more means, the plain readings are *inadmissible*, beaten everywhere by a rule that shrinks them together [129], [Wasserman §12.7, p. 204].

A picture: the readings are too far from the origin. Why should pulling towards an arbitrary point help? Look at lengths. The reading is $\boldsymbol{X} = \boldsymbol{\theta} + \boldsymbol{n}$, with noise $\boldsymbol{n}$ of mean zero and $\mathbb{E}\,|\boldsymbol{n}|^2 = k$. On average $\mathbb{E}\,|\boldsymbol{X}|^2 = |\boldsymbol{\theta}|^2 + 2\boldsymbol{\theta} \cdot \mathbb{E}\,\boldsymbol{n} + \mathbb{E}\,|\boldsymbol{n}|^2 = |\boldsymbol{\theta}|^2 + k$: the reading is systematically *longer* than the truth. In many dimensions the noise vector is almost always nearly perpendicular to $\boldsymbol{\theta}$ (there are $k - 1$ directions perpendicular to $\boldsymbol{\theta}$ and only one along it), so $\boldsymbol{X}$ sits on the far side of a right angle, as in Figure T5.8, and the point of the line from the origin to $\boldsymbol{X}$ that is closest to $\boldsymbol{\theta}$ lies *inside* $\boldsymbol{X}$. Shrinking towards the origin moves along that line towards the truth. The amount of shrinkage needed is about $k/|\boldsymbol{X}|^2$, the excess length squared relative to the total, and (T5.24) estimates it from $|\boldsymbol{X}|^2$ alone. In one or two dimensions the noise is as often along $\boldsymbol{\theta}$ as across it, the picture fails, and so does the gain.

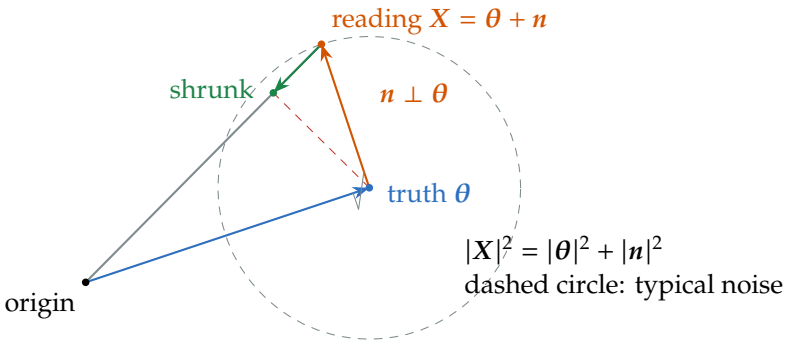


Figure T5.8: Why shrinkage helps in many dimensions. The noise $\boldsymbol{n}$ is nearly perpendicular to the truth $\boldsymbol{\theta}$, so the reading $\boldsymbol{X}$ is longer than $\boldsymbol{\theta}$ (Pythagoras). Moving from $\boldsymbol{X}$ along the line towards the origin (green) reaches a point closer to $\boldsymbol{\theta}$ than $\boldsymbol{X}$ is: here the distance drops from 1.6 to 1.43. The red dashed segment is the shortest distance from $\boldsymbol{\theta}$ to that line.

One idea, many names. Pulling many estimates towards a common value, with the strength of the pull learned from the estimates themselves, is called *shrinkage*, *partial pooling* or *borrowing strength*. In Bayesian language it is a *hierarchical* or *multilevel* model: each θ_i is drawn from a population whose spread is itself estimated [HobsonBMC (11.10), p. 254], [Kruschke §9.3]. When the population spread is estimated by a simple formula and plugged in, the method is **empirical Bayes** [130]. The everyday name for the effect it corrects is *regression to the mean*.

We now prove (T5.25) with one identity, read the estimator as empirical Bayes, check it by simulation, and then say when it should not be used.

T5.4.1 Stein's lemma and the James–Stein risk

The whole proof rests on one identity for a Gaussian variable. Let $Z \sim \mathcal{N}(\theta, 1)$ with density $\varphi(z-\theta)$, $\varphi(u) = e^{-u^2/2}/\sqrt{2\pi}$, and let g be a differentiable function that grows more slowly than $e^{u^2/2}$. Then

$$\mathbb{E}\big[(Z-\theta)\,g(Z)\big] \overset{(1)}{=} \int_{-\infty}^{\infty} g(z)\,(z-\theta)\varphi(z-\theta)\,\mathrm{d}z \overset{(2)}{=} -\int_{-\infty}^{\infty} g(z)\,\frac{\mathrm{d}}{\mathrm{d}z}\varphi(z-\theta)\,\mathrm{d}z$$
$$\overset{(3)}{=} -\big[g(z)\varphi(z-\theta)\big]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} g'(z)\varphi(z-\theta)\,\mathrm{d}z \overset{(4)}{=} \mathbb{E}\; g'(Z). \tag{T5.26}$$

(1) Definition of the expectation.
(2) The Gaussian's derivative brings down the factor: $\frac{\mathrm{d}}{\mathrm{d}z}e^{-(z-\theta)^2/2} = -(z-\theta)e^{-(z-\theta)^2/2}$.
(3) Integration by parts.
(4) The boundary term vanishes because φ falls faster than g grows; what remains is the expectation of g'.

This is **Stein's lemma**: for a Gaussian, multiplying by the deviation and averaging is the same as differentiating and averaging. With k independent components, apply it to component i with the others held fixed, then average over the others: $\mathbb{E}[(X_i-\theta_i)\,g_i(\boldsymbol{X})] = \mathbb{E}[\partial_i g_i(\boldsymbol{X})]$, with $\partial_i = \partial/\partial X_i$.

The risk of a shrinkage rule. Write the estimator as $\hat{\boldsymbol{\theta}} = \boldsymbol{X} - c\,\boldsymbol{g}(\boldsymbol{X})$ with $\boldsymbol{g}(\boldsymbol{X}) = \boldsymbol{X}/|\boldsymbol{X}|^2$. Its squared error is

$$|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}|^2 \overset{(1)}{=} |\boldsymbol{X}-\boldsymbol{\theta}|^2 - 2c\,(\boldsymbol{X}-\boldsymbol{\theta})\cdot\boldsymbol{g}(\boldsymbol{X}) + c^2|\boldsymbol{g}(\boldsymbol{X})|^2$$
$$\overset{(2)}{=} |\boldsymbol{X}-\boldsymbol{\theta}|^2 - 2c\sum_{i=1}^{k}(X_i-\theta_i)\frac{X_i}{|\boldsymbol{X}|^2} + \frac{c^2}{|\boldsymbol{X}|^2}. \tag{T5.27}$$

(1) Expand $|\boldsymbol{a}-c\boldsymbol{b}|^2 = |\boldsymbol{a}|^2 - 2c\,\boldsymbol{a}\cdot\boldsymbol{b} + c^2|\boldsymbol{b}|^2$ with $\boldsymbol{a} = \boldsymbol{X}-\boldsymbol{\theta}$.
(2) Write the dot product as a sum; $|\boldsymbol{X}/|\boldsymbol{X}|^2|^2 = 1/|\boldsymbol{X}|^2$.

Average each term. The first gives k. For the middle term, Stein's lemma with $g_i = X_i/|\boldsymbol{X}|^2$ needs the derivative

$$\partial_i \frac{X_i}{|\boldsymbol{X}|^2} \overset{(1)}{=} \frac{1}{|\boldsymbol{X}|^2} - \frac{X_i\cdot 2X_i}{|\boldsymbol{X}|^4}, \qquad \sum_{i=1}^{k}\partial_i\frac{X_i}{|\boldsymbol{X}|^2} \overset{(2)}{=} \frac{k}{|\boldsymbol{X}|^2} - \frac{2|\boldsymbol{X}|^2}{|\boldsymbol{X}|^4} = \frac{k-2}{|\boldsymbol{X}|^2}, \tag{T5.28}$$

(1) Quotient rule, with $\partial_i|\boldsymbol{X}|^2 = 2X_i$.
(2) Sum over i: $\sum_i X_i^2 = |\boldsymbol{X}|^2$.

Putting the three averages together gives (T5.25): $R = k - 2c(k-2)\,\mathbb{E}\,|\boldsymbol{X}|^{-2} + c^2\,\mathbb{E}\,|\boldsymbol{X}|^{-2}$. As a function of c the bracket $2c(k-2) - c^2$ is a downward parabola with roots 0 and $2(k-2)$ and its peak at $c = k-2$, where it equals $(k-2)^2$. The best shrinkage therefore has total risk $k - (k-2)^2\,\mathbb{E}\,|\boldsymbol{X}|^{-2}$, below k for every $\boldsymbol{\theta}$ as soon as $k \geq 3$.

Key idea: Stein's paradox

For $k \geq 3$ independent readings $X_i \sim \mathcal{N}(\theta_i, 1)$, the James–Stein estimator (T5.24) has total risk $R(\boldsymbol{\theta}) = k - (k-2)^2\,\mathbb{E}_{\boldsymbol{\theta}}\,|\boldsymbol{X}|^{-2} < k$ for every $\boldsymbol{\theta}$. Estimating each mean by its own reading is beaten everywhere, although no reading carries any information about the other means.

Why $k \geq 3$? The factor $k-2$ must be positive, and $\mathbb{E}\,|\boldsymbol{X}|^{-2}$ must be finite. Near the origin $|\boldsymbol{X}|^2 = r^2$ and the volume element is $r^{k-1}\mathrm{d}r$, so $\mathbb{E}\,|\boldsymbol{X}|^{-2}$ contains $\int_0 r^{k-3}\mathrm{d}r$, which diverges for

$k \le 2$. For $k = 1$ and $k = 2$ no shrinkage of this form helps, in agreement with the admissibility of the single average.

Example: all true means at zero. If $\boldsymbol{\theta} = \mathbf{0}$, $|\boldsymbol{X}|^2$ is a sum of k squared standard Gaussians, a χ^2_k variable. Its inverse has mean

$$\mathbb{E}\,\frac{1}{\chi^2_k} \overset{(1)}{=} \int_0^\infty \frac{1}{v}\,\frac{v^{k/2-1}e^{-v/2}}{2^{k/2}\Gamma(k/2)}\,\mathrm{d}v \overset{(2)}{=} \frac{2^{k/2-1}\Gamma(k/2-1)}{2^{k/2}\Gamma(k/2)} \overset{(3)}{=} \frac{1}{k-2}, \tag{T5.29}$$

(1) The χ^2_k density (Chapter 2).

(2) The gamma integral $\int_0^\infty v^{a-1}e^{-v/2}\mathrm{d}v = \Gamma(a)2^a$ with $a = k/2-1$, finite for $k > 2$.

(3) $\Gamma(k/2) = (k/2-1)\Gamma(k/2-1)$, so the ratio is $\frac{1}{2(k/2-1)}$; these are the steps of Eq. (2.103).

so $R(\mathbf{0}) = k - (k-2)^2/(k-2) = 2$, whatever k is. For ten means the total error falls from 10 to 2; the simulation below finds 2.040.

Example: how the gain fades as the means spread out. The risk depends on $\boldsymbol{\theta}$ only through its length $|\boldsymbol{\theta}|$: the noise is the same in every direction, so rotating $\boldsymbol{\theta}$ changes nothing. A bound follows from Jensen's inequality for the convex function $1/v$, $\mathbb{E}(1/V) \ge 1/\mathbb{E}\,V$, with $\mathbb{E}\,|\boldsymbol{X}|^2 = |\boldsymbol{\theta}|^2 + k$:

$$R(\boldsymbol{\theta}) \le k - \frac{(k-2)^2}{|\boldsymbol{\theta}|^2 + k}. \tag{T5.30}$$

For $|\boldsymbol{\theta}|^2 \ll k$ the gain is large; for $|\boldsymbol{\theta}|^2 \gg k$ the readings are already long compared with their noise, the shrinkage factor $1 - (k-2)/|\boldsymbol{X}|^2$ is close to one, and the estimator becomes the plain readings again. For $k = 10$ and $|\boldsymbol{\theta}| = 4$ the bound gives $10 - 64/26 = 7.5$; the simulation finds 7.17. (The exact curve in Figure T5.9 uses the fact that $|\boldsymbol{X}|^2$ then follows a noncentral χ^2 law, a Poisson-weighted mixture of central χ^2_{k+2j} laws, each with $\mathbb{E}(1/\chi^2_{k+2j}) = 1/(k-2+2j)$ by (T5.29).)

T5.4.2 The empirical-Bayes reading

Where does the formula come from, and why $k-2$? Suppose the true means are themselves scattered around zero, $\theta_i \sim \mathcal{N}(0, \tau^2)$, as the gains of many similar detectors are scattered around their design value. For each i this is the Gaussian prior of Section T5.3 with $s = 1$ and $\mu_0 = 0$, so the Bayes estimator is (T5.12):

$$\hat{\theta}_i = \frac{\tau^2}{1+\tau^2}\,X_i = \Big(1 - B\Big)X_i, \qquad B = \frac{1}{1+\tau^2}. \tag{T5.31}$$

We do not know τ, but the k readings together tell us about it. Each $X_i = \theta_i + n_i$ is a sum of two independent Gaussians, so over the population $X_i \sim \mathcal{N}(0, 1+\tau^2)$, and $|\boldsymbol{X}|^2/(1+\tau^2)$ is a χ^2_k variable. By (T5.29),

$$\mathbb{E}\,\frac{k-2}{|\boldsymbol{X}|^2} = \frac{k-2}{1+\tau^2}\,\mathbb{E}\,\frac{1}{\chi^2_k} = \frac{1}{1+\tau^2} = B. \tag{T5.32}$$

So $(k-2)/|\boldsymbol{X}|^2$ is an unbiased estimate of the shrinkage factor B, and plugging it into (T5.31) gives exactly the James–Stein estimator [130], [CasellaBerger Ex. 11.24, p. 459]. The data choose the amount of shrinkage: if the readings are spread much more widely than the noise, $|\boldsymbol{X}|^2$ is large, τ is evidently large, and little shrinkage is applied; if they are hardly more spread than the noise alone would make them, most of their spread is noise and they are pulled in hard. The maximum-likelihood readings are always *overdispersed*: their spread, $\sqrt{1+\tau^2}$, is larger than the spread τ of the true values [HobsonBMC p. 254].

Example: twenty detectors. The script `code/chT5/04_james_stein.py` draws 20 true gains from $\mathcal{N}(0,1)$, so $B = \frac{1}{2}$, adds unit noise, and repeats 20000 times. The estimated shrinkage factor $(k-2)/|\boldsymbol{X}|^2$ has mean 0.501 and standard deviation 0.181 over the experiments. The total squared error is 20.04 for the plain readings, 11.04 for James–Stein, 10.97 for its positive-part version (below), and 10.00 for the "oracle" that knows $\tau = 1$, (T5.16) summed over 20 means. Not knowing τ costs about one unit of total error out of the ten that shrinkage gains.

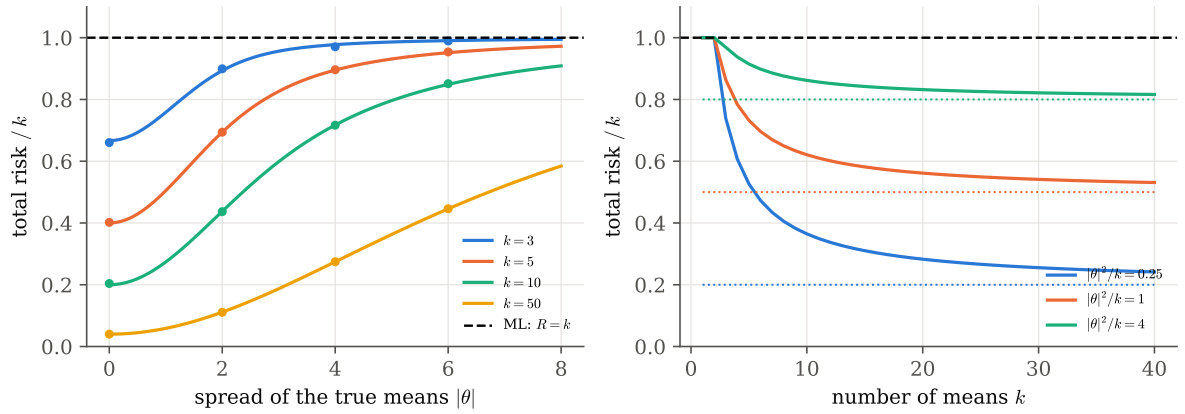


Figure T5.9: Total risk of the James–Stein estimator divided by k (the plain readings give one, dashed). Left: against the length $|\boldsymbol{\theta}|$ of the vector of true means, for $k = 3, 5, 10, 50$; curves are exact, dots are Monte Carlo. Every curve starts at $2/k$ and rises towards one as the means spread out. Right: against k, with the means spread so that $|\boldsymbol{\theta}|^2/k$ is fixed; dotted lines are the empirical-Bayes limit $1 - B = \tau^2/(1+\tau^2)$ with $\tau^2 = |\boldsymbol{\theta}|^2/k$, reached for many means.

Figure T5.9 shows the whole picture. The gain is largest when the means are close together compared with the noise and when there are many of them; it never turns into a loss. For a fixed spread per component, the risk per component falls quickly with k to the empirical-Bayes limit $\tau^2/(1+\tau^2)$, the value an analyst who knew τ would get.

T5.4.3 When shrinkage is dangerous

The guarantee is for the total, not for each mean. Take $k = 10$ means, nine of them at zero and one, θ_1, at a varying distance. The nine are well described by "all close together"; the tenth is not, but the shrinkage factor is computed from all ten and pulls it in too. Figure T5.10 shows the error of each. The nine benefit throughout. The tenth is estimated *worse* than by its own reading once θ_1 is more than about one unit away, with a mean squared error that peaks at 2.93, nearly three times the error of its own reading, at $\theta_1 \approx 3.75$. The total, 6.89 there, is still below 10, exactly as promised. If the tenth number is the one we care about, say an anomalous detector or the one band power where new physics might show up, we must not shrink it with the others.

The numbers must be exchangeable. The empirical-Bayes reading assumed that all θ_i come from one population. Shrinking a detector gain, a galaxy luminosity and a band power towards a common value is legal mathematics (the risk theorem holds for any $\boldsymbol{\theta}$) but useless: the gain is large only when the true values really are close together, on the same scale and with comparable noise. In practice one standardises each reading by its noise and shrinks only quantities of the same kind.

Shrink towards a sensible point. The origin in (T5.24) is arbitrary; any fixed point $\boldsymbol{\mu}_0$ works with $\boldsymbol{X}$ replaced by $\boldsymbol{X} - \boldsymbol{\mu}_0$. Better still is the average of the readings themselves: shrinking towards $\bar{X}$ uses one degree of freedom to estimate the centre, and the factor becomes $(k-3)/\sum_i (X_i - \bar{X})^2$, which needs $k \geq 4$ [126].

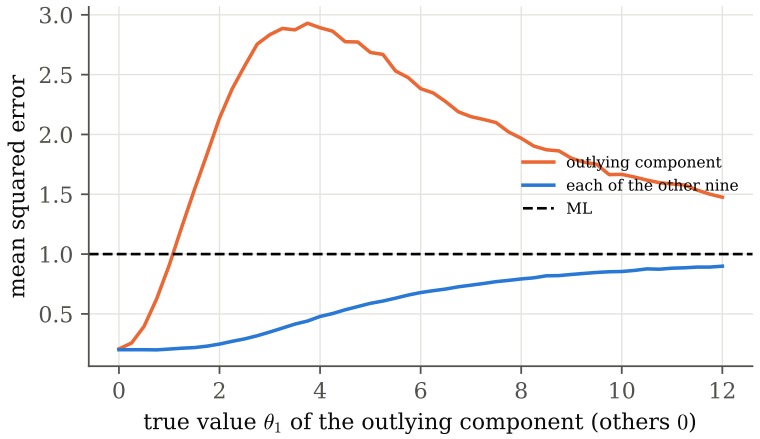


Figure T5.10: James–Stein with $k = 10$ means, nine at zero and one at θ_1. The nine (blue, each) always gain. The outlying one (orange) loses once it is more than about one unit from the others, by up to a factor of three, and recovers slowly as it moves so far away that the shrinkage factor approaches one.

Never shrink past the target. When $|\boldsymbol{X}|^2 < k - 2$ the factor $1 - (k-2)/|\boldsymbol{X}|^2$ is negative and (T5.24) flips every sign, which is absurd. The **positive-part** estimator replaces the factor by $\max(0, 1 - (k-2)/|\boldsymbol{X}|^2)$ [Wasserman (12.12), p. 204]; it has lower risk than James–Stein for every $\boldsymbol{\theta}$, and in the twenty-detector example it gained a little, 10.97 against 11.04.

Watch out: Shrinkage changes the error bars too

A shrunk estimate is biased, by $-B\,\theta_i$ on average for component i, and its error bar is not 1. Quoting $\hat{\theta}_i \pm 1$ after shrinking overstates the precision for the outliers and understates the gain for the rest. The empirical-Bayes posterior gives a consistent error bar, $\sqrt{1 - B}$ plus a correction for the uncertainty in B itself [126]; a hierarchical model sampled with MCMC (Chapter 9) gives it automatically.

When to use shrinkage. Use it when many quantities of the same kind are estimated with comparable noise, when the total or typical error is what matters (a map of calibrations, a catalogue of luminosities, the many entries of a covariance matrix estimated from simulations), and when no single quantity is singled out for discovery. Do not use it for the one number an experiment exists to measure.

T5.5 Wild readings: robust estimators

What we already have. The three laboratories of Section T5.1 showed that a few wild readings can make the sample mean five times worse than it needs to be, and that the median resists them (Section 3.13.4). We also saw the best possible answer: maximum likelihood with the exact law of the noise, which came within a few per cent of the Cramér–Rao bound in laboratory C. And we met the picture that organises everything here: each estimator is a balance of *pulls* ψ, one per reading (Figure T5.3).

When that is not enough. Maximum likelihood in laboratory C needed to know that 5% of the readings come from a Gaussian exactly ten times wider. In real data we never know that. We know only that the instrument is mostly well behaved and that something occasionally goes

wrong: a cosmic ray, a glitch in a gravitational-wave strain channel, a mistyped entry, a satellite crossing the field of view. We want a recipe that is almost as good as the mean when nothing goes wrong, and never catastrophic when something does.

A story with the whole trick in it. In a diving competition, seven judges each give a score, and the result is the average of the five middle scores: the highest and the lowest are dropped. One judge who favours a diver from their own country, or who presses the wrong key, cannot move the result much, because an extreme score is simply not counted. When all judges are honest, dropping two of seven costs a little precision. When one is not, it saves the competition. That is the whole idea of robustness: give up a little efficiency in the ideal case to limit how much any single reading can do.

In physics the dishonest judges are outliers. A CCD exposure collects a cosmic-ray hit in a few per cent of its pixels, always adding charge, never removing it. A gravitational-wave detector has short loud glitches of instrumental origin. An astronomer averaging repeated photometry "clips" readings more than three standard deviations from the median, and the Bayesian answer to the same problem is a likelihood with heavier tails than a Gaussian [HobsonBMC §2.4, pp. 52–53], [Sivia (3.72), p. 67]. The idea of the section is to make "limit the influence of any single reading" precise, measure it, and build estimators around it.

The story in symbols. Add one extra reading, of value x, to n readings with mean $\bar{x}$. The new mean is $(n\bar{x} + x)/(n + 1)$, so it moves by

$$\frac{n\bar{x} + x}{n + 1} - \bar{x} = \frac{x - \bar{x}}{n + 1}, \tag{T5.33}$$

which grows without limit as x does: one reading at a million moves the mean by a million over $n + 1$. The median moves by at most half the gap between the two middle readings, whatever x is. Between the two lies a dial: Huber's estimator lets a reading pull like the mean while it is within k standard deviations of the centre, and with a constant pull beyond. With $k = 1.345$ it is 0.950 times as efficient as the mean for Gaussian noise, a price of five per cent. In return, when 10% of the readings are wild (symmetric, ten times wider), its mean squared error is $1.48/n$ against the mean's $10.82/n$, seven times smaller. Lesson: bounding the pull of each reading costs little when the data are clean and saves a great deal when they are not.

A picture: three ways to pull. Figure T5.11 draws, for three estimators, the penalty $\rho(e)$ that each reading pays for lying a distance e from the estimate, and its slope, the pull $\psi(e) = \rho'(e)$. The mean minimises $\sum_i \rho(x_i - \mu)$ with $\rho = \frac{1}{2}e^2$, a parabola, so far readings pay enormously and pull hard. The median minimises $\sum_i |x_i - \mu|$, a V, so every reading pulls with the same strength. Huber's penalty is the parabola in the middle and the V outside: it behaves like the mean for the bulk of the data and like the median for the few readings far out.

One idea, many names. Estimators defined by a penalty ρ, or by the balance of pulls $\sum_i \psi(x_i - \mu) = 0$, are **M-estimators**, "maximum-likelihood type" because $\rho = -\ln f$ gives maximum likelihood [CasellaBerger §10.2.2, pp. 391–392]. The pull at distance x, suitably normalised, is the **influence function** of Hampel [131]. The fraction of wild readings an estimator can survive is its **breakdown point**. Sigma clipping, trimmed means, Tukey's biweight and the iteratively reweighted fits of the second half of the chapter are all members of this family, and Huber's 1964 paper started it [132].

We first measure the pull of one reading exactly, then use it to compute the variance of any M-estimator, build Huber's estimator, measure how many wild readings each estimator survives, and finally weigh the price against the gain by simulation.

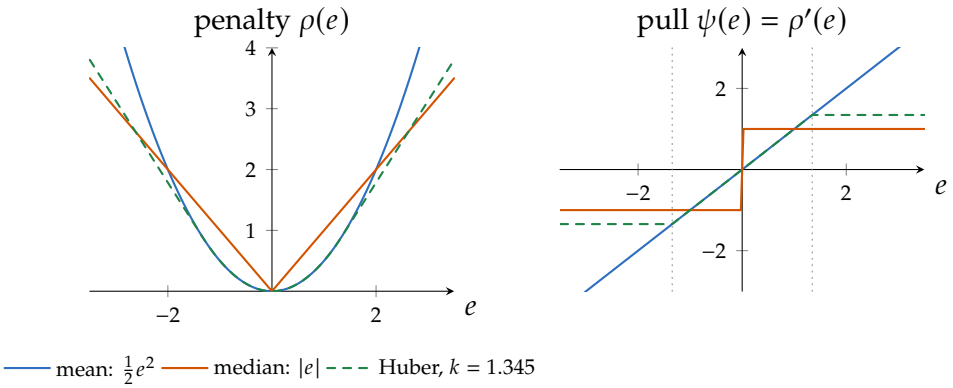


Figure T5.11: Left: the penalty each reading pays for lying a distance e (in units of the noise width) from the estimate. Right: its slope, the pull the reading exerts. The mean's pull grows without limit; the median's is the same at every distance; Huber's grows like the mean's up to $|e| = k$ (dotted lines) and is constant beyond.

T5.5.1 The influence function

Working definition T5.3 (influence function)

Let $T(F)$ be what an estimator converges to when the readings follow the distribution F [for the mean, the population mean of F]. Mix in a small fraction δ of readings that all sit at one value x, $F_\delta = (1-\delta)F + \delta\,\Delta_x$, where Δ_x puts all its probability at x. The **influence function** is $\mathrm{IF}(x) = \lim_{\delta\to 0}[T(F_\delta) - T(F)]/\delta$, the rate at which the estimate moves per unit of contamination at x [CasellaBerger Def. 10.6.1, p. 419]. Operationally: add one reading at x to a large sample of n and multiply the change in the estimate by $n+1$.

The mean. The population mean of F_δ is $(1-\delta)\mu + \delta x$, so $T(F_\delta) - T(F) = \delta(x-\mu)$ and $\mathrm{IF}(x) = x - \mu$: unbounded, the finite-sample shift (T5.33) in the limit.

The median. The median m_δ of F_δ is where its cumulative distribution reaches one half. If x lies above the median, the added mass is not yet counted at m_δ, so

$$\begin{aligned}(1-\delta)F(m_\delta) = \tfrac{1}{2} &\overset{(1)}{=} \quad F(m_\delta) = \frac{1}{2(1-\delta)} = \tfrac{1}{2} + \tfrac{1}{2}\delta + O(\delta^2) \\ &\overset{(2)}{=} \quad m_\delta - m = \frac{\delta}{2f(m)} + O(\delta^2),\end{aligned} \tag{T5.34}$$

(1) Divide by $1-\delta$ and expand $1/(1-\delta) = 1 + \delta + O(\delta^2)$.
(2) Expand F around the median m, where $F(m) = \frac{1}{2}$: $F(m_\delta) = \frac{1}{2} + f(m)(m_\delta - m) + \dots$, and solve.

If x lies below the median, the added mass is counted: $(1-\delta)F(m_\delta) + \delta = \frac{1}{2}$, and the same steps give $m_\delta - m = -\delta/(2f(m))$. So $\mathrm{IF}(x) = \mathrm{sgn}(x-m)/(2f(m))$ [CasellaBerger Ex. 10.6.2, p. 419]: bounded, and the same for a reading just above the median as for one at infinity.

Variances from influence. An estimator is, for large n, its limit plus the average of the influences of the individual readings, $T \approx T(F) + \frac{1}{n}\sum_i \mathrm{IF}(x_i)$, so its variance is $\mathbb{E}[\mathrm{IF}(X)^2]/n$. For the mean this is σ^2/n. For the median it is $\mathbb{E}[1/(4f(m)^2)]/n = 1/(4nf(m)^2)$, the formula used throughout Section T5.1. The influence function therefore measures two things at once: how much one bad reading can do (its largest value), and how precise the estimator is on good data (its mean square).

T5.5.2 M-estimators: one pull, one variance

An M-estimator solves $\sum_i \psi(x_i - \hat{\mu}) = 0$ [CasellaBerger (10.2.3), p. 392]. Take ψ odd and F symmetric about μ, so that $\mathbb{E}_F\,\psi(X-\mu) = 0$ and the population version of the equation is solved by the true centre. Under F_δ the solution μ_δ satisfies

$$\begin{aligned} 0 &\overset{(1)}{=} (1-\delta)\,\mathbb{E}_F\,\psi(X-\mu_\delta) + \delta\,\psi(x-\mu_\delta) \\ &\overset{(2)}{=} (1-\delta)\big[\mathbb{E}_F\,\psi(X-\mu) - (\mu_\delta-\mu)\,\mathbb{E}_F\,\psi'(X-\mu)\big] + \delta\,\psi(x-\mu) + O(\delta^2) \\ &\overset{(3)}{=} -(\mu_\delta-\mu)\,\mathbb{E}_F\,\psi'(X-\mu) + \delta\,\psi(x-\mu) + O(\delta^2), \end{aligned} \tag{T5.35}$$

(1) The estimating equation, averaged over the mixture F_δ.
(2) Expand $\psi(X-\mu_\delta)$ to first order in the small shift $\mu_\delta - \mu$; in the last term the shift only adds $O(\delta^2)$.
(3) $\mathbb{E}_F\,\psi(X-\mu) = 0$, and $\mu_\delta - \mu$ is itself of order δ, so $\delta(\mu_\delta-\mu)$ is $O(\delta^2)$.

Dividing by δ and letting $\delta \to 0$ gives $\mathrm{IF}(x) = \psi(x-\mu)/\mathbb{E}\,\psi'(X-\mu)$, and the variance rule of Section T5.5.1, $\mathbb{E}[\mathrm{IF}^2]/n$, then gives $n\,\mathrm{Var}\,\hat{\mu} = \mathbb{E}\,\psi^2/(\mathbb{E}\,\psi')^2$.

Key idea: Influence and variance of an M-estimator

The M-estimator with pull ψ has $\mathrm{IF}(x) = \psi(x-\mu)/\mathbb{E}\,\psi'(X-\mu)$ and, for large n, $n\,\mathrm{Var}\,\hat{\mu} = \mathbb{E}\,\psi^2/(\mathbb{E}\,\psi')^2$. Its influence is the pull itself, rescaled; a bounded pull means a bounded influence.

This variance is the formula of [CasellaBerger (10.2.6), p. 393], which they derive instead by a Taylor expansion of the estimating equation. Two checks: $\psi(e) = e$ gives $\mathbb{E}\,\psi' = 1$ and $n\,\mathrm{Var} = \sigma^2$, the mean. $\psi = \mathrm{sgn}$ gives $\psi' = 2\,\delta_{\mathrm{D}}(e)$ (the jump of height two at zero, δ_{D} the Dirac delta), so $\mathbb{E}\,\psi' = 2f(\mu)$ and $n\,\mathrm{Var} = 1/(4f(\mu)^2)$, the median.

Example: no M-estimator beats maximum likelihood. Integrating by parts, with f the noise density and $s(e) = -f'(e)/f(e)$ the maximum-likelihood pull,

$$\mathbb{E}\,\psi' \overset{(1)}{=} \int \psi'(e) f(e)\,\mathrm{d}e \overset{(2)}{=} -\int \psi(e) f'(e)\,\mathrm{d}e \overset{(3)}{=} \mathbb{E}[\psi\, s] \overset{(4)}{\le} \sqrt{\mathbb{E}\,\psi^2}\sqrt{\mathbb{E}\,s^2} = \sqrt{\mathbb{E}\,\psi^2\, I}, \tag{T5.36}$$

(1) Definition of the expectation.
(2) Integration by parts; ψf vanishes at $\pm\infty$.
(3) $-f' = s\,f$.
(4) Cauchy–Schwarz, and $\mathbb{E}\,s^2 = I$ is the Fisher information of one reading.

so $n\,\mathrm{Var}\,\hat{\mu} = \mathbb{E}\,\psi^2/(\mathbb{E}\,\psi')^2 \ge 1/I$, with equality only when $\psi \propto s$: an M-estimator reaches the Cramér–Rao bound only if its pull is the maximum-likelihood pull of the true noise law [CasellaBerger (10.2.9), p. 394]. Robustness is bought with efficiency, never the other way round.

T5.5.3 Huber's estimator

Huber's penalty is [CasellaBerger (10.2.2), p. 391]

$$\rho_k(e) = \begin{cases} \tfrac{1}{2}e^2, & |e| \le k, \\ k|e| - \tfrac{1}{2}k^2, & |e| > k, \end{cases} \qquad \psi_k(e) = \rho_k'(e) = \begin{cases} -k, & e < -k, \\ e, & |e| \le k, \\ k, & e > k, \end{cases} \tag{T5.37}$$

with the constant $-\frac{1}{2}k^2$ chosen so that the two pieces of ρ_k meet at $|e| = k$ with the same value and the same slope. The readings are measured in units of the noise width, $e = (x_i - \mu)/\hat{\sigma}$, where $\hat{\sigma}$ must itself be robust; the usual choice is the median absolute deviation from the median, divided by 0.6745 so that it estimates σ for Gaussian data. The estimate $\hat{\mu}$ minimises $\sum_i \rho_k(e_i)$.

Where does this shape come from? It is maximum likelihood for a particular noise law: $\psi_k = -(\ln f_0)'$ for $f_0(e) \propto e^{-\rho_k(e)}$, a density that is Gaussian in the centre and exponential (Laplace-like) in the tails. Huber showed that among all contaminated Gaussians $(1-\epsilon)\mathcal{N}(0,1) + \epsilon H$, with H anything at all, this f_0 is the hardest one to estimate a centre from, and that its maximum-likelihood estimator has the smallest worst-case variance over the whole family [132]. The tuning constant and the contamination are linked by $2\varphi(k)/k - 2[1-\Phi(k)] = \epsilon/(1-\epsilon)$, with φ, Φ the standard Gaussian density and distribution function; $k = 1.345$ corresponds to $\epsilon = 0.058$.

Example: the price at Gaussian noise. For $\mathcal{N}(0,1)$ noise and the variance formula of the key idea above, $\mathbb{E}\,\psi_k' = \mathrm{P}(|Z| \le k) = 2\Phi(k) - 1$, and

$$\mathbb{E}\,\psi_k^2 \overset{(1)}{=} \int_{-k}^{k} z^2 \varphi(z)\,\mathrm{d}z + 2k^2\big[1-\Phi(k)\big] \overset{(2)}{=} \big[2\Phi(k)-1\big] - 2k\varphi(k) + 2k^2\big[1-\Phi(k)\big], \tag{T5.38}$$

(1) Inside $|z| \le k$ the pull is z; outside it is $\pm k$, on two tails of probability $1-\Phi(k)$ each.

(2) Integrate by parts with $z\varphi(z) = -\varphi'(z)$: $\int_{-k}^{k} z \cdot z\varphi\,\mathrm{d}z = [-z\varphi]_{-k}^{k} + \int_{-k}^{k} \varphi\,\mathrm{d}z = -2k\varphi(k) + 2\Phi(k) - 1$.

For $k = 1.345$: $\mathbb{E}\,\psi_k' = 0.8214$ and $\mathbb{E}\,\psi_k^2 = 0.7102$, so $n\,\mathrm{Var}\,\hat{\mu} = 0.7102/0.8214^2$ and the efficiency relative to the mean is $(\mathbb{E}\,\psi_k')^2/\mathbb{E}\,\psi_k^2 = 0.950$. This is why 1.345 is the standard choice: exactly a 5% loss at the Gaussian.

Computing it. $\sum_i \rho_k(e_i)$ is a convex function of μ (each term is), so it has one minimum, where its slope $-\frac{1}{\hat{\sigma}}\sum_i \psi_k(e_i)$ crosses zero. The sum $\sum_i \psi_k((x_i-\mu)/\hat{\sigma})$ never increases as μ increases, is positive below the smallest reading and negative above the largest, so the root can be found by bisection:

Listing T5.2: Huber's estimate by bisection on the balance of pulls (`code/chT5/lib_t5a.py`)

```
39 def huber_location(x, k=1.345, scale=None, iters=60):
40     """Root of sum_i psi_k((x_i-mu)/s) = 0 by bisection. psi_k(u) = clip(u, -k, k) is
41     non-decreasing in u, so the sum is non-increasing in mu and the root is bracketed by the
42     smallest and largest reading. 60 halvings shrink the bracket by 2^-60."""
43     x = np.atleast_2d(x)
44     s = mad_scale(x) if scale is None else np.broadcast_to(scale, (len(x),))
45     lo, hi = x.min(axis=1), x.max(axis=1)
46     for _ in range(iters):
47         mid = 0.5 * (lo + hi)
48         g = np.clip((x - mid[:, None]) / s[:, None], -k, k).sum(axis=1)
49         lo = np.where(g > 0, mid, lo)        # sum still positive: the root lies above mid
50         hi = np.where(g > 0, hi, mid)
51     return 0.5 * (lo + hi)
```

Line 44 takes the scale from the median absolute deviation (or a known scale, if given). Line 45 brackets the root between the smallest and largest reading. Each pass of the loop (lines 46–50) evaluates the total pull at the midpoint and keeps the half of the bracket where the pull changes sign; sixty halvings make the bracket negligibly small. The second half of the chapter computes the same estimate a different way, as a weighted mean whose weights are recomputed until they stop changing.

Example: eight good readings and three outliers. [CasellaBerger Ex. 10.2.5, p. 391] uses eight standard Gaussian readings plus three outliers, 3, 6 and 9, with mean 1.34 and median −0.21. Minimising their criterion with the scale fixed at one gives, for $k = 1, 2, 3, 4, 5, 6$, the estimates −0.00, 0.34, 0.64, 0.86, 1.07 and 1.17, moving steadily from the median towards the mean as k grows; from $k = 8$ on, every residual is below k and the estimate is the mean.[1] The point of the example survives: k is a dial between the median ($k \to 0$) and the mean ($k \to \infty$).

T5.5.4 Breakdown: how many wild readings an estimator survives

The influence function describes one reading at a time. The opposite extreme asks how many readings can go bad together. The **breakdown point** of an estimator is the largest fraction of the readings that can be moved to arbitrary values (to infinity) while the estimate stays bounded [CasellaBerger Def. 10.2.2, p. 389]. The mean breaks with one reading, by (T5.33): its breakdown point is 0. The median survives as long as fewer than half the readings are moved, because the middle reading is then still one of the original ones, trapped between the smallest and largest of them: breakdown point $\frac{1}{2}$, the best possible (beyond half, nobody can say which readings are the good ones). A trimmed mean that drops the α fraction at each end survives until more than that fraction is moved: breakdown point α. Huber's estimator with a fixed scale has a bounded pull, but the scale must also survive; with the median absolute deviation, which itself breaks only at one half, the breakdown point is $\frac{1}{2}$.

The right panel of Figure T5.12 shows this on a sample of 20 readings, replacing m of them by 1000. The mean breaks at $m = 1$, the 10% trimmed mean at $m = 3$ (it drops two from each end), and the median and Huber's estimator only at $m = 10$, half the sample. The left panel shows the influence functions as finite-sample sensitivity curves: one reading added at x to a typical sample (the 20 quantiles of a Gaussian). The mean's curve is a straight line; the others level off, Huber's at $\pm k/\mathbb{E}\,\psi_k' = \pm 1.345/0.8214$, as the key idea predicts.

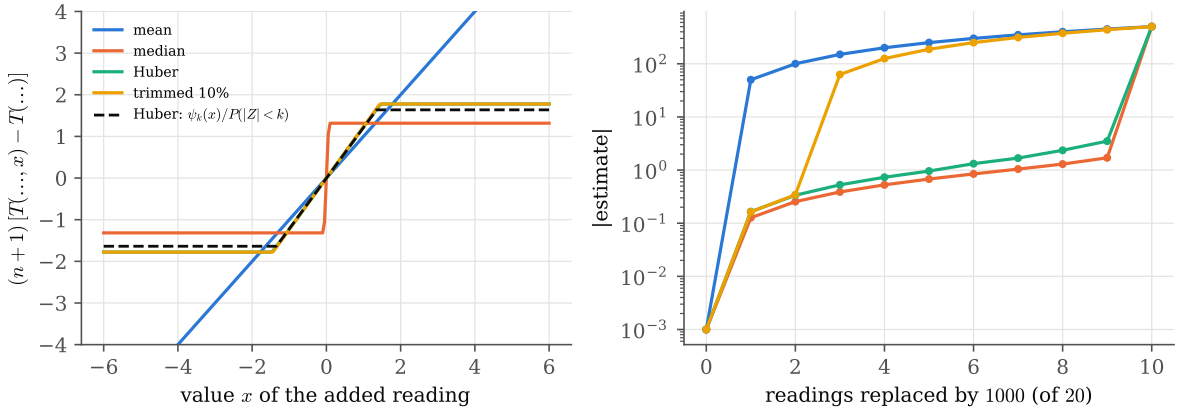


Figure T5.12: Left: how far one extra reading at x moves four estimators of the centre of 20 Gaussian readings, times $n + 1$ (the finite-sample influence function). Only the mean follows the reading out without limit; the dashed line is Huber's influence function $\psi_k(x)/\mathbb{E}\,\psi_k'$. Right: breakdown. The estimate as m of the 20 readings are replaced by 1000, on a logarithmic scale. The mean jumps at once, the trimmed mean after three, the median and Huber's estimator only at ten.

[1] The table printed with this example in [CasellaBerger Table 10.2.3] gives different intermediate values (e.g. −0.04 at $k = 2$ and 0.97 at $k = 8$). Its values cannot be minima of the stated criterion: at $k = 2$ the eight good readings minus 0.34 have pulls summing to −6.00 and the three outliers each pull +2, so 0.34 balances; and at $k = 8$ the largest residual from the mean, $9 - 1.34$, is below 8, so the minimum is the mean itself.

T5.5.5 The price and the gain, side by side

The script `code/chT5/05_robust.py` draws 20000 samples of $n = 50$ readings from a unit Gaussian, replaces a fraction ϵ of them by wild readings, and compares four estimators of the centre (true value zero). The wild readings are either symmetric, $\mathcal{N}(0, 10^2)$, or one-sided like cosmic rays, $|\mathcal{N}(0, 10^2)|$, which only ever add. Figure T5.13 shows n times the mean squared error.

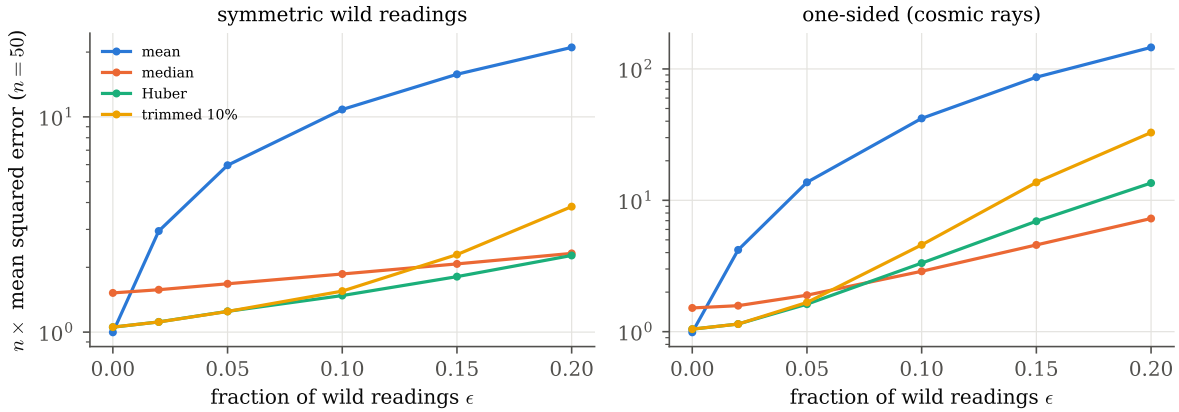


Figure T5.13: n times the mean squared error of four estimators of a centre ($n = 50$), against the fraction ϵ of wild readings. Left: wild readings symmetric about the centre. Right: wild readings that only add, like cosmic rays. At $\epsilon = 0$ the mean is best, Huber's estimator and the trimmed mean lose 5%, the median 50%. With contamination the mean is soon the worst by an order of magnitude.

No contamination. The mean gives 0.998, Huber's estimator 1.054 and the trimmed mean 1.055 (both five per cent worse, as computed above for Huber), the median 1.521, close to $\pi/2$. This is the price of protection.

Symmetric contamination. With 10% wild readings the mean reaches 10.82, while Huber's estimator gives 1.48, the trimmed mean 1.55 and the median 1.86. Symmetric outliers do not bias any of the four; they only inflate the variance, and the bounded pulls stop them from doing so.

One-sided contamination. Now every estimator is biased, because the contaminated distribution itself has moved: there is no longer a unique "centre" for an estimator to find, and each estimator settles somewhere between the core and the outliers. With 5% cosmic rays the mean's error, 13.71, is dominated by its bias; Huber's estimator (1.62), the trimmed mean (1.67) and the median (1.90) are hardly hurt. At 20% the ordering changes: the median, 7.27, beats Huber's estimator, 13.54, because the bias of an estimator under a small contamination is roughly ϵ times the largest value of its influence function, and the median's largest influence, $\sqrt{2\pi}/2 \approx 1.25$, is smaller than Huber's, $1.345/0.8214 \approx 1.64$. (The mean, at 145.8, is off the scale of usefulness.)

Watch out: Robust is not the same as right

A robust estimator limits the damage of readings that do not fit the model; it does not explain them. With one-sided contamination it is still biased, only less so. If the outliers have a physical cause (cosmic rays, a glitch population), the best answer models that cause, either by flagging the readings with independent information (a cosmic-ray detection algorithm, a glitch veto) or by a likelihood that contains the contaminating population, as maximum likelihood did in laboratory C. And never simply delete readings because they

disagree with the answer one expected: that is a selection effect, and it makes error bars too small.

When to use a robust estimator. Use it when the bulk of the noise is understood but rare failures are not: as a default for repeated readings of one quantity, for a first fit that identifies outliers before a careful analysis, and whenever an analysis must run unsupervised on data no one will inspect reading by reading. Choose the median when as many as a third of the readings may be bad, Huber's estimator (or a trimmed mean) when a few per cent may be, and maximum likelihood with an explicit contamination model when the failures are understood well enough to describe.

T5.6 When the parameter lives in the covariance: quadratic estimators

What we already have. Every estimator so far found a parameter that *shifts* the data: a level μ, a pair of masses, the coefficients of a fit. Those live in the mean of the data, $\mathbb{E}\,\boldsymbol{d} = \mathbf{A}\boldsymbol{\theta}$, and a weighted sum of the readings can pick them out. We have also met, case by case, parameters of a different kind. The angular power spectrum C_ℓ was estimated by averaging the squares of the harmonic coefficients, with the noise power subtracted (Eqs. (7.72) and (7.82)). Masked skies needed the pseudo-C_ℓ machinery of Section 8.11, and Eq. (8.3) quoted a "quadratic maximum-likelihood" formula as one Newton step of the likelihood.

When that is not enough. A parameter that sets how much the data *scatter* has no effect on their mean: the sky fluctuations average to zero whatever their amplitude. So no linear combination of the readings carries information about it on average, and the estimator must be at least quadratic. What is the best quadratic combination, when the sky is masked and the noise differs from pixel to pixel? And something new happens to the noise, which the next story makes plain.

A story with the whole trick in it. A camera on a wooden pier films a moored boat. Suppose we want the boat's average height above the water. Each frame is off a little because the pier shakes, sometimes up and sometimes down, and a long film averages the shaking away. Now suppose instead we want to know *how much the boat rocks*. Every frame shows the boat's motion and the pier's motion added together, and their spreads add too. However long we film, the pier's wobble never cancels: it adds the same amount of apparent rocking in every frame, and a longer film only measures the total, boat plus pier, more precisely. To learn how much the boat rocks we must know how much the pier shakes, and *subtract* it.

That is the difference between estimating a fixed number and estimating a variance. When the target is a fixed number, noise moves each reading in either direction and its effect averages out; it widens the error bar and nothing more. When the target is a variance or a power spectrum, the noise's own variance lands on top of the signal's in every reading, so the same noise now pushes every estimate the same way: it has become a *bias*, and it must be computed and subtracted. A CMB experiment subtracts the detector noise power from the measured spectrum (Eq. (7.82)); a lensing reconstruction subtracts the reconstruction noise N_0 from the spectrum of the lensing map (Section T1.9). The idea of the section is that both are the constant term of one general recipe, a quadratic form of the data minus a noise bias, and that the recipe has a best version whose error is set by the Fisher matrix.

The story in symbols: one pixel. A single pixel holds a signal s of variance p, the parameter, plus independent noise of known variance σ^2. Its reading $x = s + n$ has mean zero and variance $p + \sigma^2$. The natural estimate of p is the square of the reading, minus a constant b:

$$\begin{aligned} \mathbb{E}(x^2 - b) &\overset{(1)}{=} p + \sigma^2 - b, \\ \operatorname{Var}(x^2 - b) &\overset{(2)}{=} \mathbb{E}\,x^4 - (\mathbb{E}\,x^2)^2 \overset{(3)}{=} 3(p+\sigma^2)^2 - (p+\sigma^2)^2 = 2(p+\sigma^2)^2. \end{aligned} \tag{T5.39}$$

(1) $\mathbb{E}\,x^2 = \operatorname{Var} x$ because $\mathbb{E}\,x = 0$; signal and noise variances add for independent pieces.
(2) A constant does not change a variance.
(3) For a Gaussian of variance v, $\mathbb{E}\,x^4 = 3v^2$ (the case of Eq. (7.8) with all four indices equal).

The dial is b. The naive choice $b = 0$ is biased by exactly σ^2, the noise variance, however many pixels we average. The choice $b = \sigma^2$ removes the bias. The price that cannot be removed is in the variance: $2(p + \sigma^2)^2$ contains the noise too, so a noisy pixel gives a poor estimate of p even after debiasing. With $p = 1$ and $\sigma^2 = 0.5$ the script `code/chT5/06_quadratic.py` finds, over 200000 draws, a mean of 1.502 for x^2 and 1.002 for $x^2 - \sigma^2$, and a variance of 4.541 (prediction 4.500); the left panel of Figure T5.15 shows the two distributions. Lesson: for a variance parameter the noise appears twice, once as a bias we can subtract and once as extra scatter we cannot.

A picture: the cloud of readings. With two pixels, plot each pair of readings (x_1, x_2) as a point. With no noise and independent signal of variance p in each, the points form a round cloud of radius about $\sqrt{p}$; the parameter is the *size* of the cloud, and its centre is always the origin, so the average position of the points says nothing about p. Noise stretches the cloud, more along the noisier pixel's axis (Figure T5.14). An estimate of the size must use squared distances, and it should trust the direction in which the stretching by noise is small. That is the whole design problem for quadratic estimators: which squared combinations to use, with what weights, and what to subtract.

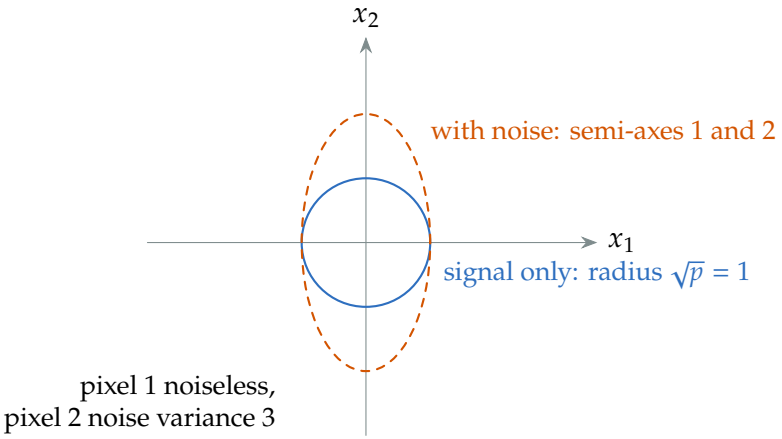


Figure T5.14: Two pixels whose signals are independent with variance $p = 1$. The signal alone fills a round cloud of radius one (blue, one standard deviation). Noise of variance 3 in the second pixel stretches the cloud to a standard deviation of 2 along x_2 (orange, dashed). Both clouds are centred at the origin, so only the squared readings carry information about p, and the squared reading of the noisy pixel carries much less of it.

One idea, many names. A quadratic form of the data minus a constant is the general shape of every variance or spectrum estimator: the sample variance of Eq. (3.27), the full-sky $\widehat{C}_\ell$, the

band powers of a galaxy survey, the pseudo-C_ℓ of a masked map, the cross-spectrum of two detectors, and the lensing reconstruction, which is quadratic in the CMB temperature [57]. Its best version is the **quadratic maximum-likelihood** (QML) estimator [133], whose weights are built from the inverse covariance and whose normalisation is the Fisher matrix. The constant subtracted goes by many names, noise bias, N_ℓ, N_0, but it is always the same object: what the quadratic form would give if the data contained noise alone.

We set up the general quadratic estimator, derive its mean (the unbiasedness condition and the bias term) and its variance for Gaussian data, find the weights that minimise the variance, and test them on the two pixels, on the full sky and on a masked field with uneven noise.

T5.6.1 Quadratic estimators and their noise bias

Let the data $\boldsymbol{d}$ (pixels, Fourier modes, any list of N numbers) have mean zero and a covariance that is a sum of known templates, one per parameter, plus noise:

$$\operatorname{Cov} \boldsymbol{d} = \mathbf{C} = \sum_a p_a \mathbf{C}_a + \mathbf{N}. \tag{T5.40}$$

For a power spectrum, p_a is the band power in band a and $\mathbf{C}_a$ is the covariance a unit band power would produce in the pixels: it encodes the geometry, the mask and the beam. $\mathbf{N}$ is the noise covariance.

Working definition T5.4 (quadratic estimator)

A **quadratic estimator** of p_a is $\hat{q}_a = \boldsymbol{d}^{\mathsf{T}} \mathbf{E}_a \boldsymbol{d} - b_a$, with a symmetric $N \times N$ weight matrix $\mathbf{E}_a$ [which products $d_i d_j$ to add up, and with what weight] and a constant b_a, the **noise bias**. For the one-pixel example, $\mathbf{E} = 1$ and $b = \sigma^2$.

Its mean. For any data of mean zero, Gaussian or not,

$$\mathbb{E}\, \boldsymbol{d}^{\mathsf{T}} \mathbf{E} \boldsymbol{d} \overset{(1)}{=} \sum_{i,j} E_{ij}\, \mathbb{E}\, d_i d_j \overset{(2)}{=} \sum_{i,j} E_{ij} C_{ji} \overset{(3)}{=} \operatorname{tr}(\mathbf{EC}), \tag{T5.41}$$

(1) Write the quadratic form as a double sum; expectation is linear.
(2) $\mathbb{E}\, d_i d_j = C_{ij} = C_{ji}$ for mean-zero data.
(3) $\sum_j E_{ij} C_{ji}$ is the (i,i) entry of $\mathbf{EC}$, and summing over i gives the trace.

so, with (T5.40), $\mathbb{E}\, \hat{q}_a = \sum_b p_b \operatorname{tr}(\mathbf{E}_a \mathbf{C}_b) + \operatorname{tr}(\mathbf{E}_a \mathbf{N}) - b_a$. The estimate is unbiased for p_a, whatever the true values of all the parameters, exactly when

$$\operatorname{tr}(\mathbf{E}_a \mathbf{C}_b) = \delta_{ab} \quad \text{for every } b, \qquad b_a = \operatorname{tr}(\mathbf{E}_a \mathbf{N}). \tag{T5.42}$$

The first condition makes the estimator respond to its own band with unit weight and be blind to the others. The second is the noise bias: the value the quadratic form takes, on average, when the data contain noise alone. It does not shrink with more data; it is a fixed offset that must be known and removed.

T5.6.2 The variance of a quadratic form of Gaussian data

For Gaussian data the four-point average of the readings is the sum over the three ways of pairing them (Isserlis' theorem, Eq. (7.8)), $\mathbb{E}\, d_i d_j d_k d_l = C_{ij}C_{kl} + C_{ik}C_{jl} + C_{il}C_{jk}$. Then, for two

symmetric weight matrices,

$$\begin{aligned}\mathbb{E}\big[(\boldsymbol{d}^{\mathsf{T}}\mathbf{E}\boldsymbol{d})(\boldsymbol{d}^{\mathsf{T}}\mathbf{E}'\boldsymbol{d})\big] &\overset{(1)}{=} \sum_{ijkl} E_{ij}E'_{kl}\big(C_{ij}C_{kl}+C_{ik}C_{jl}+C_{il}C_{jk}\big)\\ &\overset{(2)}{=} \operatorname{tr}(\mathbf{EC})\operatorname{tr}(\mathbf{E}'\mathbf{C})+2\operatorname{tr}(\mathbf{ECE}'\mathbf{C}),\\ \operatorname{Cov}\big(\boldsymbol{d}^{\mathsf{T}}\mathbf{E}\boldsymbol{d},\,\boldsymbol{d}^{\mathsf{T}}\mathbf{E}'\boldsymbol{d}\big) &\overset{(3)}{=} 2\operatorname{tr}(\mathbf{ECE}'\mathbf{C}).\end{aligned} \tag{T5.43}$$

(1) Expand both quadratic forms and insert the three pairings.
(2) The first pairing gives $\sum_{ij}E_{ij}C_{ij}\sum_{kl}E'_{kl}C_{kl}$, a product of traces. The second gives $\sum E_{ij}C_{jl}E'_{lk}C_{ki}$ after using the symmetry $E'_{kl}=E'_{lk}$, which is $\operatorname{tr}(\mathbf{ECE}'\mathbf{C})$; the third gives the same after using $C_{jk}=C_{kj}$.
(3) Subtract the product of the means (T5.41). The constants b_a do not change covariances.

For one pixel, $\mathbf{E}=1$ and $\mathbf{C}=p+\sigma^2$, this is $2(p+\sigma^2)^2$, the result of (T5.39). Note that the *total* covariance $\mathbf{C}$, noise included, appears: the noise adds scatter even after its bias is removed.

T5.6.3 The minimum-variance quadratic estimator

One parameter first. Among all symmetric $\mathbf{E}$ with $\operatorname{tr}(\mathbf{EC}_a)=1$, which has the smallest variance $2\operatorname{tr}(\mathbf{ECEC})$? Change $\mathbf{E}$ by a small symmetric $\delta\mathbf{E}$. The variance changes by $2\operatorname{tr}(\delta\mathbf{E}\,\mathbf{CEC})+2\operatorname{tr}(\mathbf{EC}\,\delta\mathbf{E}\,\mathbf{C})=4\operatorname{tr}(\delta\mathbf{E}\,\mathbf{CEC})$ (the two terms are equal by the cyclic property of the trace), and the constraint by $\operatorname{tr}(\delta\mathbf{E}\,\mathbf{C}_a)$. At the minimum, every change that keeps the constraint must leave the variance unchanged, which with a Lagrange multiplier λ (as in Section 3.7.1) means $4\mathbf{CEC}=\lambda\mathbf{C}_a$. Hence

$$\mathbf{E}_a \overset{(1)}{=} \frac{\lambda}{4}\,\mathbf{C}^{-1}\mathbf{C}_a\mathbf{C}^{-1} \overset{(2)}{=} \frac{\mathbf{C}^{-1}\mathbf{C}_a\mathbf{C}^{-1}}{\operatorname{tr}(\mathbf{C}^{-1}\mathbf{C}_a\mathbf{C}^{-1}\mathbf{C}_a)}, \qquad \operatorname{Var}\hat{q}_a \overset{(3)}{=} \frac{2}{\operatorname{tr}(\mathbf{C}^{-1}\mathbf{C}_a\mathbf{C}^{-1}\mathbf{C}_a)} = \frac{1}{F_{aa}}, \tag{T5.44}$$

(1) Multiply $4\mathbf{CEC}=\lambda\mathbf{C}_a$ by $\mathbf{C}^{-1}$ on both sides.
(2) Fix λ by the constraint $\operatorname{tr}(\mathbf{E}_a\mathbf{C}_a)=1$.
(3) Insert into $2\operatorname{tr}(\mathbf{ECEC})$: the factors $\mathbf{CC}^{-1}$ cancel, leaving $2\operatorname{tr}(\mathbf{C}^{-1}\mathbf{C}_a\mathbf{C}^{-1}\mathbf{C}_a)$ over the square of the same trace.

where $F_{ab}=\frac{1}{2}\operatorname{tr}(\mathbf{C}^{-1}\mathbf{C}_a\mathbf{C}^{-1}\mathbf{C}_b)$ is the Fisher matrix of Gaussian data whose covariance depends on the parameters (Eq. (10.20) with a mean that does not depend on them). So the best quadratic estimator of one parameter, with the others known, reaches the Cramér–Rao bound $1/F_{aa}$. The weights $\mathbf{C}^{-1}\mathbf{C}_a\mathbf{C}^{-1}$ read naturally: first whiten the data by $\mathbf{C}^{-1}$, which down-weights noisy and redundant pixels, then project onto the pattern $\mathbf{C}_a$ that the parameter produces.

All parameters together. With several band powers unknown at once, use the same building blocks for every band,

$$\tilde{q}_a=\tfrac{1}{2}\,\boldsymbol{d}^{\mathsf{T}}\mathbf{C}^{-1}\mathbf{C}_a\mathbf{C}^{-1}\boldsymbol{d}-\tfrac{1}{2}\operatorname{tr}(\mathbf{C}^{-1}\mathbf{C}_a\mathbf{C}^{-1}\mathbf{N}), \qquad \hat{\boldsymbol{p}}=\mathbf{F}^{-1}\tilde{\boldsymbol{q}}. \tag{T5.45}$$

By (T5.41), $\mathbb{E}\,\tilde{q}_a=\sum_b F_{ab}p_b$: each $\tilde{q}_a$ responds to all the bands, mixed by the Fisher matrix, and $\mathbf{F}^{-1}$ unmixes them, so $\mathbb{E}\,\hat{\boldsymbol{p}}=\boldsymbol{p}$. By (T5.43), $\operatorname{Cov}(\tilde{q}_a,\tilde{q}_b)=2\cdot\frac{1}{4}\operatorname{tr}(\mathbf{C}^{-1}\mathbf{C}_a\mathbf{C}^{-1}\mathbf{C}\,\mathbf{C}^{-1}\mathbf{C}_b\mathbf{C}^{-1}\mathbf{C})=F_{ab}$, so $\operatorname{Cov}\hat{\boldsymbol{p}}=\mathbf{F}^{-1}\mathbf{F}\mathbf{F}^{-1}=\mathbf{F}^{-1}$.

Key idea: Quadratic maximum likelihood

For Gaussian data with covariance $\sum_a p_a\mathbf{C}_a+\mathbf{N}$, the estimator (T5.45) is unbiased and has covariance $\mathbf{F}^{-1}$, the Cramér–Rao bound, when $\mathbf{C}$ is evaluated at the true parameters. It is the quadratic analogue of the Gauss–Markov estimator: whiten by $\mathbf{C}^{-1}$, project, subtract

the noise bias, and unmix with the inverse Fisher matrix.

The weights need $\mathbf{C}$, which contains the unknown $\boldsymbol{p}$. In practice one evaluates $\mathbf{C}$ at a fiducial guess. The estimate stays unbiased for any guess, since (T5.41) holds with $\mathbf{F}$ computed from the same guess, but it reaches $\mathbf{F}^{-1}$ only when the guess is right. This is Eq. (8.3) read as one Newton step of the likelihood from the fiducial spectrum, and [HobsonBMC §10.2.3, p. 237] describes the same Newton–Fisher iteration. Iterating it is the subject of the second half of the chapter.

Example: two pixels, one of them noisy. Take the two pixels of Figure T5.14: independent signal of variance $p = 1$ in each, so $\mathbf{C}_p = \mathbb{1}$, and noise variances 0 and 3, so $\mathbf{C} = \mathrm{diag}(1, 4)$. The plain average of the two debiased squares, $\mathbf{E} = \mathrm{diag}(\frac{1}{2}, \frac{1}{2})$ with $b = \frac{3}{2}$, has, by (T5.43), variance $2\,\mathrm{tr}(\mathbf{ECEC}) = 2 \cdot \frac{1}{4}(1^2 + 4^2) = 8.5$. The minimum-variance weights (T5.44) are $\mathbf{C}^{-1}\mathbf{C}_p\mathbf{C}^{-1} = \mathrm{diag}(1, \frac{1}{16})$, normalised by its trace against $\mathbf{C}_p$, $\frac{17}{16}$:

$$\mathbf{E} \overset{(1)}{=} \mathrm{diag}\left(\tfrac{16}{17}, \tfrac{1}{17}\right), \quad b \overset{(2)}{=} \mathrm{tr}(\mathbf{EN}) = \tfrac{3}{17}, \quad \mathrm{Var}\,\hat{q} \overset{(3)}{=} 2\left[\left(\tfrac{16}{17}\right)^2 \cdot 1^2 + \left(\tfrac{1}{17}\right)^2 \cdot 4^2\right] = \frac{32}{17} \approx 1.88.$$

(1) Divide $\mathrm{diag}(1, \frac{1}{16})$ by $\frac{17}{16}$.
(2) $\mathbf{N} = \mathrm{diag}(0, 3)$.
(3) (T5.43) with diagonal matrices: $2\sum_i E_{ii}^2 C_{ii}^2$. Check: $F = \frac{1}{2}(1 + \frac{1}{16}) = \frac{17}{32}$, and $1/F = \frac{32}{17}$.

The best quadratic estimator is four and a half times more precise than the plain average. It is in fact a weighted mean of the two debiased squares, $x_1^2 - 0$ and $x_2^2 - 3$, each an unbiased estimate of p with variance $2\mathbf{C}_{ii}^2$, weighted by its inverse variance as in Section 3.7.1: the weights $\frac{16}{17}$ and $\frac{1}{17}$ are proportional to $1/1^2$ and $1/4^2$. The script finds means 0.993 and 1.002 and variances 8.50 and 1.880 for the two estimators (predictions 8.50 and 1.882).

Example: the full sky. On the full sky with isotropic noise, the $2\ell + 1$ coefficients $a_{\ell m}$ of one multipole are independent with variance $C_\ell + N_\ell$, so for that multipole $\mathbf{C}_\ell = \mathbb{1}$ and $\mathbf{C} = (C_\ell + N_\ell)\mathbb{1}$, both $(2\ell + 1) \times (2\ell + 1)$. Then $\mathbf{C}^{-1}\mathbf{C}_\ell\mathbf{C}^{-1} \propto \mathbb{1}$, the trace normalisation gives $\mathbf{E} = \mathbb{1}/(2\ell + 1)$, the noise bias is $\mathrm{tr}(\mathbf{E}\,N_\ell\mathbb{1}) = N_\ell$, and

$$\hat{C}_\ell = \frac{1}{2\ell + 1}\sum_m |a_{\ell m}|^2 - N_\ell, \qquad \mathrm{Var}\,\hat{C}_\ell = 2\,\mathrm{tr}(\mathbf{ECEC}) = \frac{2(C_\ell + N_\ell)^2}{2\ell + 1}. \tag{T5.46}$$

This is the noise-debiased estimator of Eq. (7.82) and the variance of Eq. (8.29) (there with a beam). On the full sky the plain average of $|a_{\ell m}|^2$ is already optimal, because every coefficient is equally noisy and nothing couples them; the QML machinery earns its keep only when a mask or uneven noise breaks that symmetry.

Example: a masked field with uneven noise. The last part of `code/chT5/06_quadratic.py` builds that broken symmetry in one dimension: a periodic field of 64 pixels, of which a gap of 16 is unobserved, leaving 48; signal power in four bands of wavenumber with true band powers $(8, 4, 1.5, 0.5)$; and white noise of variance 0.2 on a deeply observed stretch of 24 pixels and 4 elsewhere, as a scanning strategy produces. It compares the QML estimator (T5.45) with a *debiased pseudo-spectrum*: the band-averaged power of the gapped data with uniform weights, which is also a quadratic form $\boldsymbol{d}^{\mathsf{T}}\mathbf{E}_a^{\mathrm{ps}}\boldsymbol{d}$, debiased by its own noise term and unmixed by its own coupling matrix $W_{ab} = \mathrm{tr}(\mathbf{E}_a^{\mathrm{ps}}\mathbf{C}_b)$, the one-dimensional analogue of the MASTER estimator of Section 8.11. Both are unbiased by (T5.42).

Listing T5.3: The QML ingredients: covariance, Fisher matrix, weights and noise bias (`code/chT5/06_quadratic.py`)

```
68 C = sum(p * c for p, c in zip(p_true, Ca)) + Nm
69 Ci = np.linalg.inv(C)
70 F = np.array([[0.5 * np.trace(Ci @ Ca[a] @ Ci @ Ca[b]) for b in range(4)] for a in
       range(4)])
71 E_q = [0.5 * Ci @ Ca[a] @ Ci for a in range(4)]
72 b_q = np.array([np.trace(E @ Nm) for E in E_q])
```

Line 68 assembles $\mathbf{C}$ from the four band templates and the noise. Line 70 is the Fisher matrix $F_{ab} = \frac{1}{2}\,\mathrm{tr}(\mathbf{C}^{-1}\mathbf{C}_a\mathbf{C}^{-1}\mathbf{C}_b)$, line 71 the weights $\frac{1}{2}\mathbf{C}^{-1}\mathbf{C}_a\mathbf{C}^{-1}$ and line 72 the noise biases. The script then draws 20000 Gaussian data sets with this covariance and applies both estimators.

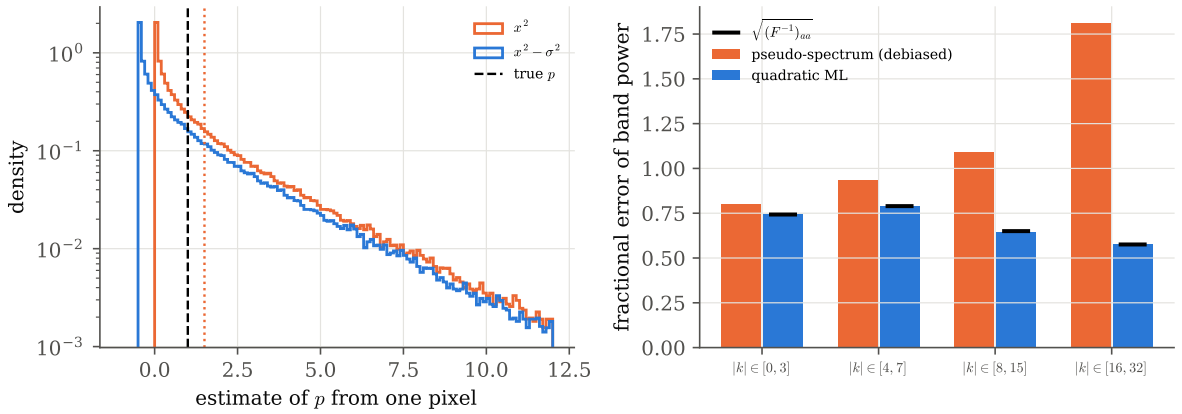


Figure T5.15: Left: one pixel with signal variance $p = 1$ and noise variance 0.5. The square of the reading (orange) is centred on $p + \sigma^2$, not on p; subtracting the noise variance (blue) centres it on p (dashed), but the spread, set by $p + \sigma^2$, stays. Right: the fractional error of four band powers of a gapped field with uneven noise, for the debiased pseudo-spectrum (orange) and the QML estimator (blue); black marks are the Fisher prediction $\sqrt{(\mathbf{F}^{-1})_{aa}}$.

Figure T5.15 (right) shows the result. Both estimators are unbiased. The QML standard deviations, 5.945, 3.148, 0.970 and 0.290 for the four bands, match the Fisher predictions 5.942, 3.158, 0.975 and 0.288. The pseudo-spectrum's variance is larger by factors 1.16, 1.42, 2.84 and 9.77. The gain grows from band to band because the higher bands have less signal and are dominated by noise there, and in a noise-dominated band uniform weights let the shallow, noisy pixels spoil the estimate, while the QML weights follow the two-pixel example and trust the deep pixels. In the low bands the signal dominates every pixel, its own random scatter (cosmic variance) dominates the error, and weighting matters less.

How this is used in real research: Verification versus research

The Monte Carlo above *verifies* a known result: the QML covariance is $\mathbf{F}^{-1}$, and a simulation that reproduces it tests the code. In real analyses the roles change. QML needs $\mathbf{C}^{-1}$ for all observed pixels, an N^3 operation, so it is used where N is small: the large angular scales of a CMB map, degraded to a few thousand pixels, where the mask couples multipoles most strongly. At high ℓ the pseudo-C_ℓ method, with weights chosen per scale (uniform where the signal dominates, inverse noise where the noise does), recovers most of the QML precision at a tiny fraction of the cost, and its covariance is calibrated with simulations, the research instrument when no formula is exact. Hybrid analyses join the two at an intermediate ℓ [134].

T5.6.4 The same structure elsewhere

Cross-spectra need no noise bias. If two detectors see the same sky with independent noises, the product of their readings has mean $\mathbb{E}(s+n_A)(s+n_B) = \mathbb{E}\,s^2$: the noise terms average to zero because the two noises are independent. A quadratic estimator built from products of *different* detectors therefore has $b = 0$, and an error in the noise model cannot bias it. This is why spectra are often computed from cross-correlations of data splits (Section 8.3.3). The noise still adds variance, by (T5.43).

The lensing quadratic estimator. Gravitational lensing correlates CMB modes that would otherwise be independent, and the reconstruction of the lensing potential is a weighted sum of products of pairs of temperature modes, a quadratic estimator whose weights are chosen, as in (T5.44), to minimise its noise [57] (Eq. (T1.24)). Its power spectrum is then *quartic* in the temperature, and the Gaussian, unlensed part of that four-point function contributes a constant, the N_0 bias (Eq. (T1.22)). It plays exactly the role of b in (T5.42): what the estimator returns from data that contain no lensing at all.

When to use a quadratic estimator. Whenever the parameter sets a variance, a power spectrum, a correlation or a noise level rather than a mean, a quadratic estimator is required, and the noise bias must be computed (or avoided with cross-correlations). Use QML when the data set is small enough to invert its covariance and its geometry is irregular; use pseudo-spectra with simulation-calibrated errors when it is large; use the plain average of squared modes when the geometry is so symmetric that the two coincide.

Recap

- An estimator is chosen by answering questions about the problem: what is known about the noise, whether a prior is worth using, whether the parameter lives in the mean or the covariance, how many parameters there are, whether data stream in, and whether the best recipe needs the answer. Same variance, different noise laws: the best estimator of a centre changed by a factor of five.
- Gauss–Markov: for $\boldsymbol{d} = \mathbf{A}\boldsymbol{\theta} + \boldsymbol{n}$ with $\operatorname{Cov}\boldsymbol{n} = \mathbf{C}$, $(\mathbf{A}^\mathsf{T}\mathbf{C}^{-1}\mathbf{A})^{-1}\mathbf{A}^\mathsf{T}\mathbf{C}^{-1}\boldsymbol{d}$ has the smallest covariance among linear unbiased estimators, whatever the noise law; for Gaussian noise it is the best unbiased estimator of all. Nonlinear recipes beat it under heavy tails.
- A Bayes estimator lowers the error where the prior is right and raises it where it is wrong: for a Gaussian pull, it wins when $|\theta - \mu_0| < \sqrt{2\tau^2 + s^2}$. A constant-risk Bayes estimator is minimax. The linear–Gaussian posterior mean is the Wiener filter (ridge regression), with error $P_s P_n/(P_s + P_n)$ per mode and a fainter, biased map.
- Stein: for $k \geq 3$ Gaussian means, shrinking all readings by $1 - (k-2)/|\boldsymbol{X}|^2$ lowers the total risk for every truth, to 2 when all means are zero; it is empirical Bayes with the shrinkage estimated from the spread. The guarantee is for the total; one outlying mean can be estimated three times worse.
- Robust estimators bound the pull of each reading. Influence function $\psi/\mathbb{E}\,\psi'$, variance $\mathbb{E}\,\psi^2/(\mathbb{E}\,\psi')^2 \geq 1/I$; Huber's $k = 1.345$ costs 5% at the Gaussian and saves an order of magnitude with 10% wild readings; the median breaks down only at one half.
- Quadratic estimators $\boldsymbol{d}^\mathsf{T}\mathbf{E}\boldsymbol{d} - b$ find parameters in the covariance. Unbiased when $\operatorname{tr}(\mathbf{E}_a\mathbf{C}_b) = \delta_{ab}$ and $b = \operatorname{tr}(\mathbf{EN})$; variance $2\operatorname{tr}(\mathbf{ECEC})$ for Gaussian data; the best weights $\mathbf{C}^{-1}\mathbf{C}_a\mathbf{C}^{-1}$ normalised by the Fisher matrix reach $\mathbf{F}^{-1}$. For a variance target the noise is a bias as well as a scatter.

T5.7 Iteration: solving the equation that defines an estimator

What we already have. Most estimators met so far are formulas. Put the readings in, get the number out. The weighted mean multiplies each reading by the inverse of its variance (Section 3.7); least squares solves one set of linear equations, the normal equations (Section 3.10); the best linear unbiased estimator of Section T5.2 is a fixed matrix times the data. Behind every one of these formulas stands an equation. The maximum-likelihood estimate is the value at which the score vanishes; the least-squares estimate is the value at which the χ^2 is smallest; the weighted mean is the value at which the weighted residuals sum to zero. For these the equation could be solved with algebra, and the solution, written out, is the formula.

When that is not enough. Often the equation that defines the estimate cannot be solved with algebra. The score of a likelihood may be a sum of nonlinear terms whose root has no closed form; an estimate may need the inverse of a function that cannot be inverted on paper; or the unknown may appear on both sides of the equation. The last case is common with counts: the variance of a Poisson count equals its mean, so the inverse-variance weights that the best combination needs contain the very rates we are trying to measure. In all these cases the estimator is still perfectly well defined: it is the solution of its equation. What is missing is only a way to compute that solution, and the standard way is to repeat a simple step until the answer stops changing. That repetition is what **iteration** means here. It is a numerical method for solving an equation, not a new kind of estimator.

A story with the whole trick in it. What number squares to two? The question defines the number completely, but no finite piece of arithmetic produces it. So we guess, and improve the guess. If x is too large, $2/x$ is too small, and their average is closer than either: $x_{k+1} = \frac{1}{2}(x_k + 2/x_k)$. Starting from 1 this gives 1.5, then 1.41667, then 1.414216, then 1.41421356, and the next pass changes nothing in the digits shown. The answer is $\sqrt{2}$ because $\sqrt{2}$ is the solution of $x^2 = 2$, not because of the particular recipe: halving an interval that brackets the root reaches the same number more slowly, and any route that converges arrives at the same place. Three things can still go wrong, and each depends on the route. Stop after two passes and the answer is 1.41667, a different number from $\sqrt{2}$. Start from -1 and the same recipe walks to $-\sqrt{2}$, the other solution of the same equation. And rewrite the equation as $x = 2/x$ and repeat that step: from 1 it jumps to 2, back to 1, and on forever, because each correction overshoots the error it corrects.

Estimators that need iteration are this story with data in it. The equation is the statistical choice: which equation we write decides the bias and the scatter of the estimate. The iteration only solves it. Newton–Raphson for a score equation, Fisher scoring, iteratively reweighted least squares and the expectation–maximisation algorithm are all recipes like the averaging step, each a different way of improving a guess. The three ways the square-root recipe can fail return as the three places where the route does matter for statistics: stopping early, several solutions, and a step that overshoots.

The story in symbols: two detectors, one source. Each detector counts for the same time and sees the same source, which contributes on average s counts to each. The backgrounds are known from observations away from the source: on average $b_1 = 30$ counts in the first detector and $b_2 = 2$ in the second. The detectors record $n_1 = 44$ and $n_2 = 10$. Subtracting the backgrounds gives two readings of s,

$$x_1 = n_1 - b_1 = 14, \qquad x_2 = n_2 - b_2 = 8,$$

each unbiased, $\mathbb{E}\, x_i = s$, because $\mathbb{E}\, n_i = s + b_i$. Their variances are those of the counts, and for a Poisson count the variance equals the mean: $\operatorname{Var} x_i = s + b_i$. Combining the readings

with weights w_i that add up to one, $\hat{s} = w_1 x_1 + w_2 x_2$, the variance is smallest for $w_i \propto 1/\mathrm{Var}\, x_i$ (Eq. (3.61)), so

$$\hat{s} \overset{(1)}{=} \frac{\sum_i x_i/(s+b_i)}{\sum_i 1/(s+b_i)}, \qquad \mathrm{Var}\,\hat{s} \overset{(2)}{=} \frac{1}{\sum_i 1/(s+b_i)}. \tag{T5.47}$$

(1) Inverse-variance weights $w_i \propto 1/(s+b_i)$, divided by their sum so that they add up to one.
(2) The variance of the inverse-variance weighted mean is one over the sum of the inverse variances (3.61).

The weights contain s. That is the dial we have to set, and there are three obvious ways to set it.

Ignore the difference between the detectors. Equal weights give $\hat{s} = (14+8)/2 = 11$. Its variance is $\frac{1}{4}[(s+b_1)+(s+b_2)]$, which at the value we will end up with is 3.56^2: the clean second detector is diluted by the noisy first.

Take the variances from the data. The observed counts are estimates of $s+b_i$, so put $w_i \propto 1/n_i$. This gives $\hat{s} = (14/44 + 8/10)/(1/44 + 1/10) = 9.111$. It looks harmless but it is not: a detector whose count fluctuated low is given a small variance and hence a large weight, so the estimate is pulled towards readings that fluctuated low. This is the Neyman χ^2 of Section 3.11.3 in disguise, and Section T5.7.2 measures its bias.

Use the current answer in the weights, and repeat. Start from the equal-weight value, $s_0 = 11$, and compute the weights it implies:

$$\begin{aligned} s_1 &\overset{(1)}{=} \frac{14/(11+30) + 8/(11+2)}{1/(11+30) + 1/(11+2)} = \frac{0.34146 + 0.61538}{0.02439 + 0.07692} \overset{(2)}{=} 9.444, \\ s_2 &\overset{(3)}{=} \frac{14/(9.444+30) + 8/(9.444+2)}{1/(9.444+30) + 1/(9.444+2)} = 9.349, \qquad s_3 = 9.343. \end{aligned} \tag{T5.48}$$

(1) (T5.47) with $s = s_0 = 11$ in the weights.
(2) Divide: 0.95684/0.10131.
(3) The same with the new value in the weights; s_3 likewise.

After three passes the value no longer moves in the third decimal. The value that reproduces itself is $\hat{s} = 9.343$, with standard error $\sqrt{1/\sum_i 1/(\hat{s}+b_i)} = 2.967$. Against equal weights the variance has dropped to 0.6948 of its value, as if we had collected almost half as much data again, and the cost was three more lines of arithmetic. The weights contain the unknown, so the estimate is the solution of an equation with s on both sides, $s = g(s)$; feeding each answer back into the weights is the square-root recipe applied to that equation.

What the iteration decides, and what it does not. If the iteration converges, the answer is the solution of the equation, and which numerical route led there does not matter. Three situations make the route matter all the same. *Stopping early*: an iteration halted before it converges returns something that is not the solution of the equation, and the number of passes becomes part of the estimator, a dial that trades bias for scatter (Section T5.7.5). *Several solutions, or none that the iteration can reach*: when the equation has more than one root, as the likelihood of a mixture often does, the starting point decides which one we get (Section T5.7.4), and when the step overshoots, the iteration never settles at all (Section T5.7.1). *The choice of equation*: weights taken from the data, weights taken from the model and minimised directly, and weights re-solved from the current answer are three different equations, hence three different estimators with different biases (Section T5.7.2); the iteration only solves whichever equation we chose, and choosing it is the statistical decision. None of this is the same as updating an estimate when new readings arrive, which consumes new data instead of reusing the same data; that is the recursion of Section T5.9.

A picture: the iteration as a staircase. Write one pass of the recipe as a function, $s_{k+1} = g(s_k)$, where g is the right-hand side of (T5.47) with s_k in the weights. Plot g against the diagonal line $y = s$ (Figure T5.16). A value that reproduces itself, $g(s^*) = s^*$, is a point where the curve crosses the diagonal. The iteration has a simple geometric form. From s_k go up to the curve to read off $g(s_k)$; go across to the diagonal to make that the new s; go up to the curve again. When the curve is flatter than the diagonal where they cross, each step lands closer, and the path is a staircase or a spiral that closes in on the crossing. When the curve is steeper than the diagonal, each step overshoots more than the last, as the over-reacting shower did, and the path spirals out. For the two detectors the curve is almost flat: one step already lands nearly on the crossing.

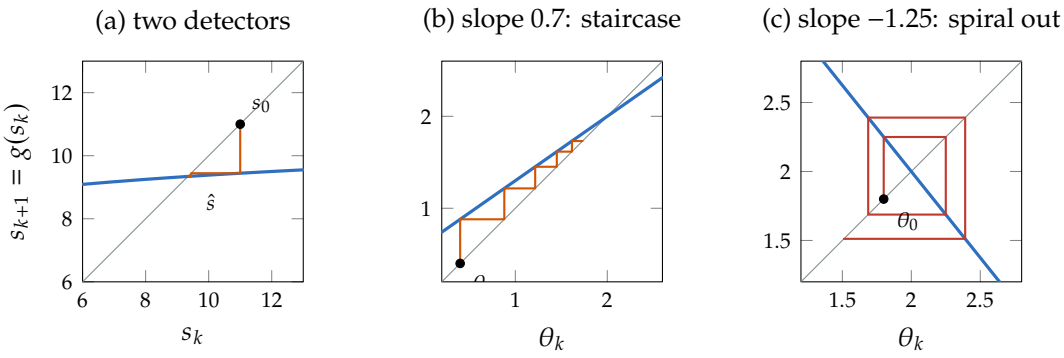


Figure T5.16: Fixed-point iteration drawn as a path between the curve g (blue) and the diagonal (grey). (a) The two-detector recipe: the curve is almost flat, so from $s_0 = 11$ the first step lands at 9.444 and the second at 9.349, already on top of the crossing at $\hat{s} = 9.343$. (b) A curve with slope 0.7 at the crossing: every step removes 30% of the remaining error, a staircase that creeps up. (c) A curve with slope −1.25: every step overshoots by more than the last, and the path spirals away from the crossing.

One idea, many names. Finding a value that reproduces itself under a recipe is called **fixed-point iteration**, and physics is full of it. In the Hartree method of atomic physics, each electron moves in the average field of all the others, but that field depends on the orbitals one is trying to find: guess the orbitals, compute the field, solve for new orbitals, repeat until they agree with the field they produce, the *self-consistent field*. In the mean-field theory of a magnet, the magnetisation sets the field that sets the magnetisation, $m = \tanh(\beta J z m)$, and the equation is solved the same way. In statistics the same idea goes under several names. When the reweighting is done in a least-squares fit, it is called iteratively reweighted least squares. When the unknowns are hidden labels, such as which event was signal, it is the expectation–maximisation (EM) algorithm. In astronomy, sharpening a blurred image this way is Richardson–Lucy deconvolution [135, 136]. In particle physics, undoing a detector's smearing this way is iterative Bayesian unfolding (Section 6.14.5). Newton–Raphson is also a fixed-point iteration, with a particularly good choice of g. The common language is calculus on a single function g, because convergence is decided by its slope.

The questions to answer are those the shower already posed. When does the iteration settle, and how fast (Section T5.7.1)? Is the value it settles on a good estimator, and in particular is it the maximum likelihood estimate (Section T5.7.2)? Then come the two most common forms, reweighted least squares (Section T5.7.3) and EM (Section T5.7.4), and the observation that stopping an iteration early can be useful in its own right (Section T5.7.5).

T5.7.1 When does the iteration settle, and how fast?

Let g be the recipe that turns a guess θ_k into the next guess, $\theta_{k+1} = g(\theta_k)$, and let θ^* be a value it reproduces, $g(\theta^*) = \theta^*$. In the two-detector example θ is the source strength s and g is the weighted mean (T5.47) with the current guess in the weights. How does the error $e_k = \theta_k - \theta^*$ change from one pass to the next?

$$e_{k+1} \overset{(1)}{=} g(\theta_k) - g(\theta^*) \overset{(2)}{=} g'(\xi_k)\,(\theta_k - \theta^*) = g'(\xi_k)\,e_k,$$
$$e_{k+1} \overset{(3)}{=} g'(\theta^*)\,e_k + \tfrac{1}{2}g''(\theta^*)\,e_k^2 + \dots \tag{T5.49}$$

(1) $\theta_{k+1} = g(\theta_k)$ by definition, and $\theta^* = g(\theta^*)$ because θ^* is a fixed point.
(2) The mean value theorem: between any two points the chord of a smooth function has the slope of the function at some point ξ_k in between.
(3) Taylor expansion of $g(\theta_k)$ about θ^*, with $g(\theta^*) = \theta^*$ subtracted.

The first line gives the global statement. If on some interval around θ^* the slope obeys $|g'| \le L < 1$, and the iteration starts inside that interval, every pass multiplies the error by at most L: $|e_k| \le L^k |e_0| \to 0$. The second line gives the speed once we are close. The error is multiplied at each pass by the number $g'(\theta^*)$, so $|e_k| \approx |g'(\theta^*)|^k\,|e_0|$, a straight line on a logarithmic plot. To gain d decimal digits takes about $d/\log_{10}(1/|g'(\theta^*)|)$ passes: one pass per digit for $|g'| = 0.1$, seven for $|g'| = 0.7$, and 230 for $|g'| = 0.99$. The sign of g' decides the shape of the path: positive gives the one-sided staircase of Figure T5.16b, negative an error that changes sign at every pass. And if $|g'(\theta^*)| > 1$ the error grows, the fixed point repels, and no starting point near it converges to it (Figure T5.16c).

Working definition T5.5 (fixed-point iteration and its contraction factor)

A **fixed-point iteration** repeats $\theta_{k+1} = g(\theta_k)$ until θ stops changing. Its **contraction factor** is $|g'(\theta^*)|$ [the fraction of the remaining error that survives one pass, near the fixed point]. Below one the iteration converges linearly, gaining a decimal digit every $1/\log_{10}(1/|g'|)$ passes. Above one it does not converge to θ^*. *Operationally:* record $|\theta_{k+1} - \theta_k|$; near convergence the ratio of successive changes is the contraction factor.

Newton's method is a fixed-point iteration with zero slope. The Newton–Raphson map of (3.84) is $g(\theta) = \theta - \ell'(\theta)/\ell''(\theta)$, with ℓ the log-likelihood. Its slope at the maximum is

$$g'(\theta) \overset{(1)}{=} 1 - \frac{\ell''\,\ell'' - \ell'\,\ell'''}{(\ell'')^2} = \frac{\ell'(\theta)\,\ell'''(\theta)}{\ell''(\theta)^2} \overset{(2)}{=} 0 \quad \text{at } \theta = \hat\theta. \tag{T5.50}$$

(1) Quotient rule for ℓ'/ℓ'': the derivative is $(\ell''\ell'' - \ell'\ell''')/(\ell'')^2$.
(2) At the maximum the score vanishes, $\ell'(\hat\theta) = 0$.

With $g'(\theta^*) = 0$ the first term of (T5.49) disappears, and $e_{k+1} \approx \frac{1}{2}g''(\theta^*)\,e_k^2$: each error is proportional to the square of the previous one, so the number of correct digits roughly doubles per pass. That is the behaviour seen in the lifetime example of Section 3.9.3, with errors 0.057, 0.0052, 0.00005. A recipe whose slope at the fixed point is merely small converges linearly, but fast; a recipe whose slope there is zero converges quadratically.

The slope of a reweighting recipe. Many iterative estimators are weighted means whose weights depend on the current guess, $g(\theta) = \sum_i w_i(\theta)x_i / \sum_i w_i(\theta)$. Their slope at the fixed point has a useful form:

$$g'(\theta) \overset{(1)}{=} \frac{\sum_i w_i' x_i \;\sum_j w_j - \sum_i w_i x_i \;\sum_j w_j'}{\left(\sum_j w_j\right)^2} = \frac{\sum_i w_i'\,(x_i - g(\theta))}{\sum_j w_j},$$

$$g'(\theta^*) \overset{(2)}{=} \frac{\sum_i w_i'(\theta^*)\,(x_i - \theta^*)}{\sum_j w_j(\theta^*)}. \tag{T5.51}$$

(1) Quotient rule, with $w_i' = \mathrm{d}w_i/\mathrm{d}\theta$; then divide numerator and denominator by $\sum_j w_j$ and recognise $g(\theta) = \sum_i w_i x_i / \sum_j w_j$ in the second term.
(2) At the fixed point $g(\theta^*) = \theta^*$.

So the slope is a weighted sum of the residuals $x_i - \theta^*$, weighted by how fast each weight changes. For the two detectors $w_i = 1/(s + b_i)$, so $w_i' = -w_i^2$, and with $\hat{s} + b_1 = 39.343$, $\hat{s} + b_2 = 11.343$,

$$g'(\hat{s}) \overset{(1)}{=} -\frac{\sum_i w_i^2 (x_i - \hat{s})}{\sum_i w_i} \overset{(2)}{=} -\frac{4.657/39.343^2 - 1.343/11.343^2}{1/39.343 + 1/11.343} \overset{(3)}{=} 0.06539. \tag{T5.52}$$

(1) (T5.51) with $w_i' = -w_i^2$.
(2) The residuals are $x_1 - \hat{s} = 14 - 9.343 = 4.657$ and $x_2 - \hat{s} = 8 - 9.343 = -1.343$.
(3) The numerator is $0.003009 - 0.010439 = -0.007430$ and the denominator 0.11358.

Each pass keeps about 6.5% of the remaining error, and the ratio of successive errors in the run of `20_iterative_weights.py` is 0.06513: a new correct digit at every pass, as in (T5.48). Why is the slope so small? At the fixed point the weighted residuals cancel exactly, $\sum_i w_i(x_i - \hat{s}) = 0$, because $\hat{s}$ is their weighted mean. The slope weights the same residuals by w_i^2 instead of w_i, so only the *difference* in weights between readings that fell high and readings that fell low survives. Residuals are noise, of random sign. When the model fits the data, reweighting recipes of this kind converge in a handful of passes.

When it does not. The binned fit of Section T5.7.2 below repeats the same recipe for a peak of known shape on a flat background, in 20000 simulated spectra at each of several exposures. With 5 background counts per bin the median contraction factor is 0.02021, and nine spectra in ten have it below 0.04999. At the lowest exposure, 1.25 background counts per bin, a few spectra are so sparse that the fitted peak sits near the edge of the allowed region, where a small change of the guess changes the weights of the nearly empty bins a lot. In 0.065% of them the iteration had not converged after sixty passes, and every one of those had $|g'(\theta^*)|$ above 0.99, while every spectrum that converged had it below one. The condition $|g'| < 1$ is not a formality.

T5.7.2 Where does it settle? At the maximum of the likelihood

A fixed point is a value that reproduces itself, but nothing so far says it is a *good* value. For the two detectors we can find out what it is by writing out the equation it satisfies. Setting $s_{k+1} = s_k = \hat{s}$ in (T5.47) and clearing the denominator,

$$\sum_i \frac{x_i - \hat{s}}{\hat{s} + b_i} \overset{(1)}{=} 0 \quad\Longleftrightarrow\quad \sum_i \frac{n_i - (\hat{s} + b_i)}{\hat{s} + b_i} \overset{(2)}{=} 0 \quad\Longleftrightarrow\quad \sum_i \left(\frac{n_i}{\hat{s} + b_i} - 1\right) \overset{(3)}{=} 0. \tag{T5.53}$$

(1) Multiply $\hat{s}\sum_i 1/(\hat{s} + b_i) = \sum_i x_i/(\hat{s} + b_i)$ out and move everything to one side.
(2) $x_i = n_i - b_i$.
(3) Split the fraction.

Now ask a different question: what is the maximum-likelihood estimate of s? The counts are independent Poisson variables with means $s + b_i$, so the log-likelihood and its derivative, the score, are

$$\ell(s) \overset{(1)}{=} \sum_i \left[n_i \ln(s + b_i) - (s + b_i) - \ln n_i!\right], \qquad \ell'(s) \overset{(2)}{=} \sum_i \left(\frac{n_i}{s + b_i} - 1\right). \tag{T5.54}$$

(1) The logarithm of the product of Poisson probabilities $(s+b_i)^{n_i}e^{-(s+b_i)}/n_i!$.
(2) Differentiate term by term; the factorials do not depend on s.

The score equation $\ell'(\hat{s})=0$ is the last form of (T5.53). So the self-consistent weighted mean is the maximum-likelihood estimate, which the script confirms by solving the score equation directly: the two agree to 10^{-10}. This explains why the iterated weights beat the two shortcuts. The equal-weight mean ignores what we know about the noise. The data-weighted mean uses the noise of the data to set its own weights. The iterated mean is maximum likelihood, efficient for large samples (Section 3.13.3), reached through a sequence of weighted averages.

The equivalence is not special to two detectors. Let counts y_i have Poisson means $\lambda_i(\theta)$, any smooth functions of the parameter. A weighted least squares fit with weights frozen at the current guess, $\sigma_i^2=\lambda_i(\theta_k)$, minimises $\sum_i (y_i-\lambda_i(\theta))^2/\lambda_i(\theta_k)$, and its stationarity condition, evaluated at a fixed point $\theta_k=\theta$, is compared below with the Poisson score:

$$
\begin{aligned}
-\frac{1}{2}\frac{\partial}{\partial\theta}\sum_i \frac{(y_i-\lambda_i(\theta))^2}{\lambda_i(\theta_k)}\Big|_{\theta_k=\theta} &\overset{(1)}{=} \sum_i \frac{y_i-\lambda_i(\theta)}{\lambda_i(\theta)}\,\frac{\partial\lambda_i}{\partial\theta},\\
\frac{\partial}{\partial\theta}\sum_i \big[y_i\ln\lambda_i(\theta)-\lambda_i(\theta)\big] &\overset{(2)}{=} \sum_i\Big(\frac{y_i}{\lambda_i}-1\Big)\frac{\partial\lambda_i}{\partial\theta} = \sum_i \frac{y_i-\lambda_i}{\lambda_i}\,\frac{\partial\lambda_i}{\partial\theta}.
\end{aligned}
\tag{T5.55}
$$

(1) Only the numerator depends on θ when the weights are frozen; the chain rule gives $-2(y_i-\lambda_i)\,\partial\lambda_i/\partial\theta$. Then set the frozen value equal to the current one.
(2) The Poisson log-likelihood without its θ-free term $-\ln y_i!$, differentiated by the chain rule.

The two right-hand sides are identical.

Key idea: Re-solved weights land on the score equation

For counts whose variance equals their mean, a least-squares fit whose weights $1/\lambda_i(\theta)$ are recomputed from the current fit and the fit repeated until nothing changes converges, if it converges, to the root of the Poisson score: the maximum-likelihood estimate. Neither weights from the data, $1/y_i$, nor weights from the model differentiated along with the residuals have this property.

The two shortcuts, met again. Section 3.11.3 fitted the normalisation of a histogram in two popular ways. The Neyman χ^2 takes each bin's variance from the data, $\sum_i (y_i-\lambda_i)^2/y_i$. The Pearson χ^2 takes it from the model, $\sum_i (y_i-\lambda_i)^2/\lambda_i(\theta)$, and minimises the whole expression in θ. There we found that Neyman's fitted total falls short of the observed total by $\chi^2_{\min}$, and Pearson's exceeds it by $\frac{1}{2}\chi^2_{\min}$, so that with twenty bins the two are off by about twenty and ten events. The same two shortcuts now sit beside a third recipe, the re-solved weights. What do they do to a parameter other than the normalisation, and why are they off in opposite directions?

Take a peak of known shape s_i (with $\sum_i s_i=1$) and unknown total θ on a known background β_i, so $\lambda_i=\theta s_i+\beta_i$ and $\partial\lambda_i/\partial\theta=s_i$. The three recipes solve three different equations:

$$
\begin{aligned}
\text{Neyman:}\quad h_N(\theta) &\overset{(1)}{=} \sum_i s_i\,\frac{y_i-\lambda_i}{y_i}=0,\\
\text{Pearson:}\quad h_P(\theta) &\overset{(2)}{=} \sum_i s_i\Big(\frac{y_i^2}{\lambda_i^2}-1\Big)=0,\\
\text{re-solved (ML):}\quad h_L(\theta) &\overset{(3)}{=} \sum_i s_i\Big(\frac{y_i}{\lambda_i}-1\Big)=0.
\end{aligned}
\tag{T5.56}
$$

(1) Derivative of $\sum_i (y_i - \lambda_i)^2/y_i$, divided by -2.
(2) Derivative of $\sum_i (y_i - \lambda_i)^2/\lambda_i$: the chain rule gives $s_i[-2(y_i - \lambda_i)\lambda_i - (y_i - \lambda_i)^2]/\lambda_i^2$, and $-(y_i - \lambda_i)(2\lambda_i + y_i - \lambda_i) = -(y_i^2 - \lambda_i^2)$; divide by -1.
(3) (T5.55) with $\partial\lambda_i/\partial\theta = s_i$.

Which way does each estimate go? Evaluate each equation at the true θ, where $\mathbb{E}\, y_i = \lambda_i$, and average over the noise. An equation whose average is zero there has its root, on average, at the truth. The likelihood equation is of this kind: $\mathbb{E}(y_i/\lambda_i - 1) = 0$. For Pearson, $\mathbb{E}\, y_i^2 = \operatorname{Var} y_i + (\mathbb{E}\, y_i)^2 = \lambda_i + \lambda_i^2$ for a Poisson count, so $\mathbb{E}\, h_P = \sum_i s_i/\lambda_i > 0$. For Neyman, the function $1/y$ is convex, so its average exceeds the inverse of the average, $\mathbb{E}(1/y_i) > 1/\lambda_i$ (Jensen's inequality), and $\mathbb{E}\, h_N = \sum_i s_i(1 - \lambda_i\, \mathbb{E}[1/y_i]) < 0$. (Neyman's χ^2 is not even defined for an empty bin, so these averages are taken over counts large enough that empty bins do not occur, as the expansion below assumes.) All three functions decrease as θ grows, since λ_i grows. A decreasing function that is positive at the truth crosses zero above it, so Pearson's estimate is biased high, and Neyman's, negative at the truth, is biased low. That is the mechanism in one line: Pearson's recipe can reduce its χ^2 by inflating the variances it divides by, and Neyman's gives the largest weights to bins that fluctuated low.

By how much? The sign came from the value of each equation at the truth; the size needs one more ingredient, because the slope of the equation is itself noisy, and a slope that happens to be steep when the value happens to be large moves the root less. Write the equation at the truth as $H = h(\theta_0)$, its slope there as H' and its curvature as H'', all random through the counts, and let $\Delta = \hat\theta - \theta_0$. For many counts per bin Δ is small, and expanding $h(\hat\theta) = 0$ to second order gives

$$0 \overset{(1)}{=} H + H'\Delta + \tfrac{1}{2}H''\Delta^2, \qquad \Delta \overset{(2)}{=} -\frac{H}{H'} - \frac{H''}{2H'}\Big(\frac{H}{H'}\Big)^2,$$
$$\mathbb{E}\,\Delta \overset{(3)}{=} -\frac{\mathbb{E}\, H}{\mathbb{E}\, H'} + \frac{\operatorname{Cov}(H, H')}{(\mathbb{E}\, H')^2} - \frac{\mathbb{E}\, H''\, \operatorname{Var} H}{2(\mathbb{E}\, H')^3}. \tag{T5.57}$$

(1) Taylor expansion of h about the truth, evaluated at the root.
(2) Solve by iteration: to first order $\Delta = -H/H'$; insert that into the quadratic term.
(3) Write $H' = \mathbb{E}\, H' + \delta H'$, expand $1/H'$ to first order in $\delta H'$, and keep the terms whose average is of the leading order: the mean of H, its correlation with the slope, and its variance times the curvature.

Now insert the three equations of (T5.56). Three sums keep recurring, so name them: $A = \sum_i s_i/\lambda_i$, $B = \sum_i s_i^2/\lambda_i$ and $C = \sum_i s_i^3/\lambda_i^2$. For a Poisson count with large mean, write $y_i = \lambda_i + \epsilon_i$ with $\operatorname{Var}\epsilon_i = \lambda_i$; to the order needed, $\operatorname{Var}(y_i^2) \approx 4\lambda_i^3$ and $\operatorname{Var}(1/y_i) \approx 1/\lambda_i^3$, by the first-order propagation of Section 3.6.3.

$$\text{Pearson:} \quad \mathbb{E}\,\Delta_P \overset{(1)}{=} \frac{A}{2B} + \frac{-8C}{4B^2} - \frac{6C\cdot 4B}{2(-2B)^3} = \tfrac{1}{2}\Big(\frac{A}{B} - \frac{C}{B^2}\Big),$$
$$\text{Neyman:} \quad \mathbb{E}\,\Delta_N \overset{(2)}{=} \frac{-A}{B} + \frac{C}{B^2} - 0 = -\Big(\frac{A}{B} - \frac{C}{B^2}\Big),$$
$$\text{re-solved (ML):} \quad \mathbb{E}\,\Delta_L \overset{(3)}{=} 0 + \frac{-C}{B^2} - \frac{2C\cdot B}{2(-B)^3} = 0. \tag{T5.58}$$

(1) $H = \sum_i s_i(y_i^2/\lambda_i^2 - 1)$ has mean A exactly; its slope $H' = -2\sum_i s_i^2 y_i^2/\lambda_i^3$ has mean $\approx -2B$; its curvature $H'' = 6\sum_i s_i^3 y_i^2/\lambda_i^4$ has mean $\approx 6C$; $\operatorname{Var} H \approx \sum_i s_i^2\, 4\lambda_i^3/\lambda_i^4 = 4B$; and $\operatorname{Cov}(H, H') = -2\sum_i s_i^3 \operatorname{Var}(y_i^2)/\lambda_i^5 \approx -8C$.
(2) $H = \sum_i s_i(1 - \lambda_i/y_i)$ has mean $\approx -A$ (from $\mathbb{E}(1/y_i) \approx (1 + 1/\lambda_i)/\lambda_i$, the second-order bias of Section 3.6.3); $H' = -\sum_i s_i^2/y_i$ has mean $\approx -B$; $H'' = 0$ because h_N is linear in θ; $\operatorname{Cov}(H, H') = \sum_i s_i^3 \lambda_i \operatorname{Var}(1/y_i) \approx C$.

(3) $H = \sum_i s_i(y_i/\lambda_i - 1)$ has mean zero; $H' = -\sum_i s_i^2 y_i/\lambda_i^2$ has mean $-B$; $H'' = 2\sum_i s_i^3 y_i/\lambda_i^3$ has mean $2C$; $\mathrm{Var}\, H = B$; $\mathrm{Cov}(H, H') = -\sum_i s_i^3 \lambda_i/\lambda_i^3 = -C$. The two terms cancel.

So Neyman's bias is exactly -2 times Pearson's at this order, and the likelihood equation has no bias at this order at all. The value of each equation at the truth gives the term A/B; the noisy slope and curvature give the correction C/B^2, which is not small: for a single bin ($s = 1$) the two cancel, as they must, because then both shortcuts return $y - \beta$, which is unbiased. In units of the statistical error $(\sum_i s_i^2/\lambda_i)^{-1/2}$, the Fisher error of θ (Section 3.12.3), both biases fall as one over the square root of the counts per bin. The simulation in `20_iterative_weights.py` fits a Gaussian peak of 20 expected events on 100 background events spread over 20 bins, scaled to several exposures. With 5 background counts per bin the three biases are -0.996, $+0.4381$ and 0.01409 standard errors; with sixteen times the exposure they are -0.2143, $+0.1053$ and -6.637×10^{-4} (Figure T5.17). Neyman's bias is close to twice Pearson's, with the opposite sign, and the ratio approaches -2 as the counts grow, as (T5.58) says. The re-solved weights stay within a few hundredths of a standard error of the truth at every exposure.

Listing T5.4: Re-solved weights for the binned fit, all toy spectra at once (`code/chT5/20_iterative_weights.py`)

```
104     th = th_ney.copy()
105     for k in range(60):
106         lam_k = np.maximum(th[:, None] * shape + beta, 1e-9)
107         th = np.sum(shape * (y - beta) / lam_k, 1) / np.sum(shape**2 / lam_k, 1)
```

Line 104 starts from the Neyman estimate. Line 106 computes the predicted counts $\lambda_i = \theta_k s_i + \beta_i$ for the current guess, kept positive. Line 107 is the weighted least-squares solution for one parameter with weights $1/\lambda_i$: $\theta = \sum_i s_i(y_i - \beta_i)/\lambda_i \,/\, \sum_i s_i^2/\lambda_i$. Sixty passes are far more than needed; the script checks that the result agrees with a direct root search of the Poisson score in every spectrum whose contraction factor is below 0.7.

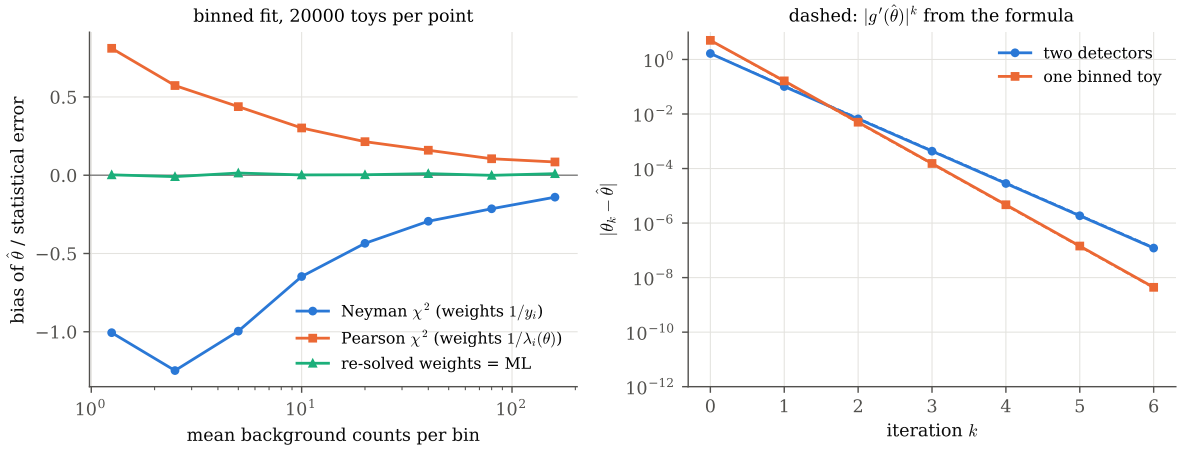


Figure T5.17: Left: the bias of three recipes for the total of a peak on a known background, in units of the statistical error, against the background counts per bin. Weights from the data (Neyman) pull the estimate low, weights from the model minimised directly (Pearson) push it high by about half as much, and weights re-solved from the current fit (the maximum likelihood estimate) are unbiased at every exposure. Right: the error of the iteration against the pass number for the two-detector example and for one simulated spectrum; each is a straight line on the logarithmic scale, with the slope predicted by the contraction factor $|g'(\hat{\theta})|$ (dashed, on top of the points).

The same move elsewhere in the book. Later chapters use re-solved weights without stopping to name them. The fit of the Hubble constant and the infall towards the Virgo cluster weights

each galaxy by the inverse of its velocity variance, which contains H_0 through the distance error, and iterates the weights five times (Section 11.16.3). The cosmic-chronometer fit has a systematic error proportional to the model, so its covariance contains H_0, and it is iterated from a starting value of 70 (Section 11.19.2). Both are the re-solved-weight recipe, and (T5.51) explains why five passes are enough: each variance changes only a little when H_0 moves by its error, so the slope of the recipe is tiny. Fisher scoring, which fits CMB parameters to a power spectrum (Section 10.6.6), is the many-parameter form of Section T5.7.3 below.

T5.7.3 Iteratively reweighted least squares

The two-detector recipe is one case of a pattern that recurs throughout statistics. Many estimators are defined as the solution of an equation of the form

$$\sum_i \psi\big(r_i(\theta)\big)\,\frac{\partial \mu_i}{\partial \theta} = 0, \qquad r_i(\theta) = \frac{y_i - \mu_i(\theta)}{\sigma_i}, \tag{T5.59}$$

where $\mu_i(\theta)$ is the model's prediction for reading y_i, r_i the standardised residual, and ψ a function that says how strongly a residual of a given size pulls on the fit. For least squares $\psi(r) = r$: every residual pulls in proportion to its size. Write any other ψ as $\psi(r) = w(r)\,r$, with $w(r) = \psi(r)/r$. If the weights $w(r_i)$ were constants, (T5.59) would be the normal equation of a weighted least squares fit (Eq. (3.87)), solved in one step. They are not constants, because the residuals depend on θ; but we can freeze them at the current fit, solve, recompute the residuals and the weights, and repeat. This is **iteratively reweighted least squares** (IRLS) [137, 138]. The two-detector recipe was IRLS with $w_i = 1/(s + b_i)$; two more examples show its range.

A robust mean. Section T5.5 introduced the Huber estimator [132] for readings that are mostly well behaved with a few wild ones. It minimises $\sum_i \rho_k(x_i - \theta)$ with

$$\rho_k(r) = \begin{cases} \tfrac{1}{2}r^2, & |r| \le k, \\ k|r| - \tfrac{1}{2}k^2, & |r| > k, \end{cases} \qquad \psi_k(r) = \rho_k'(r) = \begin{cases} r, & |r| \le k, \\ k\,\mathrm{sign}(r), & |r| > k, \end{cases}$$

[CasellaBerger (10.2.2), p. 391]: a residual pulls in proportion to its size up to k, and no harder beyond. The weight is $w(r) = \psi_k(r)/r = \min(1, k/|r|)$, so one IRLS pass for a location ($\mu_i = \theta$, $\partial\mu_i/\partial\theta = 1$, $\sigma_i = 1$) is a weighted mean,

$$\theta_{j+1} = \frac{\sum_i w_i x_i}{\sum_i w_i}, \qquad w_i = \min\Big(1, \frac{k}{|x_i - \theta_j|}\Big). \tag{T5.60}$$

A reading within k of the current fit counts fully; one further out counts in inverse proportion to its distance, so that its pull, weight times residual, is exactly k.

Casella and Berger try this on eleven numbers, eight standard normal readings and three outliers [CasellaBerger Ex. 10.2.5, p. 391]:

$$-1.28,\ -0.96,\ -0.46,\ -0.44,\ -0.26,\ -0.21,\ -0.063,\ 0.39,\ 3,\ 6,\ 9.$$

Their mean is 1.338, dragged up by the outliers, and their median is -0.21. With $k = 1.5$ and the median as starting point, (T5.60) gives $-0.21 \to 0.1165 \to 0.1484 \to \ldots$ and settles at $\hat\theta = 0.1521$ after 10 passes (to 10^{-10}). At the fixed point the eight inliers have weight one and the three outliers have weights $1.5/|x_i - \hat\theta|$, that is 0.527, 0.256 and 0.170 (Figure T5.18, left).

How fast? Use the slope formula (T5.51). An inlier's weight is constant, $w_i' = 0$. An outlier's weight is $w_i = k/|x_i - \theta|$, and

$$w_i'\,(x_i - \theta) \overset{(1)}{=} \frac{k\,\mathrm{sign}(x_i - \theta)}{(x_i - \theta)^2}\,(x_i - \theta) \overset{(2)}{=} \frac{k}{|x_i - \theta|} = w_i, \qquad \text{so} \qquad g'(\hat\theta) \overset{(3)}{=} \frac{\sum_{\text{clipped}} w_i}{\sum_{\text{all}} w_i}. \tag{T5.61}$$

(1) $\mathrm{d}|x-\theta|/\mathrm{d}\theta = -\operatorname{sign}(x-\theta)$, so $\mathrm{d}(k/|x-\theta|)/\mathrm{d}\theta = k\operatorname{sign}(x-\theta)/(x-\theta)^2$.
(2) $\operatorname{sign}(r)\, r = |r|$.
(3) Insert into (T5.51); only clipped readings contribute to the numerator.

The contraction factor is the share of the total weight carried by the clipped readings: $(0.527 + 0.256 + 0.170)/(8 + 0.953) = 0.1064$, and the ratio of successive errors in the run is 0.1064. A few outliers that are far out carry little weight, and the iteration converges fast.

Why it never goes uphill. The function ρ_k depends on r only through $t = r^2$: $\rho_k = \phi(t)$ with $\phi(t) = \frac{1}{2}t$ for $t \le k^2$ and $\phi(t) = k\sqrt{t} - \frac{1}{2}k^2$ beyond. Its slope $\phi'(t)$ is $\frac{1}{2}$ and then $k/(2\sqrt{t})$, never increasing, so ϕ is concave and lies below each of its tangent lines. Writing $r_{i,j}$ for the residuals at θ_j, and noting that $\phi'(r^2) = \frac{1}{2}w(r)$,

$$\sum_i \rho_k(r_{i,j+1}) \overset{(1)}{\le} \sum_i \Big[\rho_k(r_{i,j}) + \tfrac{1}{2}w_i\big(r_{i,j+1}^2 - r_{i,j}^2\big)\Big] \overset{(2)}{\le} \sum_i \rho_k(r_{i,j}). \tag{T5.62}$$

(1) The tangent line of the concave ϕ at $t_0 = r_{i,j}^2$ lies above ϕ everywhere: $\phi(t) \le \phi(t_0) + \phi'(t_0)(t - t_0)$.
(2) θ_{j+1} minimises $\sum_i w_i(x_i - \theta)^2$ with the weights frozen at θ_j, so $\sum_i w_i r_{i,j+1}^2 \le \sum_i w_i r_{i,j}^2$.

The frozen-weight quadratic touches the Huber objective at the current fit and lies above it elsewhere, so going down the quadratic goes down the objective. The same argument works for any ρ that is a concave function of r^2, and the same shape of argument, a bound that touches at the current point, will prove that EM never goes downhill.

Figure T5.18, right, shows the estimate as k runs from 0 to 10: from the median towards the mean, which it reaches at $k = 8$ because then no residual exceeds k. The table printed with this example in Casella and Berger differs at intermediate k; Section T5.5.3 shows why its values cannot be minima. The script checks every value by IRLS against a direct minimisation.

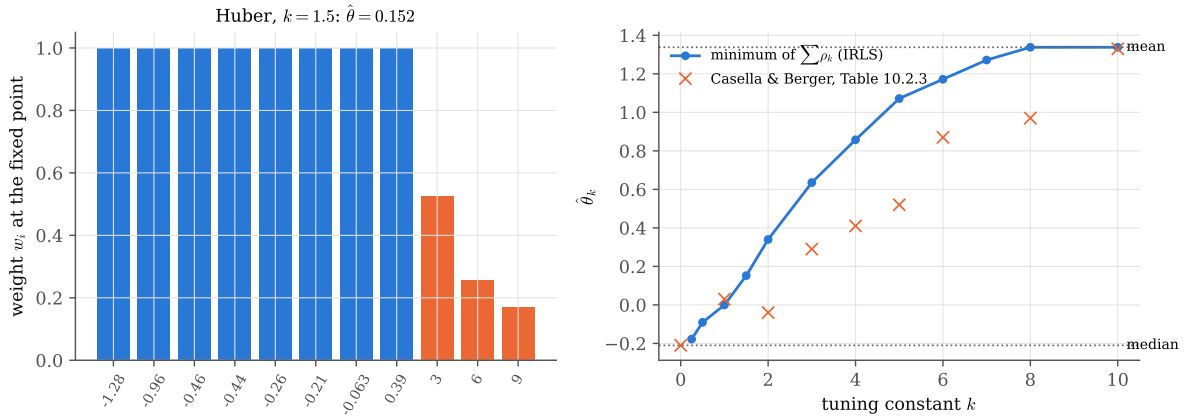


Figure T5.18: The Huber estimate of Casella and Berger's eleven numbers by reweighted least squares. Left: the weights at the fixed point for $k = 1.5$; the eight readings near the centre count fully, the outliers 3, 6 and 9 (orange) count less the further out they lie. Right: the estimate against the tuning constant k, moving from the median ($k \to 0$) to the mean (no residual larger than k), with the values of Casella and Berger's Table 10.2.3 for comparison.

A decay curve with empty bins. A source decays with lifetime τ, and we count its decays in twenty time bins of unit width. The expected count in bin t_i is $\lambda_i = e^{a+bt_i}$, with $b = -1/\tau$: an

exponential, a straight line on a logarithmic scale. The tempting fit is a straight line through $\ln y_i$. But late bins hold a handful of counts, some hold none, and $\ln 0$ does not exist. What does the Poisson likelihood ask for instead? With $\eta_i = a + bt_i = \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}$, $\boldsymbol{x}_i = (1, t_i)^{\mathsf{T}}$ and $\boldsymbol{\beta} = (a, b)^{\mathsf{T}}$,

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} \overset{(1)}{=} \sum_i \Big(\frac{y_i}{\lambda_i} - 1\Big)\frac{\partial \lambda_i}{\partial \boldsymbol{\beta}} \overset{(2)}{=} \sum_i (y_i - \lambda_i)\,\boldsymbol{x}_i, \qquad -\frac{\partial^2 \ell}{\partial \boldsymbol{\beta}\,\partial \boldsymbol{\beta}^{\mathsf{T}}} \overset{(3)}{=} \sum_i \lambda_i\,\boldsymbol{x}_i \boldsymbol{x}_i^{\mathsf{T}} \equiv X^{\mathsf{T}} W X,$$

$$\boldsymbol{\beta}_{\text{new}} \overset{(4)}{=} \boldsymbol{\beta} + (X^{\mathsf{T}} W X)^{-1} X^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{\lambda}) \overset{(5)}{=} (X^{\mathsf{T}} W X)^{-1} X^{\mathsf{T}} W \boldsymbol{z}, \qquad z_i = \eta_i + \frac{y_i - \lambda_i}{\lambda_i}. \tag{T5.63}$$

(1) The Poisson score (T5.55), with a vector parameter.
(2) $\partial \lambda_i / \partial \boldsymbol{\beta} = e^{\eta_i} \boldsymbol{x}_i = \lambda_i \boldsymbol{x}_i$.
(3) Differentiate $(y_i - \lambda_i)\boldsymbol{x}_i$ once more: $-\lambda_i \boldsymbol{x}_i \boldsymbol{x}_i^{\mathsf{T}}$. The rows $\boldsymbol{x}_i^{\mathsf{T}}$ form the matrix X, and $W = \operatorname{diag}(\lambda_i)$. The data y_i have dropped out, so this observed curvature equals its average, the Fisher matrix.
(4) A Newton step (Section 3.9.3): current point plus inverse curvature times score.
(5) Write $\boldsymbol{\beta} = (X^{\mathsf{T}} W X)^{-1} X^{\mathsf{T}} W X \boldsymbol{\beta}$ and $\boldsymbol{y} - \boldsymbol{\lambda} = W W^{-1}(\boldsymbol{y} - \boldsymbol{\lambda})$, and collect: $X\boldsymbol{\beta} = \boldsymbol{\eta}$, and $W^{-1}(\boldsymbol{y} - \boldsymbol{\lambda})$ has components $(y_i - \lambda_i)/\lambda_i$.

So each Newton step is a weighted least-squares fit of a straight line to the numbers z_i, with weights λ_i. Those are exactly the two things the naive fit got wrong. The *working response* z_i is the logarithm of the count, linearised about the current fit: the first-order Taylor expansion of $\ln y$ about λ is $\ln \lambda + (y - \lambda)/\lambda$, and it is finite for $y = 0$. The weight λ_i is the inverse of $\operatorname{Var}(\ln y_i) \approx \operatorname{Var} y_i / \lambda_i^2 = 1/\lambda_i$ by the delta method (Section 3.6.3): late bins with few expected counts count less. Here, because the curvature does not depend on the data, the Newton step, the Fisher-scoring step of Section 10.6.6 and the IRLS step are one and the same. The slope of the recipe at the fixed point is zero, (T5.50), so it converges quadratically.

Listing T5.5: Poisson regression by reweighted least squares (`code/chT5/21_irls.py`)

```
113 def poisson_irls(y, iters=50, keep=False):
114     z0 = np.log(y + 0.5)                  # starting point: straight line through log counts
115     beta = np.linalg.lstsq(X, z0, rcond=None)[0]
116     path = [beta.copy()]
117     for _ in range(iters):
118         eta = X @ beta
119         lam = np.exp(eta)
120         z = eta + (y - lam) / lam         # working response
121         W = lam                           # working weights
122         beta_new = np.linalg.solve(X.T @ (W[:, None] * X), X.T @ (W * z))
123         path.append(beta_new.copy())
124         if np.max(np.abs(beta_new - beta)) < 1e-14:
125             beta = beta_new
126             break
127         beta = beta_new
128     return (beta, np.array(path)) if keep else beta
```

Line 114 starts from a straight line through $\ln(y_i + \frac{1}{2})$, a crude but finite guess. Each pass computes the current linear predictor η_i and expected counts λ_i (lines 118–119), the working response and weights (lines 120–121), and solves the weighted normal equations (line 122). The script checks the result against a general-purpose minimiser of $-\ell$.

For one simulated curve with $\tau = 4$, 50 expected counts in the first bin and 4 empty bins, the largest change of the parameters falls as 0.00221, 2.268×10^{-6}, 2.225×10^{-12} over the passes: the exponent roughly doubles each time. The fit gives $\hat{\tau} = 3.922 \pm 0.2825$; the straight line through the logarithms of the non-empty bins gives 4.537. Over 5000 simulated curves (Figure T5.19) the Poisson fit averages 4.006 with a spread of 0.2948, while the line through $\ln y_i$ averages

4.485 with a spread of 0.3789; weighting that line by y_i helps only a little (4.393 ± 0.3375). The naive fits are biased towards long lifetimes for two reasons. Dropping empty bins keeps, in the tail, only the bins that fluctuated up, which flattens the tail. And the logarithm is concave, so $\mathbb{E}\ln y < \ln \mathbb{E}\, y$, by more in bins with fewer counts, which tilts the line. The bias, about one and a half standard errors, is made by the recipe, not by the data.

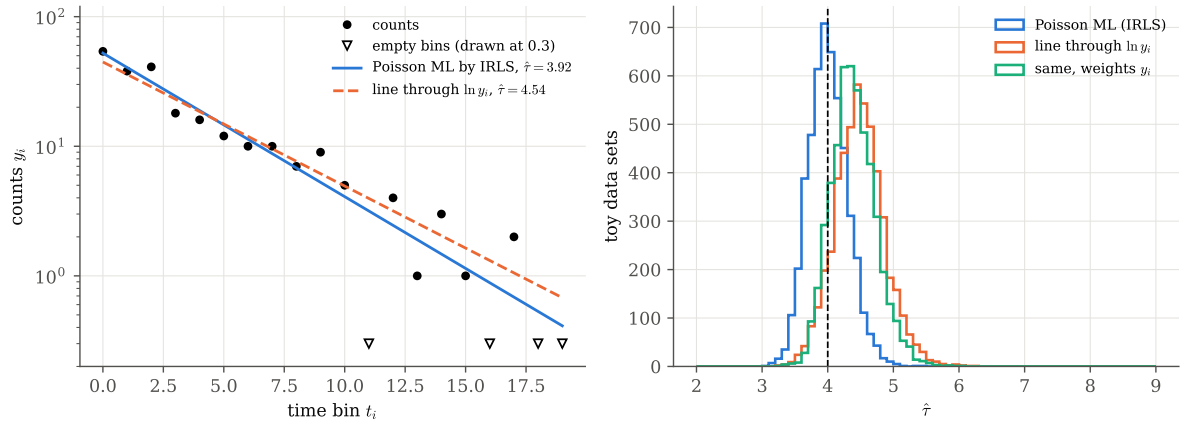


Figure T5.19: Left: one simulated decay curve (points; empty bins marked by open triangles) with the Poisson maximum-likelihood fit by reweighted least squares (solid) and a straight line through the logarithms of the non-empty bins (dashed). Right: the lifetimes from many simulated curves. The Poisson fit is centred on the true value (dashed line); fits through the logarithms are shifted to longer lifetimes and are wider.

This model, a count whose logarithm is linear in the parameters, is the simplest *generalised linear model*, and statistical software fits all such models (logistic regression for yes/no outcomes is another) by exactly this loop [138].

T5.7.4 The EM algorithm: when the labels are missing

The problem. A mass spectrum between 100 and 160 GeV holds $n = 1000$ events. A fraction f of them comes from a resonance at $\mu = 125$ GeV, smeared by the detector into a Gaussian of width $\sigma = 2$ GeV, as in a diphoton search; the rest is background, which we take as flat over the window for simplicity, with density $u = 1/60\,\text{GeV}^{-1}$. Each event's density is therefore the mixture

$$p(x \mid f) = f\,\varphi(x) + (1 - f)\,u, \qquad \varphi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}.$$

If every event carried a label saying "signal" or "background", the estimate of f would be a count: the number of signal labels divided by n. The labels are missing. The likelihood $\prod_i [f\varphi(x_i) + (1-f)u]$ is a product of sums, and its score equation $\sum_i (\varphi_i - u)/m_i = 0$, with $\varphi_i = \varphi(x_i)$ and $m_i = f\varphi_i + (1-f)u$, has no closed-form solution. When μ and σ are unknown too it is worse. The **expectation–maximisation (EM) algorithm** [139] replaces this one hard problem by a sequence of easy ones: guess the labels, count, and repeat [Wasserman pp. 143–146].

Guessing the labels: Bayes for one event. Take one event, at mass x_i, and a current guess f_k of the fraction. *The event.* An event was recorded at x_i. *What could have produced it?* Either the resonance or the background: these are the two hypotheses. *How probable was the event under each?* Under the signal hypothesis the density at x_i is φ_i, high near the peak and tiny far from it; under the background hypothesis it is u, the same everywhere. *Only then the priors:* before

looking at x_i, an event is signal with probability f_k. *Normalise:*

$$\tau_i \equiv \mathrm{P}(\text{signal} \mid x_i) = \frac{f_k\,\varphi_i}{f_k\,\varphi_i + (1-f_k)\,u}. \tag{T5.64}$$

This is the *E-step*: each event gets a soft label τ_i between zero and one. An event at the peak is probably signal; an event at 150 GeV is almost certainly background; an event on the shoulder of the peak is uncertain.

Counting the soft labels. With hard labels $z_i \in \{0,1\}$ the *complete-data* log-likelihood, the one we would have with labels, is a sum of easy pieces:

$$\ell_c(f,\mu,\sigma) \overset{(1)}{=} \sum_i \Big\{ z_i \big[\ln f + \ln \varphi(x_i;\mu,\sigma)\big] + (1-z_i)\big[\ln(1-f) + \ln u\big] \Big\}. \tag{T5.65}$$

(1) An event with label $z_i = 1$ contributes the probability f of being signal times the signal density; one with $z_i = 0$ contributes $(1-f)u$. The logarithm of the product over events is the sum.

We do not know the z_i, but we know their probabilities, so we average ℓ_c over them: since ℓ_c is linear in each z_i, the average replaces z_i by τ_i. The *M-step* maximises this average. Setting its derivatives to zero,

$$\begin{aligned}
\frac{\partial}{\partial f}: \ \frac{\sum_i \tau_i}{f} - \frac{\sum_i (1-\tau_i)}{1-f} = 0 &\overset{(1)}{\Longrightarrow} f_{k+1} = \frac{1}{n}\sum_i \tau_i, \\
\frac{\partial}{\partial \mu}: \ \sum_i \tau_i \frac{x_i-\mu}{\sigma^2} = 0 &\overset{(2)}{\Longrightarrow} \mu_{k+1} = \frac{\sum_i \tau_i x_i}{\sum_i \tau_i}, \\
\frac{\partial}{\partial \sigma}: \ \sum_i \tau_i \Big[-\frac{1}{\sigma} + \frac{(x_i-\mu)^2}{\sigma^3}\Big] = 0 &\overset{(3)}{\Longrightarrow} \sigma^2_{k+1} = \frac{\sum_i \tau_i (x_i-\mu_{k+1})^2}{\sum_i \tau_i}.
\end{aligned} \tag{T5.66}$$

(1) Multiply by $f(1-f)$: $(1-f)\sum \tau_i = f\sum(1-\tau_i)$, so $\sum \tau_i = fn$.
(2) $\ln\varphi = -(x-\mu)^2/2\sigma^2 - \ln\sigma - \frac{1}{2}\ln 2\pi$; differentiate in μ.
(3) Differentiate the same in σ and multiply by σ^3.

The new fraction is the average soft label: the expected number of signal events divided by n. The new peak position and width are the mean and variance of the masses, each event weighted by its probability of being signal. Then the soft labels are recomputed with the new parameters, and so on. The E-step is Bayes' theorem; the M-step is counting.[2]

Why the likelihood never goes down. Write the hidden label of event i as z and any probability distribution for it as $q_i(z)$. For one event,

$$\begin{aligned}
\ln p(x_i \mid \theta) &\overset{(1)}{=} \ln \sum_z p(x_i, z \mid \theta) \overset{(2)}{=} \ln \sum_z q_i(z)\,\frac{p(x_i, z \mid \theta)}{q_i(z)} \\
&\overset{(3)}{\geq} \sum_z q_i(z) \ln \frac{p(x_i, z \mid \theta)}{q_i(z)} \equiv F_i(q_i, \theta).
\end{aligned} \tag{T5.67}$$

(1) Marginalise over the label: $p(x \mid \theta) = \sum_z p(x, z \mid \theta)$, with z = signal or background.

[2] In [Wasserman p. 145, Ex. 9.50] the complete-data log-likelihood is printed without its squares; it should read $-\frac{1}{2}\sum_i (1-z_i)(y_i-\mu_0)^2 - \frac{1}{2}\sum_i z_i (y_i-\mu_1)^2$. On p. 144 the left-hand side $\mathcal{L}(\theta^{j+1})/\mathcal{L}(\theta^j)$ should be its logarithm.

(2) Multiply and divide each term by $q_i(z)$.

(3) Jensen's inequality for the concave logarithm: the logarithm of an average is at least the average of the logarithms, $\ln \sum_z q_z a_z \geq \sum_z q_z \ln a_z$ for weights $q_z \geq 0$ adding to one. For two terms it says the chord of ln lies below its curve. Equality holds when all the a_z are equal.

So for any guess of the labels, $F = \sum_i F_i$ is a lower bound on the log-likelihood. The bound touches when the ratio $p(x_i, z \mid \theta)/q_i(z)$ is the same for both labels, that is when $q_i(z)$ is proportional to $p(x_i, z \mid \theta)$: when q_i is the posterior probability of the label, which is (T5.64). Now follow one EM pass, starting from θ_k. The E-step sets q_i to the posterior at θ_k, so the bound touches there: $F(q, \theta_k) = \ln \mathcal{L}(\theta_k)$. In $F(q, \theta) = \sum_i \sum_z q_i(z) \ln p(x_i, z \mid \theta) - \sum_i \sum_z q_i \ln q_i$, only the first sum depends on θ, and it is the averaged complete-data log-likelihood, so the M-step maximises F over θ. Then

$$\ln \mathcal{L}(\theta_{k+1}) \overset{(1)}{\geq} F(q, \theta_{k+1}) \overset{(2)}{\geq} F(q, \theta_k) \overset{(3)}{=} \ln \mathcal{L}(\theta_k). \tag{T5.68}$$

(1) The bound (T5.67) holds at every θ.

(2) θ_{k+1} maximises $F(q, \cdot)$, so it does at least as well as θ_k.

(3) The E-step made the bound touch at θ_k.

Figure T5.20 draws it for four events: a curve that is easy to climb is laid under the likelihood, touching at the current point; we climb it to its top, and the likelihood there is at least as high. The proof is [Wasserman p. 144] and [CasellaBerger Thm. 7.2.20, p. 266] in our notation.

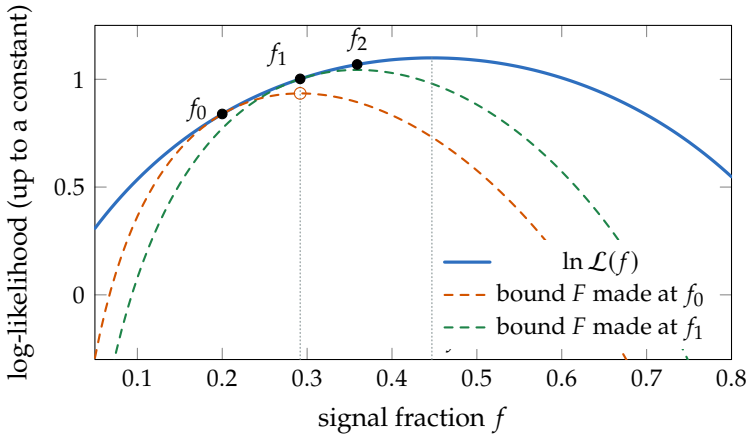


Figure T5.20: One EM pass for four events whose signal-to-background density ratios are 8, 3, 0.2 and 0.1. At $f_0 = 0.2$ the E-step builds a lower bound (orange, dashed) that touches the log-likelihood (blue). Its top, at $f_1 = 0.292$ (open circle), is the M-step. The likelihood at f_1 lies above the bound, so it went up. The bound built at f_1 (green) carries the iteration to $f_2 = 0.359$, and so on towards the maximum at $\hat{f} = 0.447$.

The spectrum. The whole E-step and M-step for the fraction fit into two lines of the script `22_em_mixture.py`.

Listing T5.6: EM for the signal fraction (`code/chT5/22_em_mixture.py`)

```
def em_fraction(phi, f=0.5, iters=2000, tol=1e-15):
    path, ll = [f], [np.sum(np.log(f * phi + (1 - f) * u))]
    for _ in range(iters):
        tau = f * phi / (f * phi + (1 - f) * u)          # E-step: soft labels
        f_new = tau.mean()                                # M-step: count them
```

Line 53 is (T5.64) for all events at once, and line 54 counts the soft labels. With the peak known and the fraction to be found, EM started at $f = 0.5$ goes $0.5 \to 0.2146 \to 0.153 \to 0.1302 \to \dots$ and reaches $\hat{f} = 0.1073 \pm 0.01508$, the root of the score equation, in 38 passes (to 10^{-10}). The simulated spectrum held 97 signal events. With the position and width unknown as well, started deliberately badly at $f = 0.5$, $\mu = 120$, $\sigma = 5$, the full EM of (T5.66) reaches $\hat{f} = 0.1043$, $\hat{\mu} = 125.3\,\text{GeV}$ and $\hat{\sigma} = 1.811\,\text{GeV}$ in 219 passes, the same point a general minimiser finds, and the log-likelihood rises at every pass (Figure T5.21).

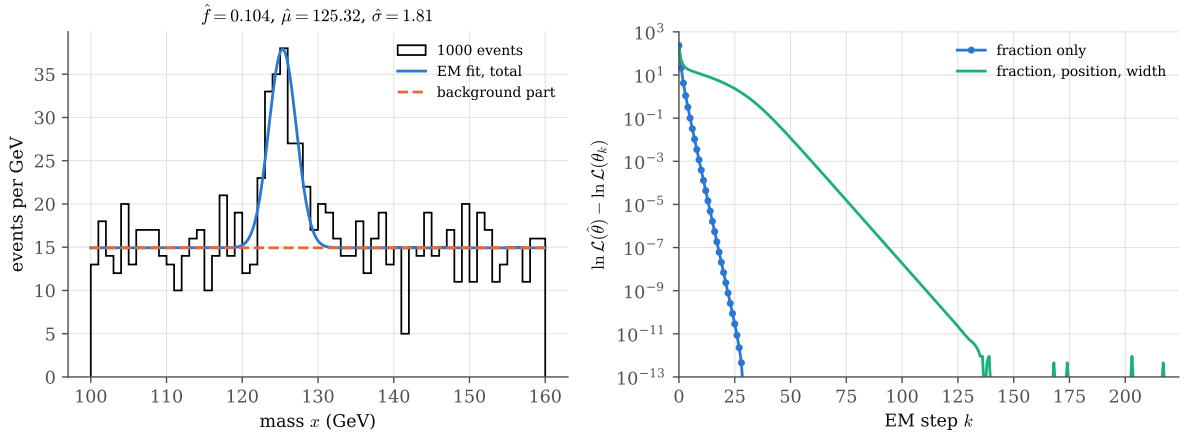


Figure T5.21: Left: the simulated mass spectrum with the EM fit of fraction, position and width (solid) and its background part (dashed). Right: how far the log-likelihood is below its maximum, pass by pass, for the fit of the fraction alone (points) and of all three parameters (line). It never increases, as (T5.68) requires; the first falls by a constant factor per pass, the second crawls.

How fast? The share of information that is missing. EM needed 38 passes where reweighting needed a handful. Its contraction factor can be computed. One pass is $g(f) = \frac{1}{n}\sum_i f\varphi_i/m_i$ with $m_i = f\varphi_i + (1-f)u$, so

$$
\begin{aligned}
g'(f) &\overset{(1)}{=} \frac{1}{n}\sum_i \frac{\varphi_i m_i - f\varphi_i(\varphi_i - u)}{m_i^2} \overset{(2)}{=} \frac{1}{n}\sum_i \frac{\varphi_i u}{m_i^2}, \\
\varphi_i u &\overset{(3)}{=} m_i\tilde{m}_i - f(1-f)(\varphi_i - u)^2, \qquad \tilde{m}_i \equiv (1-f)\varphi_i + fu = m_i + (1-2f)(\varphi_i - u), \\
g'(\hat{f}) &\overset{(4)}{=} 1 + \frac{1-2\hat{f}}{n}\sum_i \frac{\varphi_i - u}{m_i} - \frac{\hat{f}(1-\hat{f})}{n}\sum_i \frac{(\varphi_i - u)^2}{m_i^2} \overset{(5)}{=} 1 - \frac{I_{\text{obs}}}{I_{\text{com}}}.
\end{aligned}
\tag{T5.69}
$$

(1) Quotient rule: $\mathrm{d}m_i/\mathrm{d}f = \varphi_i - u$.
(2) $\varphi_i m_i - f\varphi_i(\varphi_i - u) = \varphi_i[f\varphi_i + (1-f)u - f\varphi_i + fu] = \varphi_i u$.
(3) Multiply out: $m_i\tilde{m}_i = f(1-f)(\varphi_i^2 + u^2) + [f^2 + (1-f)^2]\varphi_i u$, and $f^2 + (1-f)^2 = 1 - 2f(1-f)$; subtracting $f(1-f)(\varphi_i - u)^2$ leaves $\varphi_i u$. The second form of $\tilde{m}_i$ follows by collecting terms.
(4) Divide the identity by m_i^2 and average.
(5) At the estimate the score $\sum_i(\varphi_i - u)/m_i$ vanishes. The observed information is $I_{\text{obs}} = -\ell''(\hat{f}) = \sum_i(\varphi_i - u)^2/m_i^2$. The information with labels, from $\ln f$ for each of nf signal events and $\ln(1-f)$ for each background event, is $I_{\text{com}} = n/(f(1-f))$, the binomial result.

The contraction factor of EM is the fraction of the information that the missing labels take away [139]. Here it is 0.5791, and the ratio of successive errors in the run is 0.5791. The same number appears in the error bars: with labels the error on f would be $1/\sqrt{I_{\text{com}}} = 0.009787$, without them it is $1/\sqrt{I_{\text{obs}}} = 0.01508$, and $1 - (0.009787/0.01508)^2$ is again the contraction factor. EM is

slow exactly when the labels are hard to guess. Figure T5.22 shows it for peaks of different widths in the same window: a narrow peak separates signal from background well, and EM contracts by 0.2766 per pass, so six digits, which take $\ln 10^6/\ln(1/\text{factor})$ passes, need about a dozen; a peak twelve GeV wide overlaps the background everywhere, the factor is 0.958, and six digits take several hundred passes.

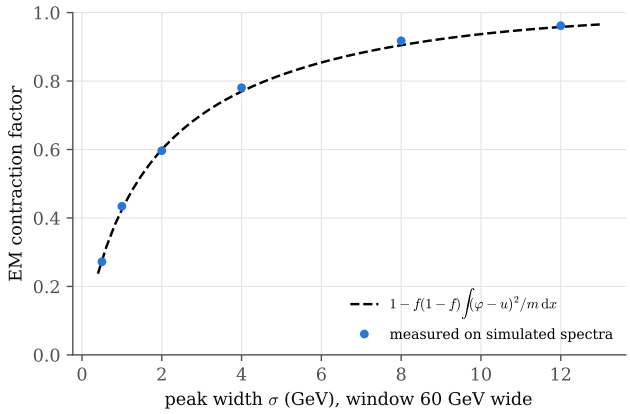


Figure T5.22: The contraction factor of EM for the signal fraction against the width of the peak, in a window 60 GeV wide with 10% signal. Dashed: the population value of $1 - I_{\text{obs}}/I_{\text{com}}$; points: the median over twenty simulated spectra of the measured ratio of successive errors. The more the peak overlaps the background, the less each event says about its own label, and the slower EM converges.

Watch out: EM climbs, but maybe the wrong hill

Equation (T5.68) guarantees that the likelihood never decreases, not that EM finds the highest maximum. It converges to a stationary point of the likelihood under mild conditions [140] [CasellaBerger §7.5.4, p. 299], which can be a local maximum. For mixtures there are several: a peak placed on a background fluctuation, or the two components swapped. Start from several points, as for any maximiser, and compare the likelihoods reached [Wasserman p. 143].

EM or Newton? Newton's method converges quadratically but needs second derivatives and can step downhill far from the maximum; EM needs only the complete-data maximisation, which is often a weighted average, and never steps downhill, but converges linearly at the rate of missing information. A common practice is to run EM first and finish with a few Newton steps. The iterative unfolding of Section 6.14.5 is EM for Poisson counts in true bins, with "which true bin did this event come from?" as the missing label: its fixed point, (6.78), is the maximum-likelihood condition, as (T5.68) leads us to expect.

T5.7.5 Stopping early on purpose

So far the iteration was a means to an end: run it until it stops changing. Sometimes the end point is the problem. When a detector's smearing is undone by unfolding, the maximum-likelihood solution amplifies the noise in the rough components of the spectrum, and Section 6.14.5 finds that a few EM passes, stopped early, give a smaller error than running to convergence. Why should the number of passes act like a regularisation strength?

The simplest iteration that shows it is gradient descent on a least-squares problem. We want $\boldsymbol{\theta}$ from $\boldsymbol{y} = A\boldsymbol{\theta} + \text{noise}$, and step downhill on $\frac{1}{2}|\boldsymbol{y} - A\boldsymbol{\theta}|^2$ with a small step ω, starting from

$\boldsymbol{\theta}_0 = \mathbf{0}$. In the eigenvectors of $A^{\mathsf{T}}A$, with eigenvalues s_j^2, each component evolves on its own:

$$\begin{aligned} \boldsymbol{\theta}_{k+1} &\overset{(1)}{=} \boldsymbol{\theta}_k + \omega A^{\mathsf{T}}(\boldsymbol{y} - A\boldsymbol{\theta}_k), \\ \boldsymbol{\theta}_{k+1} - \hat{\boldsymbol{\theta}} &\overset{(2)}{=} (\mathbb{1} - \omega A^{\mathsf{T}}A)(\boldsymbol{\theta}_k - \hat{\boldsymbol{\theta}}), \\ \theta_{k,j} &\overset{(3)}{=} \left[1 - (1 - \omega s_j^2)^k\right] \hat{\theta}_j . \end{aligned} \tag{T5.70}$$

(1) The gradient of $\frac{1}{2}|\boldsymbol{y} - A\boldsymbol{\theta}|^2$ is $-A^{\mathsf{T}}(\boldsymbol{y} - A\boldsymbol{\theta})$.
(2) Subtract the least-squares solution $\hat{\boldsymbol{\theta}}$, which obeys $A^{\mathsf{T}}\boldsymbol{y} = A^{\mathsf{T}}A\hat{\boldsymbol{\theta}}$ (Eq. (3.87)).
(3) Along eigenvector j the matrix multiplies by $1 - \omega s_j^2$ at each step; after k steps from $\boldsymbol{\theta}_0 = \mathbf{0}$ the error is $-(1 - \omega s_j^2)^k \hat{\theta}_j$.

A component with a large s_j^2, which the data measure well, converges in a few steps. A component with a tiny s_j^2, which the data barely see and where $\hat{\theta}_j$ is mostly noise, has moved only the fraction $1 - (1 - \omega s_j^2)^k \approx k\omega s_j^2$ of the way. Compare the Tikhonov (ridge) estimate, which minimises $|\boldsymbol{y} - A\boldsymbol{\theta}|^2 + \tau|\boldsymbol{\theta}|^2$, so $(A^{\mathsf{T}}A + \tau\mathbb{1})\boldsymbol{\theta} = A^{\mathsf{T}}\boldsymbol{y}$ and $\theta_{\tau,j} = s_j^2/(s_j^2 + \tau)\,\hat{\theta}_j$. For well-measured components both factors are near one; for poorly measured ones, s_j^2/τ against $k\omega s_j^2$. The two filters agree when

$$\tau \approx \frac{1}{\omega k} : \tag{T5.71}$$

stopping after k steps is regularisation with strength $1/(\omega k)$ (Figure T5.23). More passes, weaker regularisation; running to convergence is no regularisation at all. EM for unfolding behaves the same way, multiplicatively instead of additively, which is why its pass count has to be chosen with the same care as τ [141].

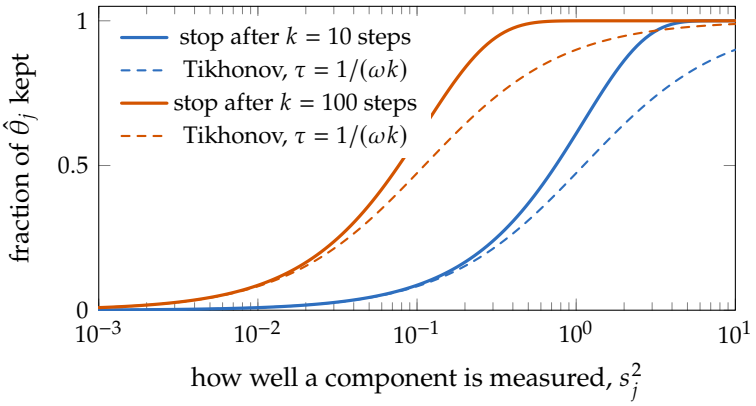


Figure T5.23: How much of each component of the least-squares solution survives, for gradient descent stopped after k steps (solid; step $\omega = 0.09$) and for Tikhonov regularisation with $\tau = 1/(\omega k)$ (dashed). Well-measured components (right) are kept by both; poorly measured ones (left), mostly noise, are suppressed by both, in nearly the same way. Ten times more steps moves the cut ten times further to the left.

What the iterative estimators have in common. Every iterative recipe met so far had the same three ingredients. A best estimator whose form needs the parameter itself: weights $1/(s + b_i)$, clipped weights, soft labels. A map g that improves a guess. And a fixed point that is the estimator we wanted, usually the maximum-likelihood one, reached at a speed set by $|g'|$ at the fixed point: tiny for reweighting when the model fits, zero for Newton, the missing information

for EM. The practical checks follow: watch the ratio of successive changes, start from more than one point when the likelihood may have several maxima, and decide whether you want the fixed point at all.

T5.8 Estimators that correct their own bias

What we already have. An estimator's bias is the gap between its average over repeated experiments and the truth, and its mean squared error is the variance plus the square of the bias (Section 3.2.1). Some estimators are unbiased by construction: the sample mean, the variance with $n-1$ in the denominator. For others, Section 3.6.3 found a small systematic offset from the curvature of a nonlinear function: $\mathbb{E}(1/x) \approx (1/\mu)(1+\sigma^2/\mu^2)$, above $1/\mu$. Maximum-likelihood estimates are consistent, so their bias disappears as the sample grows (Section 3.13.1), and for most purposes that is the end of the matter.

When that is not enough. A small bias matters in two situations. When each estimate rests on few data, a bias of order $1/n$ can be a sizeable fraction of the answer. And when many such estimates are combined, the scatter averages down while the bias does not, so after enough combining the bias is all that is left. In both cases we would like to remove the bias, ideally without a formula for it, since for a complicated estimator there is none.

A story with the whole trick in it. A GPS watch records a runner's position every ten seconds and adds up the straight lines between the points. On a winding trail the straight lines cut the corners, so the watch always reads short, and it reads shorter the fewer points it uses. We cannot see the true length, but we can see how the answer changes with the number of points: recompute the length from every other point and see how much shorter it comes out. If we know that the shortfall shrinks in a definite way as the points get denser, the two answers can be extrapolated to infinitely many points, and most of the shortfall is gone. Nothing outside the watch's own data was needed.

An estimator from n readings is in the same position. Its bias usually shrinks like $1/n$, so leaving one reading out and recomputing shows how the answer moves when the sample shrinks, and the two answers can be extrapolated to an infinite sample. Lifetimes are the physics version: the decay constant λ of a particle is estimated from n measured lifetimes t_i by $\hat{\lambda} = 1/\bar{t} = n/\sum_i t_i$, the maximum-likelihood estimate, and it is biased high.

The story in symbols: five lifetimes. Take $n = 5$ lifetimes, $t = (0.5,\ 1.0,\ 1.5,\ 2.0,\ 3.0)$ (in some unit of time), with sum $S = 8$. The estimate is $\hat{\lambda} = 5/8 = 0.625$. Leave out each lifetime in turn and recompute with the remaining four:

$$\hat{\lambda}_{(i)} = \frac{4}{S - t_i} = 0.5333,\ 0.5714,\ 0.6154,\ 0.6667,\ 0.8000, \qquad \bar{\lambda}_{(\cdot)} = \frac{1}{5}\sum_i \hat{\lambda}_{(i)} = 0.6374.$$

With four lifetimes the estimate is on average 0.0124 higher than with five. If the bias is a/n, going from five readings to four adds $a/4 - a/5 = a/20$; so $a/20 \approx 0.0124$, the bias of the five-reading estimate is $a/5 \approx 4 \times 0.0124 = 0.0495$, and the corrected value is $0.625 - 0.0495 = 0.5755$. That is the **jackknife** [142, 143]:

$$\hat{\lambda}_{\mathrm{JK}} = n\hat{\lambda} - (n-1)\,\bar{\lambda}_{(\cdot)} = 5(0.625) - 4(0.6374) = 0.5755. \tag{T5.72}$$

The dial is how far to extrapolate. Doing nothing leaves the full bias, which for this estimator is exactly $\lambda/(n-1)$, a quarter of λ at $n = 5$ (Section T5.8.1 derives it). The jackknife removes the

$1/n$ part and leaves an error of order $1/n^2$. The price is a little extra scatter, and Section T5.8.4 weighs it. Lesson: a bias that shrinks in a known way with the sample size can be measured by shrinking the sample, and extrapolated away.

A picture: extrapolating in $1/n$. Plot the estimate against $1/n$ (Figure T5.24). The full sample gives one point, at $1/5$; the average of the leave-one-out estimates gives another, at $1/4$. If the bias is proportional to $1/n$, the points lie on a straight line that reaches the truth at $1/n = 0$, an infinite sample. The line through $(1/n, \hat{\lambda})$ and $(1/(n-1), \bar{\lambda}_{(\cdot)})$ has slope $(\bar{\lambda}_{(\cdot)} - \hat{\lambda})\, n(n-1)$, so its value at zero is $\hat{\lambda} - (n-1)(\bar{\lambda}_{(\cdot)} - \hat{\lambda})$, which is (T5.72).

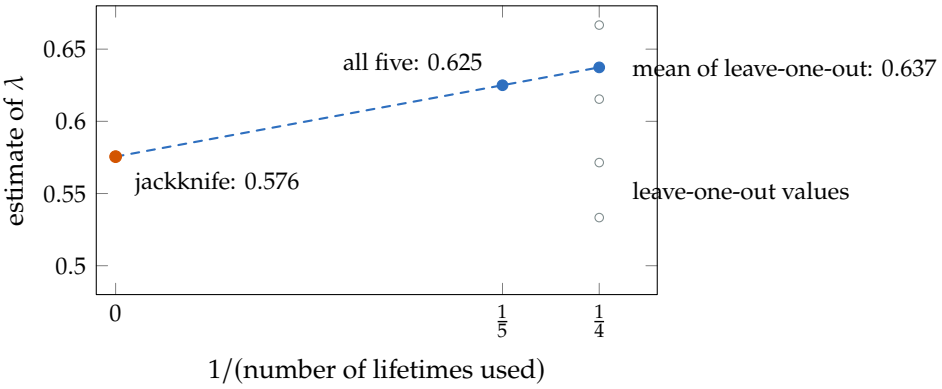


Figure T5.24: The jackknife as an extrapolation. The two blue points are the estimate from all five lifetimes at $1/n = 1/5$ and the mean of the five leave-one-out estimates at $1/n = 1/4$. The individual leave-one-out values are plotted in grey, with one value (0.80) above the frame. Extending the line through the blue points to $1/n = 0$ gives the extrapolation to infinite sample size (orange).

One idea, many names. Measuring an error by changing a resolution parameter and extrapolating to the limit is an old idea in physics and numerical analysis. Richardson extrapolation combines a numerical derivative or integral computed with step h and with step $h/2$ so that the leading error term cancels, and Romberg integration repeats this. Lattice field theory extrapolates results at several lattice spacings to the continuum limit, and several volumes to infinite volume. The jackknife is Richardson extrapolation with $1/n$ as the step. The bootstrap estimates the same bias differently, by simulating from the data [23], and repeating either correction gives the *iterated* bias corrections, which turn out to be the fixed-point iterations of Section T5.7 in disguise.

Four questions follow. Why is the bias so often of order $1/n$ (Section T5.8.1)? Does the jackknife really remove it, and what does it leave (Section T5.8.2)? How does the bootstrap do it, and what does iterating buy (Section T5.8.3)? What does correcting cost, and when is it worth it (Section T5.8.4)? The section ends where the idea is used in research, the iterative reconstruction of CMB lensing (Section T5.8.5).

T5.8.1 Why the bias is so often of order $1/n$

Many estimators are a smooth function of an average, $T = g(\bar{x})$ with $\bar{x} = \frac{1}{n}\sum_i x_i$, the x_i independent with mean μ and variance σ^2. Expand about μ and average:

$$\begin{aligned}\mathbb{E}\; g(\bar{x}) &\overset{(1)}{=} g(\mu) + g'(\mu)\;\mathbb{E}(\bar{x}-\mu) + \tfrac{1}{2}g''(\mu)\;\mathbb{E}(\bar{x}-\mu)^2 + \tfrac{1}{6}g'''(\mu)\;\mathbb{E}(\bar{x}-\mu)^3 + \dots \\ &\overset{(2)}{=} g(\mu) + \frac{g''(\mu)\,\sigma^2}{2n} + O\Big(\frac{1}{n^2}\Big).\end{aligned} \tag{T5.73}$$

(1) Taylor expansion of g about μ, averaged term by term.
(2) $\mathbb{E}(\bar{x}-\mu) = 0$; $\mathbb{E}(\bar{x}-\mu)^2 = \operatorname{Var}\bar{x} = \sigma^2/n$; the third central moment of a mean of n independent readings is μ_3/n^2 (the cube of $\sum_i(x_i-\mu)$ keeps only the n terms with one index, each μ_3, divided by n^3), and the fourth is $3\sigma^4/n^2 + O(n^{-3})$.

The bias is the curvature of g times the variance of the average, and that variance is σ^2/n. A straight-line g has no bias at all; a curved one always has a bias that starts at $1/n$ and continues in powers of $1/n$. Two examples.

The decay constant. For exponential lifetimes $\mu = 1/\lambda$ and $\sigma^2 = 1/\lambda^2$, and $g(\mu) = 1/\mu$ has $g'' = 2/\mu^3$, so the bias is $\frac{1}{2}(2\lambda^3)(1/\lambda^2)/n = \lambda/n$. Here the exact answer is available. The sum $S = \sum_i t_i$ of n exponential lifetimes has the Gamma density $\lambda^n s^{n-1}e^{-\lambda s}/(n-1)!$, so

$$\mathbb{E}\,\frac{1}{S} \overset{(1)}{=} \int_0^\infty \frac{1}{s}\,\frac{\lambda^n s^{n-1}e^{-\lambda s}}{(n-1)!}\,\mathrm{d}s \overset{(2)}{=} \frac{\lambda^n}{(n-1)!}\,\frac{(n-2)!}{\lambda^{n-1}} = \frac{\lambda}{n-1}, \qquad \mathbb{E}\,\hat{\lambda} = \mathbb{E}\,\frac{n}{S} = \lambda\,\frac{n}{n-1}. \tag{T5.74}$$

(1) The density of a sum of n independent exponentials (the Gamma distribution with shape n, Section 2.15.2).
(2) $\int_0^\infty s^{n-2}e^{-\lambda s}\mathrm{d}s = (n-2)!/\lambda^{n-1}$, the Gamma integral.

So the bias is $\lambda/(n-1) = \lambda/n + \lambda/n^2 + \ldots$, whose first term is the λ/n of (T5.73). This is Cowan's (6.9) [Cowan §6.2, p. 73]: $\hat{\tau} = \bar{t}$ is unbiased for the lifetime, and $1/\hat{\tau}$ is not unbiased for $1/\tau$.

The standard deviation. The sample variance s^2 with $n-1$ is unbiased, but its square root is not: $g(v) = \sqrt{v}$ is curved. Apply (T5.73) with s^2 in the role of the average, mean σ^2 and, for Gaussian readings, variance $2\sigma^4/(n-1)$. With $g''(v) = -\frac{1}{4}v^{-3/2}$, the bias of s is $\frac{1}{2}(-\frac{1}{4}\sigma^{-3})\,2\sigma^4/(n-1) = -\sigma/(4(n-1))$. The exact factor is $\mathbb{E}\,s = c_4\sigma$ with $c_4 = \sqrt{2/(n-1)}\,\Gamma(n/2)/\Gamma((n-1)/2)$; for $n = 5$, $c_4 = 0.94$ against $1 - 1/16 = 0.9375$ from the expansion. Unbiasedness does not survive a nonlinear change of variable.

T5.8.2 The jackknife removes the $1/n$ term

Let T_n be any estimator computed from n independent readings, with $\mathbb{E}\,T_n = \theta + a/n + b/n^2 + \ldots$, where a and b may depend on θ but not on n. Leave out reading i, recompute to get $T_{(i)}$, a statistic of $n-1$ readings, and average the n values: $\bar{T}_{(\cdot)} = \frac{1}{n}\sum_i T_{(i)}$. Casella and Berger define the jackknife estimator as $\mathrm{JK}(T_n) = nT_n - \frac{n-1}{n}\sum_i T_{(i)} = nT_n - (n-1)\bar{T}_{(\cdot)}$ [CasellaBerger Ex. 7.66, p. 297]. Its average is

$$\begin{aligned}\mathbb{E}\,\mathrm{JK}(T_n) &\overset{(1)}{=} n\Big(\theta + \frac{a}{n} + \frac{b}{n^2}\Big) - (n-1)\Big(\theta + \frac{a}{n-1} + \frac{b}{(n-1)^2}\Big) + \ldots \\ &\overset{(2)}{=} \theta + \frac{b}{n} - \frac{b}{n-1} + \ldots \overset{(3)}{=} \theta - \frac{b}{n(n-1)} + \ldots\end{aligned} \tag{T5.75}$$

(1) Each $T_{(i)}$ is the same estimator applied to $n-1$ readings, so it has the same expansion with $n-1$ in place of n, and so does their average.
(2) The θ's give $n\theta - (n-1)\theta = \theta$; the a's give $a - a = 0$.
(3) $1/n - 1/(n-1) = -1/(n(n-1))$.

Key idea: The jackknife removes the $1/n$ bias

If $\mathbb{E}\,T_n = \theta + a/n + O(n^{-2})$, then $nT_n - (n-1)\bar{T}_{(\cdot)}$ has bias $O(n^{-2})$, whatever a is. It needs no formula for the bias, only the ability to recompute the estimator with one reading left out. If the bias is exactly a/n, the jackknife is exactly unbiased.

Example: the square of a success probability. Casella and Berger's exercise makes the last sentence concrete. Let $x_1, \dots, x_n$ be Bernoulli trials with success probability θ, and estimate θ^2 by $T_n = \bar{x}^2$. With $S = \sum_i x_i$ binomial, $\mathbb{E}\,T_n = (\operatorname{Var} S + (\mathbb{E}\,S)^2)/n^2 = \theta^2 + \theta(1-\theta)/n$: a bias of exactly a/n. The leave-one-out estimate is $T_{(i)} = (S - x_i)^2/(n-1)^2$; S of them have $x_i = 1$ and the other $n - S$ have $x_i = 0$. So

$$\mathrm{JK} \overset{(1)}{=} \frac{S^2}{n} - \frac{S(S-1)^2 + (n-S)S^2}{n(n-1)} \overset{(2)}{=} \frac{(n-1)S^2 - (n-2)S^2 - S}{n(n-1)} = \frac{S(S-1)}{n(n-1)},$$
$$\mathbb{E}\,\mathrm{JK} \overset{(3)}{=} \frac{\operatorname{Var} S + (\mathbb{E}\,S)^2 - \mathbb{E}\,S}{n(n-1)} = \frac{n\theta(1-\theta) + n^2\theta^2 - n\theta}{n(n-1)} = \theta^2. \tag{T5.76}$$

(1) $nT_n = S^2/n$ and $(n-1)\bar{T}_{(\cdot)} = \frac{n-1}{n}\sum_i (S - x_i)^2/(n-1)^2$, with the sum split into the S successes and the $n - S$ failures.
(2) Expand: $S(S-1)^2 + (n-S)S^2 = nS^2 - 2S^2 + S$; put everything over $n(n-1)$.
(3) $\mathbb{E}[S(S-1)] = \mathbb{E}\,S^2 - \mathbb{E}\,S$, with $\mathbb{E}\,S = n\theta$ and $\operatorname{Var} S = n\theta(1-\theta)$.

The jackknife of the maximum-likelihood estimate is exactly unbiased: it is the count of ordered pairs of successes divided by the number of ordered pairs of trials, the natural unbiased estimate of θ^2.

Example: the decay constant. For $T_n = \hat{\lambda} = n/S$ the bias $\lambda/(n-1)$ is not exactly a/n, so a remainder is left. It can be computed exactly. Each $\hat{\lambda}_{(i)} = (n-1)/(S - t_i)$ is built from $n-1$ exponentials, so by (T5.74) with $n-1$ in place of n its average is $\lambda(n-1)/(n-2)$, and

$$\mathbb{E}\,\hat{\lambda}_{\mathrm{JK}} \overset{(1)}{=} n\,\frac{\lambda n}{n-1} - (n-1)\,\frac{\lambda(n-1)}{n-2} \overset{(2)}{=} \lambda\,\frac{n^2(n-2) - (n-1)^3}{(n-1)(n-2)} \overset{(3)}{=} \lambda - \frac{\lambda}{(n-1)(n-2)}. \tag{T5.77}$$

(1) Average of $n\hat{\lambda}$ minus $n-1$ times the average leave-one-out value.
(2) Common denominator.
(3) $n^3 - 2n^2 - (n^3 - 3n^2 + 3n - 1) = n^2 - 3n + 1 = (n-1)(n-2) - 1$.

A bias of $+\lambda/(n-1)$ has become $-\lambda/((n-1)(n-2))$: at $n = 5$ from $+0.25\lambda$ to -0.083λ, at $n = 10$ from $+0.111\lambda$ to -0.014λ. The simulation of 40000 experiments per sample size in `23_bias_correction.py` finds -0.08422 and -0.01532 (Figure T5.25). The same leave-one-out values also give an error bar, the jackknife variance of Section 13.18, which a later chapter uses for correlation functions measured from one sky.

T5.8.3 The bootstrap estimate of bias, and iterating it

The bootstrap of Section 3.15 replaced the unknown distribution F of the data by the empirical distribution $\hat{F}$, which puts probability $1/n$ on each observed value, and simulated from it. The same move estimates a bias. In the real world the bias is $\mathbb{E}_F\,T - \theta$. In the bootstrap world the data come from $\hat{F}$, and the quantity being estimated is the same function of $\hat{F}$ that θ is of F: the *plug-in* value, which for $\hat{\lambda} = 1/\bar{t}$ is $1/\bar{t}$ itself, the estimate T. So the bootstrap world's bias is $\mathbb{E}^*\,T^* - T$, where T^* is the estimator recomputed on a resample and $\mathbb{E}^*$ the average over resamples [23], and the corrected estimator is

$$T_1 = T - \left(\mathbb{E}^*\,T^* - T\right) = 2T - \mathbb{E}^*\,T^*. \tag{T5.78}$$

For a smooth function of a mean, the bootstrap world has its own version of (T5.73), $\mathbb{E}^*\,g(\bar{x}^*) - g(\bar{x}) \approx g''(\bar{x})\hat{\sigma}^2/2n$, the true bias with μ, σ replaced by their estimates. The error of that replacement is of order $1/\sqrt{n}$ times a term of order $1/n$, and averages to order $1/n^2$: like the jackknife, one bootstrap correction leaves a bias of order $1/n^2$.

A parametric version that can be done by hand. If we trust the exponential model, we can resample from the fitted exponential instead of from the data. Then each resampled estimate has, by (T5.74), the average $\mathbb{E}^* T^* = cT$ with $c = n/(n-1)$, and (T5.78) gives $T_1 = (2-c)T$, whose average is $(2-c)c\lambda$. Writing $c = 1+\varepsilon$ with $\varepsilon = 1/(n-1)$, $(1-\varepsilon)(1+\varepsilon) = 1-\varepsilon^2$: the bias has dropped from $\varepsilon\lambda = \lambda/(n-1)$ to $-\varepsilon^2\lambda = -\lambda/(n-1)^2$.

Iterating. T_1 is itself an estimator with a bias, which the bootstrap can estimate in the same way: recompute T_1 on resamples, $T_1^* = 2T^* - \mathbb{E}^{**} T^{**}$, where the double star marks resamples drawn from a resample. One point needs care: in the bootstrap world the truth is still the plug-in value T, not T_1. So the bias of T_1 is estimated by $\mathbb{E}^* T_1^* - T$, and

$$T_2 = T_1 - \big(\mathbb{E}^* T_1^* - T\big) \overset{(1)}{=} 2T - \mathbb{E}^* T^* - 2\,\mathbb{E}^* T^* + \mathbb{E}^* \mathbb{E}^{**} T^{**} + T = 3T - 3\,\mathbb{E}^* T^* + \mathbb{E}^* \mathbb{E}^{**} T^{**}. \quad \text{(T5.79)}$$

(1) Insert $T_1 = 2T - \mathbb{E}^* T^*$ and $\mathbb{E}^* T_1^* = 2\,\mathbb{E}^* T^* - \mathbb{E}^* \mathbb{E}^{**} T^{**}$.

This is the **iterated** (double) bootstrap [144]. In the parametric version every level multiplies by c, so $T_2 = (3 - 3c + c^2)T$ and $\mathbb{E}\,T_2 = (1 - \varepsilon + \varepsilon^2)(1+\varepsilon)\lambda = (1+\varepsilon^3)\lambda$: each level gains one more power of $1/(n-1)$.

The fixed point behind the ladder. What is the iteration converging to? We are looking for the parameter value θ_c that would, on average, produce the estimate we observed: $\theta_c + b(\theta_c) = T$, where $b(\theta)$ is the bias when the truth is θ. Written as $\theta_c = T - b(\theta_c)$, this is a fixed-point equation, and the natural iteration is

$$\theta^{(0)} = T, \qquad \theta^{(j+1)} = T - b\big(\theta^{(j)}\big). \quad \text{(T5.80)}$$

For the decay constant $b(\theta) = \varepsilon\theta$, so $\theta^{(1)} = (1-\varepsilon)T$, the one-level parametric bootstrap; $\theta^{(2)} = T - \varepsilon(1-\varepsilon)T = (1 - \varepsilon + \varepsilon^2)T$, the two-level one. The slope of the map is $-\varepsilon$, so by Section T5.7.1 it converges, alternating in sign, to $\theta_c = T/(1+\varepsilon) = T\,(n-1)/n$. That is the exactly unbiased estimator, since $\mathbb{E}[T(n-1)/n] = \lambda$ by (T5.74). Each level of bootstrap is one pass of a fixed-point iteration on the bias equation, and the iteration helps when the bias is a smooth, well-estimated function of the parameter with $|b'| < 1$. The same equation, with many parameters at once, fixes the power spectrum of a reconstructed lensing map: its bias $N^{(1)}$ is linear in the very spectrum being measured, so the debiased spectrum solves $\boldsymbol{c} = \boldsymbol{d} - \mathbf{K}\boldsymbol{c}$, and Section T1.9.5 solves it both directly and by this iteration.

What the simulation shows. Figure T5.25, left, compares the corrections over 40000 simulated experiments per sample size, the bootstrap with 100 resamples and, for the double bootstrap, 25 second-level resamples of each. The plain estimate follows the exact $\lambda/(n-1)$; the jackknife follows (T5.77). The nonparametric bootstrap does better than expected: at $n = 10$ its bias is $-7.750 \times 10^{-5} \pm 0.001821$, consistent with zero. Iterating it does not help here. At $n = 10$ the double bootstrap is off by 0.009755 ± 0.00189, no better than one level. At $n = 5$ it misbehaves. A resample of a resample of five lifetimes can consist of the shortest lifetime five times over, which gives a huge $1/\bar{t}^{**}$, so a few experiments produce enormous corrections and the average is dominated by them. Its median is off by only -0.06696 and its interquartile range, 0.7191, is like the plain estimator's 0.6859, but its mean is unstable. This is the weakness of Section 3.15.4 again: resampling copes badly with statistics that a single extreme reading can control. The theorem that each level gains a power of $1/n$ is about large n. The parametric ladder, which uses the model instead of resampling, does gain a power at each level.

T5.8.4 What a correction costs, and when it is worth it

Removing a bias changes the estimator, and therefore its variance. By (3.13) the mean squared error is the variance plus the squared bias, and a correction of order $1/n$ changes the variance

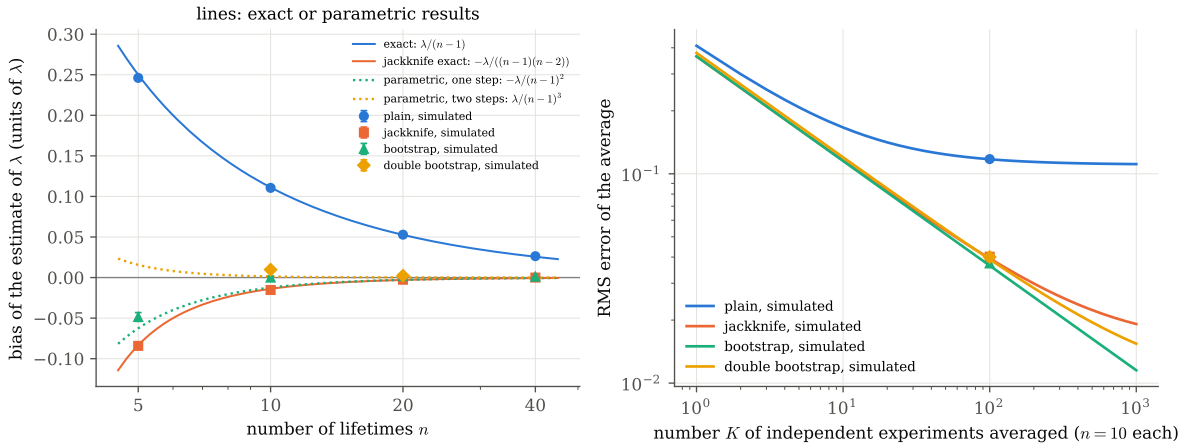


Figure T5.25: Left: the bias of the decay constant estimated from n lifetimes, in units of λ, for the plain estimate and three corrections (points, simulated; error bars are Monte Carlo errors), against the exact plain and jackknife results and the one- and two-level parametric bootstrap (lines). The double bootstrap at $n = 5$ is off the scale and omitted. Right: the error of the average of K independent estimates, each from $n = 10$ lifetimes. The scatter averages down as $1/\sqrt{K}$; the bias of the plain estimate does not, and beyond $K \approx 12$ it is all that is left. Points: direct simulation at $K = 100$.

at order $1/n^2$, the same order as the squared bias it removes. Whether the error goes down or up depends on the case.

When the correction shrinks the estimate, it wins twice. The exactly unbiased $T(n-1)/n$ for the decay constant has variance $((n-1)/n)^2$ times that of T: smaller. The jackknife behaves similarly: at $n = 10$ its mean squared error is $0.1322\lambda^2$ against $0.1672\lambda^2$ for the plain estimate.

When it stretches the estimate, it can lose. The sample standard deviation is biased low, $\mathbb{E}\,s = c_4\sigma$ with $c_4 < 1$, and the unbiased version s/c_4 stretches it. Since $\mathbb{E}\,s^2 = \sigma^2$, $\mathrm{Var}\,s = \sigma^2(1 - c_4^2)$, and

$$\mathrm{MSE}(s) \overset{(1)}{=} \sigma^2(1-c_4^2) + \sigma^2(1-c_4)^2 = 2\sigma^2(1-c_4), \qquad \mathrm{MSE}(s/c_4) \overset{(2)}{=} \frac{\mathrm{Var}\,s}{c_4^2} = \sigma^2\Big(\frac{1}{c_4^2} - 1\Big). \quad \text{(T5.81)}$$

(1) Variance plus squared bias, and $(1 - c_4^2) + (1 - c_4)^2 = 2 - 2c_4$.
(2) s/c_4 is unbiased, so its error is all variance.

For $n = 5$, $c_4 = 0.94$, and the two are $0.120\sigma^2$ and $0.132\sigma^2$: the unbiased estimator is worse. The simulation with 200000 samples agrees (0.1201 and 0.1317) and adds the jackknife of s: it removes most of the bias (-0.0608σ becomes 0.01384σ) but raises the variance from $0.1164\sigma^2$ to $0.1391\sigma^2$, and its mean squared error, $0.1393\sigma^2$, is the largest of the three. For one experiment with a small sample, removing a bias can make the answer worse.

Watch out: Unbiased is not the same as better

A bias correction trades a systematic error for extra scatter. For a single measurement judged by its mean squared error, the trade can go either way, and (T5.81) shows a case where it loses. Correct a bias when the estimate will be combined with many others, or when a systematic offset would be misread as physics; do not correct it by reflex.

When the bias matters: combining many estimates. Average K independent estimates, each from n lifetimes. The variance falls as $1/K$, the bias stays:

$$\mathrm{MSE}(K) \overset{(1)}{=} \frac{\mathrm{Var}\,\hat\lambda}{K} + b^2, \qquad K_{\mathrm{cross}} \overset{(2)}{=} \frac{\mathrm{Var}\,\hat\lambda}{b^2} \overset{(3)}{=} \frac{n^2}{n-2}. \quad \text{(T5.82)}$$

(1) The average of independent estimates has the same bias and $1/K$ of the variance.
(2) The two terms are equal at K_{cross}.
(3) $\operatorname{Var}\hat{\lambda} = n^2\,\mathbb{E}(1/S^2) - (\mathbb{E}\,\hat{\lambda})^2$ with $\mathbb{E}(1/S^2) = \lambda^2/((n-1)(n-2))$ (the Gamma integral of (T5.74) with s^{-2}), so $\operatorname{Var}\hat{\lambda} = \lambda^2 n^2/((n-1)^2(n-2))$; and $b = \lambda/(n-1)$.

For $n = 10$, $K_{\text{cross}} = 12.5$, and the simulation gives 12.68. Average more than about a dozen ten-event experiments and the bias of the plain estimate dominates. At $K = 100$ its error is 0.1176λ, almost all bias, against 0.04017λ after the jackknife and 0.03731λ after the bootstrap (Figure T5.25, right). The rule of thumb is that a $1/n$ bias matters once we average more than about n estimates, or once the bias is a sizeable fraction of the final error bar. That is the normal situation for a power spectrum, built from hundreds of multipoles that each have only a few modes, and for any stacked measurement made of many small samples.

T5.8.5 In research: iterating a lensing reconstruction

The reconstruction of the CMB lensing potential in Section T1.9 combines two problems met above: an estimator whose best weights depend on what it estimates, and a bias that has to be removed. There we built the quadratic estimator $\hat{\psi}$ from pairs of temperature modes and found that the spectrum of the reconstruction is the signal plus two biases, $C_L^{\psi\psi} + N_L^{(0)} + N_L^{(1)}$ (T1.27). The noise term $N^{(0)}$ equals the normalisation of the best weights, $N_L^{(0)} = \left[\int f^2/(2C_{\ell_1}^{\text{tot}}C_{\ell_2}^{\text{tot}})\right]^{-1}$ (T1.24): the smaller the total power C^{tot} in the modes being paired, the smaller the noise. For temperature we found that $N^{(0)}$ comes mostly from the chance pairing of the unlensed CMB with itself, which no amount of cleverness can remove.

Their question. Hirata and Seljak asked whether the quadratic estimator is the best one can do, by writing down the full likelihood of the lensed sky as a function of ψ [145, 146]. For temperature at the noise of Planck the answer was essentially yes. For polarisation it was no, and the reason is a fixed point.

The mechanism. The most sensitive polarisation estimator pairs E modes with B modes. Primordial B modes are faint or absent, so on a low-noise map the observed B is mostly lensing: E modes deflected by the very ψ we want. The EB version of the noise formula has $C^{EE,\text{tot}}C^{BB,\text{tot}}$ in the denominator [57], and $C^{BB,\text{tot}}$ is lensing power plus instrument noise. So the reconstruction noise is set by the lensing signal. Now iterate. Reconstruct ψ; use it to undo the lensing of the observed maps (*delensing*), which removes most of the lensing B modes; the residual B power is smaller, so the best weights change and $N^{(0)}$ drops; reconstruct again with the cleaner maps; repeat (Figure T5.26). The fixed point is the maximum-likelihood ψ of the full likelihood, and in the language of Section T5.7 it is a reweighting recipe whose weights depend on the answer through the residual C^{BB}. Hirata and Seljak found that the mean squared error of the reconstruction drops by as much as a factor of two at a noise level of $1.4\,\mu$K-arcmin and by as much as ten at $0.25\,\mu$K-arcmin, depending on the angular scale [146]. The temperature case differs exactly where Section T1.9 found: there $N^{(0)}$ is unlensed CMB, which delensing does not touch.

In practice. Modern analyses implement the iteration as a maximum-a-posteriori reconstruction, solving for ψ and the delensed CMB together with the lensing remapping done exactly [147], and they are needed for the next generation of low-noise polarisation experiments. Every bias of Section T1.9 then has an iterated counterpart: $N^{(0)}$ and $N^{(1)}$ must be recomputed for the iterated estimator, by simulations that go through the same iterations as the data. The lesson about bias corrections applies in full: the correction is estimated from simulations that mimic the estimator, and the iteration is stopped once the change in $\hat{\psi}$ is small compared with its noise.

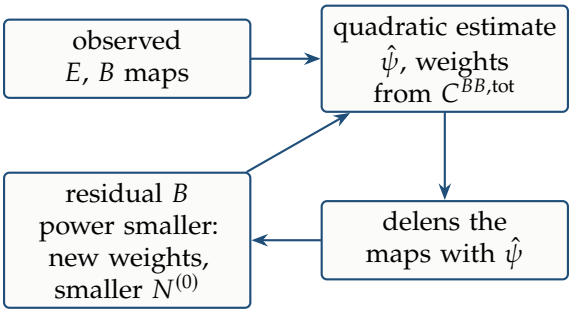


Figure T5.26: The iteration of a polarisation lensing reconstruction. The noise of the estimate depends on the B-mode power, most of which is produced by the lensing being estimated; removing it with the current estimate lowers the noise of the next one. The loop settles at the maximum-likelihood reconstruction.

T5.9 Estimators that update with every reading

What we already have. Every estimator so far took the whole data set at once. The mean adds all the readings and divides; least squares builds $A^{\mathsf{T}}\mathbf{C}^{-1}A$ from all the rows of the design matrix and inverts it (Eq. (3.87)). If one more reading arrives, we start again.

When that is not enough. Some data never stop arriving. A detector fires calibration pulses every second for months; a telescope's pointing is checked against stars all night; an online trigger must decide about each event before the next one comes. Keeping every reading and recomputing from scratch costs memory and time that grow without limit. Worse, the quantity may not stand still. A photomultiplier's gain drifts with temperature; a mean over all the pulses since the start of the run then describes the detector as it was on average, not as it is now.

A story with the whole trick in it. A ship's navigator before satellites kept one number for the ship's position, with an idea of how uncertain it was. Between star sightings she advanced the position by the ship's speed and heading, and because currents and steering errors push the ship unpredictably, her uncertainty grew with every hour. When a star sighting came, she did not throw away her estimate and use the sighting alone, nor ignore the sighting: she moved her position towards it, a long way if her own estimate had grown uncertain and the sighting was good, a little if the sighting was poor. Then the uncertainty shrank, and the cycle began again. She never needed the log of all past sightings, only the current position and its uncertainty.

That is what a recursive estimator does with a stream of calibration readings: keep the current estimate and its variance; let the variance grow by the amount the quantity can drift between readings; when a reading arrives, move the estimate towards it by a fraction set by the two variances. When nothing drifts, this reproduces the ordinary mean exactly, reading by reading.

The story in symbols: a running mean. Readings 10, 12, 11, 13 arrive one at a time. Writing m_N for the mean of the first N, each new reading moves the mean by a fraction $1/N$ of the *surprise*, the difference between the reading and the current mean:

$$m_N \overset{(1)}{=} \frac{(N-1)\,m_{N-1} + x_N}{N} \overset{(2)}{=} m_{N-1} + \frac{1}{N}\bigl(x_N - m_{N-1}\bigr). \tag{T5.83}$$

(1) The first $N-1$ readings add up to $(N-1)m_{N-1}$; add the new one and divide by N.
(2) Split $(N-1)m_{N-1}/N = m_{N-1} - m_{N-1}/N$.

So $m_1 = 10$, $m_2 = 10 + \frac{1}{2}(12 - 10) = 11$, $m_3 = 11 + \frac{1}{3}(11 - 11) = 11$, $m_4 = 11 + \frac{1}{4}(13 - 11) = 11.5$, which is $46/4$, the batch mean. The dial is the fraction $1/N$, the **gain**: how far the estimate moves towards a new reading. Here it shrinks as readings accumulate, which is right when the quantity is constant, because the estimate is already backed by $N - 1$ readings. It is wrong when the quantity drifts: after a thousand readings a new one moves the estimate by a thousandth of the surprise, and the estimate can no longer follow. On the drifting detector gain simulated in Section T5.9.3, the running mean of all readings is off by 0.06061 on average, six times worse than the best recursive estimator, whose gain stays near 0.03921 for ever. Lesson: an estimate can be updated with one number per reading, and the whole art is choosing how far each reading moves it.

A picture: predict and update. Figure T5.27 shows one cycle when the belief about the quantity is a Gaussian. Between readings, the quantity may drift: the belief keeps its centre and widens (left). A reading is a second, independent Gaussian piece of information; multiplying the two (Bayes' theorem with the widened belief as the prior) gives a narrower Gaussian whose centre lies between them, closer to whichever is narrower (right).

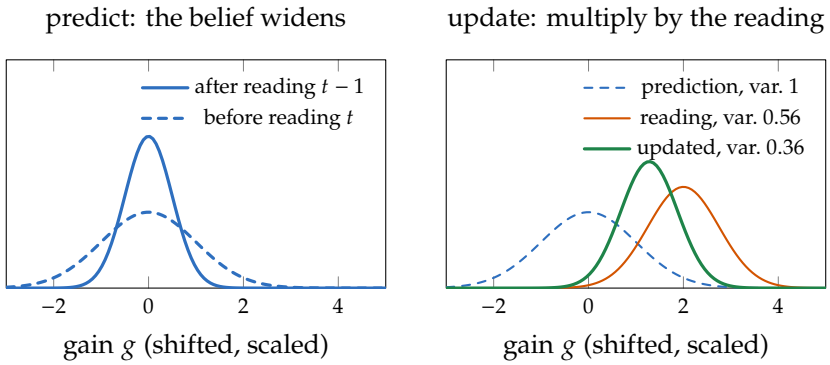


Figure T5.27: One cycle of the Kalman filter for a Gaussian belief. Left: between readings the quantity may drift, so the variance grows by the drift variance (here from 0.25 to 1) while the centre stays. Right: the prediction (variance 1, centre 0) and a reading at 2 with variance 0.5625 combine into a Gaussian with variance $1/(1 + 1/0.5625) = 0.36$ and centre $0 + 0.64 \times 2 = 1.28$: the gain $K = 1/(1 + 0.5625) = 0.64$ moves the estimate most of the way towards the better of the two.

One idea, many names. Updating an estimate by a fraction of the surprise is everywhere. The running mean and its exponentially weighted cousin, which keeps the gain fixed, are used in every control room. Recursive least squares updates a fit point by point. The **Kalman filter** [148], which navigated the Apollo spacecraft, is the general Gaussian version with a drifting state. In Bayesian language all of them are one rule: today's posterior is tomorrow's prior, the sequential use of Bayes' theorem that Chapter 5 makes central. The mathematics is the algebra of Gaussians, because a Gaussian belief multiplied by a Gaussian likelihood is again Gaussian, so the whole belief is carried by two numbers.

We first make the running mean and variance exact and safe in floating-point arithmetic (Section T5.9.1), then extend the update to least-squares fits (Section T5.9.2), and finally let the quantity drift (Section T5.9.3).

T5.9.1 The running mean and variance

(T5.83) gives the mean exactly. For the variance we need the running sum of squared deviations, $M_N = \sum_{i \le N}(x_i - m_N)^2$, so that $s_N^2 = M_N/(N - 1)$. When a reading arrives, the mean moves, so

every old deviation changes. How does the sum change? Write $\delta = x_N - m_{N-1}$ for the surprise; by (T5.83), $m_N - m_{N-1} = \delta/N$ and $x_N - m_N = \delta(N-1)/N$. Then

$$\begin{aligned} M_N &\overset{(1)}{=} \sum_{i<N} (x_i - m_{N-1})^2 + (N-1)(m_{N-1} - m_N)^2 + (x_N - m_N)^2 \\ &\overset{(2)}{=} M_{N-1} + \frac{(N-1)\,\delta^2}{N^2} + \frac{(N-1)^2\,\delta^2}{N^2} \overset{(3)}{=} M_{N-1} + \frac{N-1}{N}\,\delta^2 \\ &\overset{(4)}{=} M_{N-1} + (x_N - m_{N-1})(x_N - m_N). \end{aligned} \tag{T5.84}$$

(1) For the old readings write $x_i - m_N = (x_i - m_{N-1}) + (m_{N-1} - m_N)$ and square; the cross term is $2(m_{N-1} - m_N)\sum_{i<N}(x_i - m_{N-1}) = 0$, because deviations from their own mean add up to zero. The new reading adds its own term.
(2) Insert $m_N - m_{N-1} = \delta/N$ and $x_N - m_N = \delta(N-1)/N$.
(3) $(N-1)/N^2 + (N-1)^2/N^2 = (N-1)N/N^2$.
(4) $\delta \cdot \delta(N-1)/N$ is the surprise before the update times the surprise after it.

This is **Welford's update** [149]: one pass, two stored numbers, and nothing but small differences multiplied together.

Why not the textbook formula? The sum of squares can also be written as $\sum x_i^2 - (\sum x_i)^2/N$, which also needs only two running sums. In exact arithmetic the two are identical. In a computer they are not. The script `24_recursive.py` reads $N = 100000$ values of a frequency counter at 1 GHz with 1 Hz of jitter. Each x_i^2 is about 10^{18}, and a double-precision number carries sixteen significant digits, so each square is rounded by about 100; the difference of the two sums, which should be N times one, is the difference of two numbers near 10^{23} and keeps none of its digits. The textbook formula returns a variance of 21140 Hz^2. Welford's update returns 1.0000116, against 1.0000116 from two passes with accurate summation: a relative error of 3.213×10^{-8}. Subtracting the current mean before squaring is what saves it.

Listing T5.7: Welford's running mean and variance (`code/chT5/24_recursive.py`)

```
for k, xk in enumerate(x, start=1):          # Welford: one pass, one reading at a time
    d = xk - mean_run
    mean_run += d / k
    M2 += d * (xk - mean_run)
```

Readings of different quality. If reading N has variance σ_N^2, weight it by $w_N = 1/\sigma_N^2$. The weighted mean of Section 3.7 updates the same way. With $P_N = \sum_{i \le N} w_i$, the total weight (the inverse of the variance of the mean),

$$P_N = P_{N-1} + w_N, \qquad m_N = m_{N-1} + \frac{w_N}{P_N}\big(x_N - m_{N-1}\big), \tag{T5.85}$$

by the same algebra as (T5.83) with $N-1$ replaced by P_{N-1}. Precisions add, and the gain is the new reading's share of the total precision. For equal weights it is $1/N$.

T5.9.2 Recursive least squares

A straight-line calibration $y = a + bx$ is fitted to points that arrive one at a time. After N points the least-squares solution (unit weights for simplicity) is $\hat{\boldsymbol{\theta}}_N = P_N A_N^{\mathsf{T}} \boldsymbol{y}_N$ with $P_N = (A_N^{\mathsf{T}} A_N)^{-1}$, where row i of A_N is $\boldsymbol{a}_i^{\mathsf{T}} = (1, x_i)$ (Eq. (3.87)). A new point adds one row: $A_{N+1}^{\mathsf{T}} A_{N+1} = P_N^{-1} + \boldsymbol{a}\boldsymbol{a}^{\mathsf{T}}$, with $\boldsymbol{a}$ the new row. Inverting a matrix plus a rank-one term is the Sherman–Morrison formula,

which Section 11.19.2 checks by multiplication: $(M+\boldsymbol{a}\boldsymbol{a}^{\mathsf{T}})^{-1} = M^{-1} - M^{-1}\boldsymbol{a}\boldsymbol{a}^{\mathsf{T}}M^{-1}/(1+\boldsymbol{a}^{\mathsf{T}}M^{-1}\boldsymbol{a})$. With $\boldsymbol{k} \equiv P_N\boldsymbol{a}/(1+\boldsymbol{a}^{\mathsf{T}}P_N\boldsymbol{a})$, the *gain vector*,

$$
\begin{aligned}
&P_{N+1} \overset{(1)}{=} P_N - \boldsymbol{k}\,\boldsymbol{a}^{\mathsf{T}}P_N, \qquad P_{N+1}\boldsymbol{a} \overset{(2)}{=} \boldsymbol{k},\\
&\hat{\boldsymbol{\theta}}_{N+1} \overset{(3)}{=} P_{N+1}\big(P_N^{-1}\hat{\boldsymbol{\theta}}_N + \boldsymbol{a}\,y\big) \overset{(4)}{=} (1-\boldsymbol{k}\boldsymbol{a}^{\mathsf{T}})\hat{\boldsymbol{\theta}}_N + \boldsymbol{k}\,y = \hat{\boldsymbol{\theta}}_N + \boldsymbol{k}\,\big(y - \boldsymbol{a}^{\mathsf{T}}\hat{\boldsymbol{\theta}}_N\big).
\end{aligned}
\tag{T5.86}
$$

(1) Sherman–Morrison with $M = P_N^{-1}$.
(2) $P_N\boldsymbol{a} - \boldsymbol{k}\,(\boldsymbol{a}^{\mathsf{T}}P_N\boldsymbol{a}) = \boldsymbol{k}(1+\boldsymbol{a}^{\mathsf{T}}P_N\boldsymbol{a}) - \boldsymbol{k}\,\boldsymbol{a}^{\mathsf{T}}P_N\boldsymbol{a} = \boldsymbol{k}$.
(3) $A_{N+1}^{\mathsf{T}}\boldsymbol{y}_{N+1} = A_N^{\mathsf{T}}\boldsymbol{y}_N + \boldsymbol{a}\,y$, and $A_N^{\mathsf{T}}\boldsymbol{y}_N = P_N^{-1}\hat{\boldsymbol{\theta}}_N$.
(4) $P_{N+1}P_N^{-1} = \mathbb{1} - \boldsymbol{k}\boldsymbol{a}^{\mathsf{T}}$ by the first line, and $P_{N+1}\boldsymbol{a} = \boldsymbol{k}$.

The fit moves by the gain vector times the surprise $y - \boldsymbol{a}^{\mathsf{T}}\hat{\boldsymbol{\theta}}_N$, the new point's residual from the current line, exactly as the running mean did. Nothing is approximate: the script starts from the exact fit to the first two points, feeds in the other $200-2$ one by one, and after every point compares with the batch solution; the largest difference over the whole stream is 2.748×10^{-14}, rounding error.

Listing T5.8: Recursive least squares for a straight line (`code/chT5/24_recursive.py`)

```
for i in range(2, Nl):
    a = A[i]
    Pa = P @ a
    k = Pa / (1.0 + a @ Pa)                  # gain vector
    theta = theta + k * (ys[i] - a @ theta)  # correct by the prediction error
    P = P - np.outer(k, Pa)                  # Sherman-Morrison update of (A^T A)^{-1}
```

T5.9.3 The Kalman filter: a quantity that drifts

The model. A detector's gain g_t wanders slowly: from one calibration reading to the next it changes by a random amount w_t with mean zero and variance q, a random walk. Each calibration reading measures the current gain with noise: $y_t = g_t + v_t$, with v_t of variance r. In the simulation $\sqrt{q} = 0.002$ (the gain changes by 0.2% per reading) and $\sqrt{r} = 0.05$ (each reading is good to 5%). Both kinds of noise are taken Gaussian and independent. We carry a Gaussian belief about the gain, mean m_{t-1} and variance P_{t-1}, after reading $t-1$.

Predict. The next gain is $g_t = g_{t-1} + w_t$. Its mean is unchanged and its variance is the sum of two independent variances:

$$
m_t^- = m_{t-1}, \qquad P_t^- = P_{t-1} + q. \tag{T5.87}
$$

Update. The reading y_t arrives. By Bayes' theorem the new belief is the prediction times the likelihood of the reading, $\exp[-(y_t-g)^2/2r]$, as a function of g. Take $-2\ln$ of the product and collect powers of g:

$$
\begin{aligned}
&\frac{(g-m_t^-)^2}{P_t^-} + \frac{(y_t-g)^2}{r} \overset{(1)}{=} g^2\Big(\frac{1}{P_t^-}+\frac{1}{r}\Big) - 2g\Big(\frac{m_t^-}{P_t^-}+\frac{y_t}{r}\Big) + \text{const} \overset{(2)}{=} \frac{(g-m_t)^2}{P_t} + \text{const},\\
&\qquad \frac{1}{P_t} = \frac{1}{P_t^-} + \frac{1}{r}, \qquad m_t \overset{(3)}{=} P_t\Big(\frac{m_t^-}{P_t^-}+\frac{y_t}{r}\Big) \overset{(4)}{=} m_t^- + K_t\,\big(y_t - m_t^-\big), \qquad K_t = \frac{P_t^-}{P_t^- + r}.
\end{aligned}
\tag{T5.88}
$$

(1) Expand both squares; terms without g go into the constant.
(2) Complete the square: $Ag^2 - 2Bg = A(g-B/A)^2 - B^2/A$, with $A = 1/P_t$ and $B/A = m_t$.
(3) Read off the centre B/A.

(4) With $P_t = P_t^- r/(P_t^- + r)$: $m_t = (r\, m_t^- + P_t^- y_t)/(P_t^- + r) = m_t^- + P_t^-(y_t - m_t^-)/(P_t^- + r)$.

The precisions add, as for the weighted mean, and the estimate moves towards the reading by the *Kalman gain* K_t, which is the prediction's variance divided by the total: near one when the prediction is poor, near zero when the reading is poor. The new variance is $P_t = (1-K_t)P_t^- = P_t^- r/(P_t^- + r)$. The same completion of the square gives the Gaussian posterior of Section 5.8 with one reading.

Listing T5.9: The Kalman filter for a drifting gain (`code/chT5/24_recursive.py`)

```
83 def kalman(y, q, r, m0, P0):
84     m, Pk = m0, P0
85     est, var = np.empty(len(y)), np.empty(len(y))
86     for t, yt in enumerate(y):
87         Pk = Pk + q                          # predict: the gain may have wandered
88         K = Pk / (Pk + r)                    # update: weigh prediction against reading
89         m = m + K * (yt - m)
90         Pk = Pk * r / (Pk + r)          # = (1 - K) P, written without cancellation
91         est[t], var[t] = m, Pk
92     return est, var
```

Line 85 is the prediction (T5.87), lines 86–87 the gain and the update of the mean, line 88 the update of the variance. The script writes the variance as $P^- r/(P^- + r)$ rather than $(1-K)P^-$: the two are equal, but with a very large starting variance $1 - K$ is the difference of two numbers close to one, and the second form avoids that cancellation.

No drift: the running mean, again. With $q = 0$ and a starting variance so large that the prior carries no information, $1/P_0 \to 0$, the precisions simply add: $1/P_t = t/r$. Then $K_t = P_{t-1}/(P_{t-1} + r) = (r/(t-1))/(r/(t-1) + r) = 1/t$, and (T5.88) is the running mean (T5.83). The script confirms it reading by reading on a stream of 5000, to 4.663×10^{-15}.

With drift: a gain that stops shrinking. With $q > 0$ the variance no longer falls to zero. Each prediction adds q, each update removes a fraction. The variance settles where the two balance, a fixed point of the map $P^- \mapsto P^- r/(P^- + r) + q$, which is itself a fixed-point iteration of Section T5.7. Writing M for the settled prediction variance,

$$M \overset{(1)}{=} \frac{Mr}{M+r} + q \quad\Longleftrightarrow\quad M^2 \overset{(2)}{=} qM + qr \quad\Longleftrightarrow\quad M \overset{(3)}{=} \frac{q + \sqrt{q^2 + 4qr}}{2},$$
$$K_\infty \overset{(4)}{=} \frac{M}{M+r} \approx \sqrt{\frac{q}{r}}, \qquad P_\infty \overset{(5)}{=} \frac{Mr}{M+r} \approx \sqrt{qr} \qquad (q \ll r). \tag{T5.89}$$

(1) The prediction variance after an update and a prediction equals itself.
(2) Multiply by $M + r$: $M^2 + Mr = Mr + qM + qr$.
(3) The positive root of the quadratic.
(4) For $q \ll r$, $M \approx \sqrt{qr} \ll r$, so $M/(M+r) \approx M/r$.
(5) The updated variance; again $M \ll r$.

In the steady state the filter is $m_t = m_{t-1} + K_\infty(y_t - m_{t-1})$: an exponentially weighted average in which reading $t - j$ counts with weight $K_\infty(1-K_\infty)^j$. Its memory is about $1/K_\infty \approx \sqrt{r/q}$ readings: the number of readings over which the drift, $\sqrt{q}$ per reading in a random walk, grows as large as the noise averaged over the same readings. For our detector $K_\infty = 0.03921$, a memory of 25.5 readings, and the error is $\sqrt{P_\infty} = 0.009901$, about $(qr)^{1/4}$.

The rivals. Two simpler estimators are the obvious alternatives. The running mean of all readings so far ignores the drift. A *trailing window* averages the last W readings; how wide should it be? The estimate at time t has two independent errors. The noise of W averaged readings has variance r/W. And the window's average gain lags the current gain: $g_t - \frac{1}{W}\sum_{j=0}^{W-1} g_{t-j} = \frac{1}{W}\sum_{l=0}^{W-2}(W-1-l)\,w_{t-l}$, since the increment w_{t-l} separates g_t from the $W-1-l$ older gains $g_{t-l-1}, \dots, g_{t-W+1}$. So

$$\text{error}^2(W) \overset{(1)}{=} \frac{r}{W} + \frac{q}{W^2}\sum_{m=1}^{W-1} m^2 \approx \frac{r}{W} + \frac{qW}{3}, \qquad W^* \overset{(2)}{=} \sqrt{\frac{3r}{q}}, \qquad \text{error}^2(W^*) \overset{(3)}{=} 2\sqrt{\frac{rq}{3}}. \quad \text{(T5.90)}$$

(1) Independent increments: the variance of the lag is q/W^2 times the sum of the squared coefficients, and $\sum_{m=1}^{W-1} m^2 \approx W^3/3$.
(2) Set the derivative $-r/W^2 + q/3$ to zero.
(3) Both terms equal $\sqrt{rq/3}$ at W^*.

For our detector $W^* = 43.3$ readings and the error is 0.01075, slightly worse than the Kalman filter's 0.009901. The simulation, 200 streams of 5000 readings each, with the first 500 discarded, measures all of them (Figures T5.28 and T5.29). The Kalman filter is off by 0.009867 on average, as (T5.89) predicts. The best trailing window, found by trying every width, has $W = 42$ and an error of 0.01064, 7.238% larger; the filter wins on 100% of the streams. The running mean of all readings is off by 0.06061: it is biased by the drift, and more data make it worse, not better.

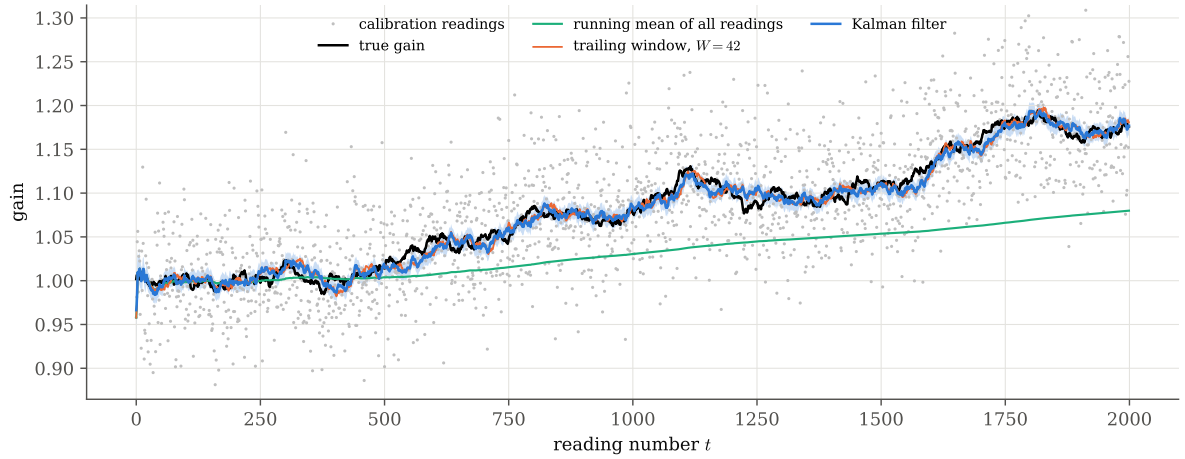


Figure T5.28: One simulated stream of calibration readings (grey dots) of a gain that wanders as a random walk (black). The running mean of all readings (green) falls behind the drift and never catches up. A trailing window of the best width (orange) and the Kalman filter (blue, with its $\pm 1\sigma$ band) both follow it; the filter is slightly closer, with no window to choose.

What it needs. The filter needed q and r. The noise r of a calibration reading is usually known from the calibration itself. The drift q is rarely known, but the filter is forgiving (Figure T5.29, right): telling it a drift variance ten times too small or too large raises its error from 0.009867 to 0.01308 and 0.01308. A wrong guess of the drift is equivalent to a wrong window width, and both are cured the same way: estimate it from the data.

How this is used in real research: How research uses recursive estimators

Our filter tracks one number with known noise and drift. Real tracking problems carry a state vector (gain, offset, temperature coefficient), a model of how it evolves, and covariance matrices Q and R; the predict and update steps are the matrix versions of (T5.87)–(T5.88),

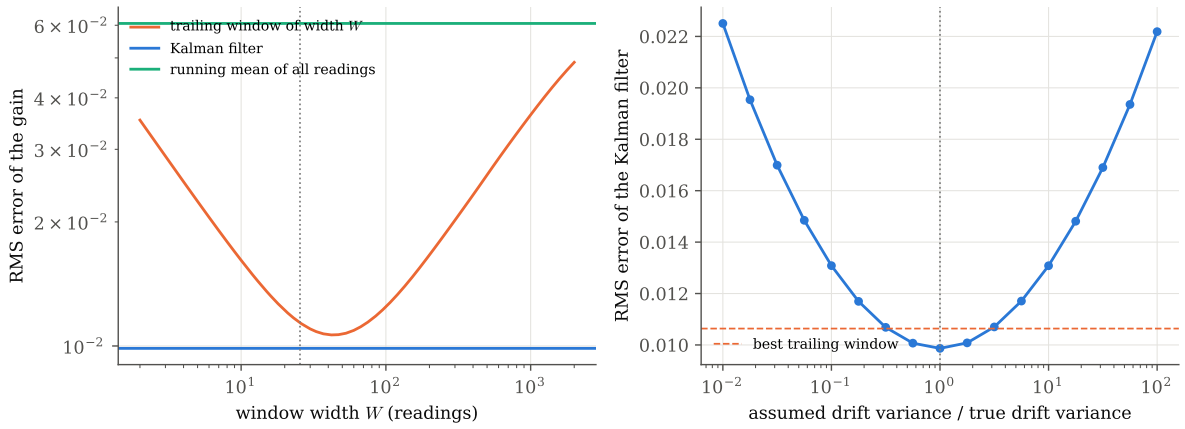


Figure T5.29: Left: the error of a trailing window against its width (orange): narrow windows are noisy, wide ones lag. Its minimum is above the error of the Kalman filter (blue); the running mean of all readings (green) is six times worse. The dotted line marks the filter's memory $1/K_\infty$. Right: the error of the Kalman filter when it is told a drift variance that is wrong by the factor on the axis. A factor of ten either way costs about a third.

with the gain $K = P^- H^\mathsf{T}(HP^- H^\mathsf{T} + R)^{-1}$ for readings $\boldsymbol{y} = H\boldsymbol{g} + \boldsymbol{v}$ [148]. Q and R are fitted by maximising the likelihood of the surprises $y_t - m_t^-$, which the filter delivers as independent Gaussians with variance $P_t^- + r$. When the whole stream is available afterwards, a backward pass (the Kalman smoother) uses later readings as well, and its error is smaller than the filter's. The same recursion fits particle tracks in a tracking detector, hit by hit, and follows instrument gains through long observing campaigns.

Iterative or recursive? The two families look alike, since both repeat an update, but they answer different questions. An iterative estimator repeats on the *same* data until a self-consistent answer stops changing. A recursive estimator does one update per *new* reading and never revisits the old ones. The Kalman filter's variance settling to its steady state is where the two meet: a recursion whose variance runs to a fixed point.

T5.10 A field guide: which estimator, when

What we already have. Six questions about a problem each point to a family of estimators (Section T5.1.1), and a table records what each estimator assumes and what it optimises (Section T5.1.2). Three more families have joined them: estimators that iterate to a self-consistent answer, estimators that remove their own bias, and estimators that update with each reading.

When that is not enough. A real analysis rarely asks one question. A power spectrum from a masked map is a parameter in the covariance, estimated at hundreds of multipoles, with weights that need the spectrum itself, and it will be averaged into bands. The answers interact, some questions should come first, and each choice has a price that is worth stating before committing to it.

A story with the whole trick in it. A good mechanic facing a car that will not start does not try every repair in the manual. She asks questions in an order: is there fuel, is there a spark,

does the starter turn. Each answer rules out a whole branch, the cheap questions come first, and the repair at the end follows from the answers. She also knows what each repair costs and what it can break. Choosing an estimator works the same way. We ask, in order, where the parameter lives, what we know about the noise, whether the recipe needs its own answer, what else we know, how the data arrive, and how the estimate will be used; the estimator is assembled from the answers, and its cost is checked by simulation.

The roadmap figure (Figure T5.1) listed the questions and the family each answer points to. Figure T5.30 turns them into a procedure that can be followed: the first three questions are a tree that ends in the form of the estimator, and the other four are modifiers asked of whatever form came out. Table T5.1 lists situations with the estimator each calls for, what it assumes, what it costs, and where the book develops it. The six cases after them are worked in the order of the questions.

T5.10.1 Six cases

Each case follows the order of reasoning of the Monty Hall problem (Chapter 1): what was observed, what we know about how it was produced, the candidate estimators one at a time with the reason for or against each, and only then the choice.

Case 1: counts on a background. *The data.* A search counts events in a signal window in several detectors or bins; each has a background expectation b_i known from sidebands, and the expected signal $s\,\epsilon_i$ differs between them through the efficiency ϵ_i. *What do we know?* Counts are Poisson: the noise law is known exactly, and its variance is the unknown mean. *The candidates.* Background-subtracted counts averaged with equal weights ignore that some channels are cleaner. Weights $1/n_i$ from the observed counts favour channels that fluctuated low, the Neyman bias of Section T5.7.2. Minimising Pearson's χ^2 directly pushes the estimate up. Weights $1/(s\epsilon_i + b_i)$ re-solved from the current estimate converge to the Poisson maximum-likelihood estimate, which question 4 told us to expect. *The choice:* the Poisson likelihood, maximised by re-solved weights or by any maximiser, and an interval from the likelihood as in Section 4.13.4 when the counts are small.

Case 2: the angular power spectrum of the CMB. *The data.* A map, decomposed into $a_{\ell m}$, with instrument noise of known spectrum N_ℓ. *What do we know?* The parameter C_ℓ is the variance of the $a_{\ell m}$: their mean is zero whatever C_ℓ is. So question 1 already decides: no linear combination of the data can see C_ℓ, and the estimator is quadratic. *The candidates.* The plain average of $|a_{\ell m}|^2$ over m includes the noise variance, a bias equal to N_ℓ, which is why $\hat{C}_\ell = \frac{1}{2\ell+1}\sum_m |a_{\ell m}|^2 - N_\ell$ subtracts it (Eq. (7.82)). On a masked sky the modes are coupled, and the best quadratic weights depend on C_ℓ itself (question 4): either a fiducial spectrum is used, or the estimate is iterated with its own output as the fiducial. Question 7 matters too: each $\hat{C}_\ell$ at low ℓ rests on $2\ell + 1$ modes, so any nonlinear function of it, such as $\ln \hat{C}_\ell$ or $1/\hat{C}_\ell$, carries a bias of order $1/(2\ell + 1)$ that does not average away over many multipoles (Section T5.8.4). *The choice:* the unbiased quadratic estimator (a pseudo-C_ℓ with its mode-coupling correction, Section 8.11, or the quadratic maximum-likelihood estimator of Section T5.6), with the noise bias subtracted, and nonlinear functions of it left to the likelihood instead of being averaged.

Case 3: photometry with cosmic rays. *The data.* Twenty exposures of a star, each giving a flux. *What do we know?* Most exposures have Gaussian noise from photon counting and read-out; a few are hit by cosmic rays, which only ever *add* charge. This is laboratory C of Section T5.1, with a one-sided contamination. *The candidates.* The mean is pulled up by every hit and never pulled

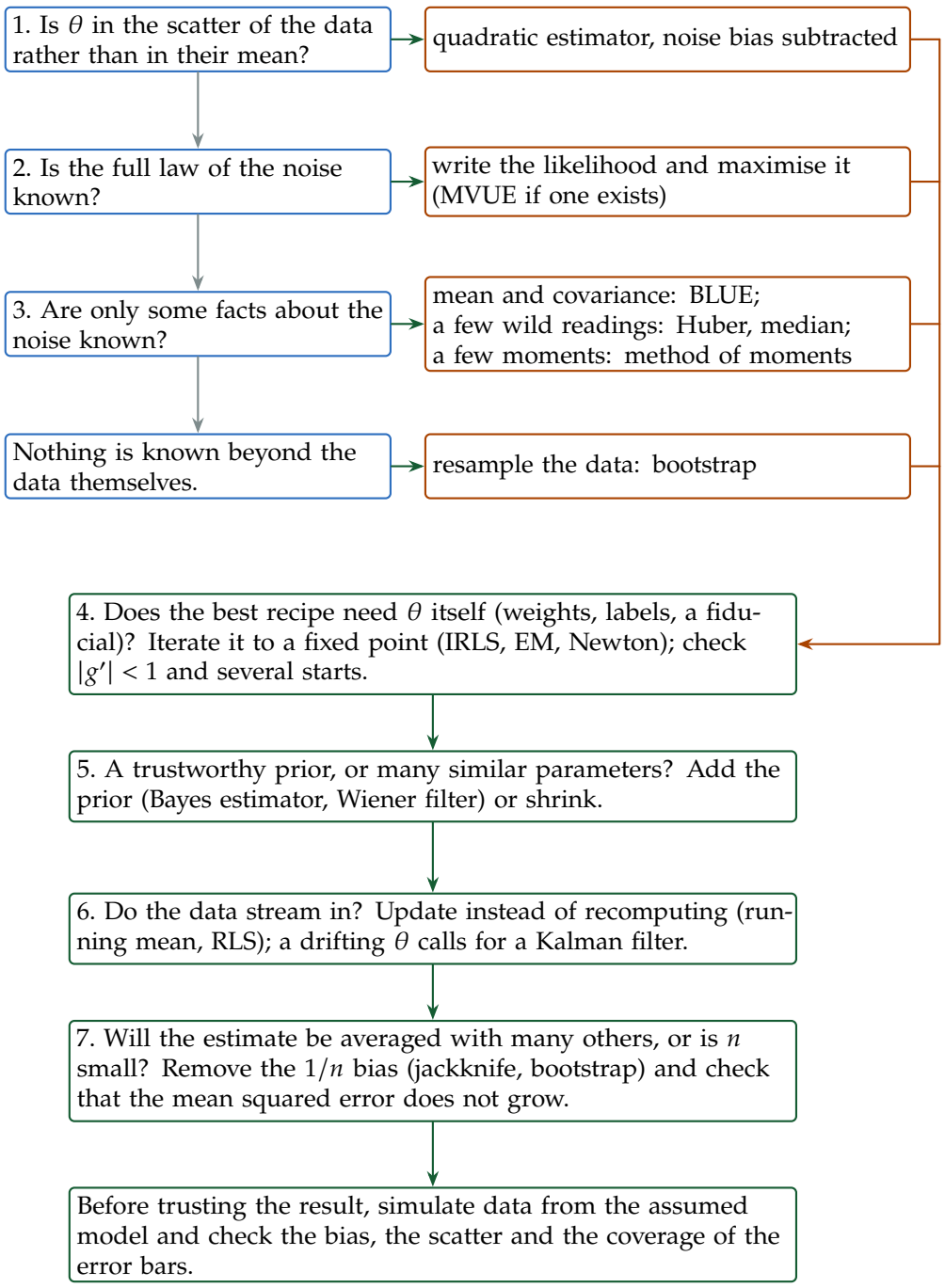


Figure T5.30: The operational procedure for choosing an estimator, drawn from the questions of the chapter roadmap (Figure T5.1). Top: a tree. Green arrows answer yes, grey arrows answer no and lead to the next question, and every branch ends in the form of the estimator (orange). Bottom: four modifiers, each asked of whatever form came out, in order. A yes adds an ingredient to the recipe and the procedure goes on to the next question. The last box is not optional.

Table T5.1: Situations and the estimators they call for.

situation	estimator	assumes	costs	where
Gaussian readings, known errors	weighted mean, least squares	known σ_i	nothing	Sections 3.7 and 3.10
full noise law, a complete sufficient statistic	minimum-variance unbiased	the full law	rarely available in closed form	Section 3.16.5
full noise law, large n	maximum likelihood	the full law	biased at order $1/n$	Section 3.9
a few moments known as functions of θ	method of moments	those moments only	less efficient than ML; a good starting point for iterations	Section 3.9.2
only mean and covariance of the noise known	BLUE	linear model	inefficient for heavy tails	Section T5.2
counts, few per bin	Poisson ML (re-solved weights)	Poisson law	an iteration; never weights $1/y_i$	Sections 3.11 and T5.7.2
a few wild readings	Huber or median, by IRLS	most readings well behaved	a little efficiency; choice of k	Sections T5.5 and T5.7.3
θ in the covariance	quadratic estimator	Gaussian field, known noise	noise bias to subtract; needs a fiducial	Section T5.6
prior information	Bayes estimator, Wiener filter	a trustworthy prior	biased if the prior is wrong	Section T5.3
many similar parameters	shrinkage	exchangeable parameters	each estimate biased	Section T5.4
weights need θ	fixed-point iteration, IRLS	contraction, $\lvert g' \rvert < 1$	iterations; may not converge	Section T5.7
hidden labels, mixtures	EM	a complete-data model	slow when labels are unclear; local maxima	Section T5.7.4
ill-posed inversion	regularisation or stopping early	smooth truth	bias towards the start	Sections T5.7.5 and 6.14.3
small n, estimate to be combined	jackknife or bootstrap correction	bias in powers of $1/n$	more variance	Section T5.8
stream, constant θ	running mean, Welford, RLS	nothing new	nothing	Sections T5.9.1 and T5.9.2
stream, drifting θ	Kalman filter	drift and noise variances	needs q, r	Section T5.9.3

down, so it is biased, not merely inefficient. The median is robust but throws away about a third of the information when there are no hits. The Huber estimator, computed by reweighted least squares with k about one and a half standard deviations (Section T5.7.3), keeps almost all the information and limits the pull of each hit to k; the popular *sigma clipping*, which repeatedly discards readings more than a few standard deviations from the current mean, is the same fixed-point iteration with weights that jump from one to zero. Because the hits are one-sided, even a robust estimator keeps a small upward bias, of order the fraction of hits times k. *The choice:* Huber by IRLS, with k and the remaining bias checked on simulated exposures with the observed hit rate.

Case 4: many band powers. *The data.* Fifty band powers of a spectrum, each with its own error. *What do we know?* Each is a fine estimate on its own, but neighbouring bands of a smooth spectrum are similar (question 5). *The candidates.* Shrinking every band towards a smooth fit lowers the total squared error of the fifty, as James and Stein showed (Section T5.4), but makes each band biased towards the fit. If the band powers are the input to a cosmological fit, that bias is harmful: the likelihood already combines the bands with the right weights, and shrinking would pull the result towards the smooth model assumed in the shrinkage. If the purpose is a picture or a denoised spectrum, shrinkage is the right tool. *The choice* depends on the use: unbiased band powers with their covariance for inference, shrinkage for display or for a first look.

Case 5: a detector's calibration, streamed. *The data.* A calibration pulse every second, each measuring the gain with a known noise. *What do we know?* The gain may drift with temperature, slowly but without a known law. *The candidates.* The running mean of all pulses is exact if nothing drifts and biased if anything does (Section T5.9.3). A trailing window of the right width works, but the width has to be tuned. The Kalman filter with a random-walk drift needs the drift variance, can estimate it from its own surprises, and beat the best window on every simulated stream. *The choice:* the Kalman filter for live monitoring; the Kalman smoother, which also uses later pulses, when the calibration is applied afterwards.

Case 6: the lensing potential. *The data.* A CMB map in temperature and polarisation. *What do we know?* Lensing correlates modes that would be independent on an unlensed sky: ψ lives in the off-diagonal part of the covariance (question 1), so the estimator is quadratic in the map, and its spectrum carries the biases $N^{(0)}$ and $N^{(1)}$ (Section T1.9). *The candidates.* The quadratic estimator with the Hu–Okamoto weights is optimal at first order and fast. Its biases must be removed (question 7), by formulas checked on simulations and by the realisation-dependent $N^{(0)}$ of Section T1.9. For polarisation at low noise its weights depend on the lensing B modes it is estimating (question 4), and the iterated reconstruction of Section T5.8.5 reaches noise levels the quadratic estimator cannot. *The choice:* the quadratic estimator for temperature and for noisy polarisation; the iterative or maximum-a-posteriori reconstruction for deep polarisation maps; in both cases bias subtraction calibrated on simulations that pass through the same pipeline.

What the cases share. No case needed an exotic estimator. Each needed the questions asked in the right order, and in four of the six the decisive question was one the data alone cannot answer: what the noise is, whether the parameter drifts, how the estimate will be used. The final box of Figure T5.30 applied in every case: when the estimator is chosen, simulate data under the assumptions that chose it and measure its bias and its scatter.

Recap

- When the best weights need the parameter, iterate $\theta_{k+1} = g(\theta_k)$. It converges when $|g'(\theta^*)| < 1$, linearly with that factor per pass; Newton has $g' = 0$ and converges quadratically. For two detectors, $11 \to 9.444 \to 9.349 \to 9.343$, with factor 0.065.
- Re-solved weights $1/\lambda_i(\theta)$ for counts land on the Poisson score equation: the maximum-likelihood estimate. Weights from the data (Neyman) bias low, Pearson's minimisation biases high by about half as much.
- IRLS solves $\sum_i w(r_i) r_i \, \partial\mu_i/\partial\theta = 0$ by weighted least squares: Huber's robust mean (rate = share of weight on clipped points, never uphill) and Poisson regression (working

response $\eta + (y - \lambda)/\lambda$, weights λ; equal to Newton and to Fisher scoring).

- EM: soft labels by Bayes' theorem, then count. Jensen's inequality makes the likelihood rise at every pass; the rate is the fraction of information the missing labels take away, $1 - I_{\text{obs}}/I_{\text{com}}$.
- Stopping an iteration after k passes regularises like Tikhonov with $\tau \approx 1/(\omega k)$.
- A smooth function of a mean is biased by $g''\sigma^2/2n$; $1/\bar{t}$ by exactly $\lambda/(n-1)$. The jackknife $nT - (n-1)\bar{T}_{(\cdot)}$ removes the $1/n$ term; the bootstrap $2T - \mathbb{E}^* T^*$ does too; iterating is a fixed-point iteration of the bias equation. A correction can raise the mean squared error; it pays when more than about n estimates are averaged.
- Recursive estimators update by a gain times the surprise: running mean (gain $1/N$), Welford's variance, RLS by Sherman–Morrison, and the Kalman filter (predict: variances add; update: precisions add), whose gain settles at $\approx \sqrt{q/r}$ and beats the best moving window.
- Choosing an estimator: ask where θ lives, what the noise is, whether the recipe needs θ, what else is known, how the data arrive and how the result will be used; then check by simulation.

Chapter summary

concept	what it means in practice	where it is used
the questions that choose an estimator	where θ lives, what the noise is, prior, how many parameters, stream, self-dependence, use of the result	every analysis; the field guide, Section T5.10
best linear unbiased estimator	weights $(A^\mathsf{T}C^{-1}A)^{-1}A^\mathsf{T}C^{-1}$: needs only the mean and covariance of the noise	generalised least squares in Section 11.19.2
Bayes risk, Wiener filter	average error over a prior; Gaussian prior and noise give the Wiener filter	priors in Chapter 5; filtered maps
shrinkage (James–Stein)	many similar means estimated together beat one-by-one ML in total squared error	hierarchical models in Chapter 5
robust (Huber, median)	bounded pull of each reading; computed by IRLS	outlier-prone readings; robust summaries in Section 12.9.4
quadratic estimator, noise bias	a parameter in the covariance needs a quadratic estimator; noise adds a bias to subtract	$\hat{C}_\ell$ in Eq. (7.82); Section 8.3; lensing Section T1.9
fixed-point iteration, contraction factor	repeat $\theta \to g(\theta)$; converges if $\lvert g'(\theta^*)\rvert < 1$, one digit per $1/\log_{10}(1/\lvert g'\rvert)$ passes	re-solved weights in Section 11.16.3
re-solved Poisson weights	weights $1/\lambda_i(\theta)$ recomputed until fixed: the Poisson ML estimate	binned fits; Section 3.11
IRLS	frozen weights, weighted least squares, repeat; Newton and Fisher scoring as special cases	Fisher scoring in Section 10.6.6
EM algorithm	soft labels by Bayes, then count; likelihood never falls; rate $= 1 - I_{\rm obs}/I_{\rm com}$	iterative unfolding, Section 6.14.5
stopping early	k passes act like regularisation of strength $\approx 1/(\omega k)$	choosing the number of unfolding iterations, Section 6.14.5
$1/n$ bias	a curved function of a mean is biased by $g''\sigma^2/2n$	any nonlinear summary of few data
jackknife bias correction	$nT - (n-1)\bar{T}_{(\cdot)}$ removes the $1/n$ term	jackknife errors in Section 13.18
bootstrap bias correction, iterated	$2T - \mathbb{E}^*\,T^*$; $3T - 3\,\mathbb{E}^*\,T^* + \mathbb{E}^{**}\,T^{**}$; fixed-point view of the bias equation	with the bootstrap of Section 3.15
price of a correction	variance may rise; worth it when more than $\sim n$ estimates are averaged	band powers, stacked measurements
running mean, Welford	update by a gain times the surprise; subtract the mean before squaring	running averages of Markov chains, Figure 9.16
recursive least squares	Sherman–Morrison update, identical to the batch fit	streamed calibration fits
Kalman filter	predict (variances add), update (precisions add); gain $\approx \sqrt{q/r}$ for a drifting quantity	monitoring detector gains
iterative lensing reconstruction	delens, re-estimate: noise falls when B modes are lensing-dominated	CMB lensing, Section T1.9

Reading the literature

- L. Wasserman, *All of Statistics*, ch. 8 (the bootstrap, with the jackknife in its appendix) and §9.13 (Newton–Raphson and the EM algorithm).
- G. Casella and R. L. Berger, *Statistical Inference*, §7.2.4 (EM, with its monotonicity theorem), §7.5.4, Exercise 7.66 (the jackknife), §10.1.4 (bootstrap standard errors), §10.2 (robustness, M-estimators and the Huber estimator).
- G. Cowan, *Statistical Data Analysis*, ch. 6–7: the bias of maximum-likelihood estimators and least squares with binned data.
- Linear, Bayes and shrinkage estimators: [126–130]; robust estimation: [131, 132]; quadratic estimators of a power spectrum: [133].
- Iteration: EM [139] and its convergence [140]; reweighted least squares [137, 138]; iterative deconvolution and unfolding [135, 136, 141].
- Bias correction: the jackknife [142, 143], the bootstrap [23] and its iteration [144].
- Recursive estimation: [148, 149].
- Lensing reconstruction beyond the quadratic estimator: [145–147].

Part III

Simulation, Fields and Forecasts

Chapter 6

Monte Carlo: Drawing Samples on a Deterministic Machine

Where this chapter is going

A call such as `rng.random(N)` supplies N numbers that we use as independent draws from the uniform distribution on $[0, 1)$. By transforming these draws we can sample other distributions, then estimate probabilities, expectations and integrals by averaging. Yet a generator follows a fixed rule for updating its internal *state*, so starting from the same state reproduces the entire sequence. Its suitability for a simulation therefore depends on whether the regularities in that sequence affect the quantities we intend to estimate.

One regularity is unavoidable. A generator has a finite number of states, so it must eventually repeat, and after that no new information enters a simulation. The simplest generators, the linear congruential ones, put their consecutive outputs on a crystal lattice: in three dimensions the triples of IBM's famous RANDU lie on just fifteen planes. Their lowest bits repeat with tiny periods. Modern generators (xorshift, the Mersenne Twister, and numpy's default PCG64) keep a larger state, update it by a rule that is cheap and has an enormous period, and pass it through an output function that hides the regularities the update leaves behind. Seeds and streams let us rerun any simulation bit for bit and run thousands of independent copies side by side. Whether a generator is good enough is a hypothesis test: H_0 is "independent and uniform", and the test statistics are histograms, correlations, runs and tuples. Finally, when the goal is an integral rather than a random process, we may drop randomness altogether: quasi-random points fill space more evenly than independent points and beat the $1/\sqrt{N}$ error.

From uniform numbers we then build draws from any distribution (inverse transform, Box–Muller, rejection and importance sampling), and averages over such draws obey a $\sigma/\sqrt{N}$ law that controls every Monte Carlo estimate. On the book's pipeline (Chapter 0), where a statistic S computed from the data must have its scatter $P(S \mid \theta)$ worked out before any physics can be inferred, Monte Carlo is how we find that scatter when no formula exists: we simulate the data many times and histogram the summary.

6.1 Random numbers from a machine that cannot be random

A simulation is only as good as the random numbers that drive it, and a computer cannot make random numbers. In Chapter 1 we settled the Monty Hall problem twice: once with a probability tree and once by letting the computer play hundreds of thousands of games and counting. The counting looked the more trustworthy of the two, because it needed no formula. But every one of those games was decided by numbers that a program produced, and

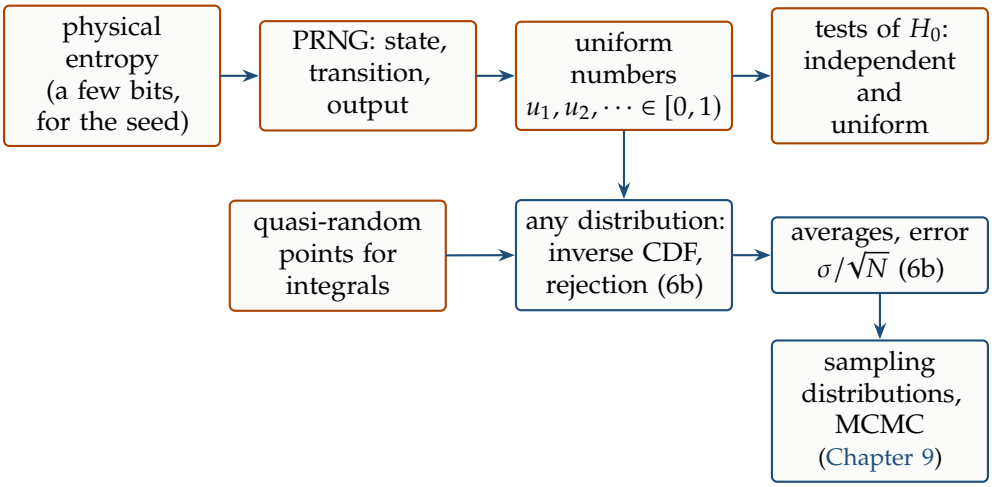


Figure 6.1: How the ideas of the chapter connect. A few bits of physical entropy seed a deterministic generator, whose output is accepted as uniform only after it survives statistical tests; quasi-random points replace the generator when the goal is an integral (orange boxes). Uniform numbers are then turned into draws from other distributions and averaged, with an error that falls as $1/\sqrt{N}$ (blue boxes). The results feed every sampling method of later chapters.

a program is a list of instructions carried out exactly. So two questions arise. Where did the randomness come from? And how would we know if it were not good enough? Monty Hall is a good place to find out. Its answer rests on three numbers, the probability of what Monty did under each possible location of the car, and a simulation must get each of them right.

6.1.1 What a simulation of Monty Hall has to reproduce

Recall the game. A car is behind one of three doors A, B, C, each with probability $\frac{1}{3}$; goats are behind the other two. We pick door A. The host, Monty, knows where the car is. He always opens one of the two doors we did not pick, never the one with the car, and when both of those hide goats he chooses between them at random, like tossing a fair coin [Wasserman p. 14, Ex. 10]. He opens door C. Should we switch to B?

The event. Monty opened door C. Write E for it; every simulated game that ends this way counts towards it.

What could have made him open C? Monty's move depends on one thing we cannot see, where the car is. So the candidate explanations are the three places the car can be: H_A, H_B and H_C, the car behind A, B or C. For each one we ask: if the car were there, how likely was Monty to open C?

How probable is the event under each location? This is the step a simulation has to reproduce game by game. Monty knows where the car is, he never reveals it, and he never opens our door.

- Under H_A the car is behind our door A, so both B and C hide goats. Monty may open either and tosses his coin: $\mathrm{P}(E \mid H_A) = \frac{1}{2}$.
- Under H_B the car is behind B. He may not open our door A and may not reveal the car behind B, so C is the only door he can open: $\mathrm{P}(E \mid H_B) = 1$.
- Under H_C the car is behind C. Opening C would reveal it, so he never does: $\mathrm{P}(E \mid H_C) = 0$.

These three numbers are the second layer of the tree in Figure 6.2.

Only then the priors, and the normalisation. Before Monty acted, each location had prior probability $\mathrm{P}(H_A) = \mathrm{P}(H_B) = \mathrm{P}(H_C) = \frac{1}{3}$. We multiply each prior by the probability of the event

under it, and divide by the sum of these products:

$$\mathrm{P}(H_B \mid E) \overset{(1)}{=} \frac{\mathrm{P}(E \mid H_B)\,\mathrm{P}(H_B)}{\mathrm{P}(E \mid H_A)\,\mathrm{P}(H_A) + \mathrm{P}(E \mid H_B)\,\mathrm{P}(H_B) + \mathrm{P}(E \mid H_C)\,\mathrm{P}(H_C)}$$
$$\overset{(2)}{=} \frac{1 \cdot \frac{1}{3}}{\frac{1}{2} \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3}} = \frac{1/3}{1/2} = \frac{2}{3}. \tag{6.1}$$

(1) Bayes' theorem (Chapter 5) for the three hypotheses, which are mutually exclusive and exhaustive: the denominator is the total probability $\mathrm{P}(E)$ of the observation, summed over every path in the tree that ends in "Monty opened C".

(2) The three probabilities of E found above and the equal priors. The denominator is $\mathrm{P}(E) = \frac{1}{6} + \frac{1}{3} = \frac{1}{2}$.

In the same way $\mathrm{P}(H_A \mid E) = \frac{1}{6}/\frac{1}{2} = \frac{1}{3}$ and $\mathrm{P}(H_C \mid E) = 0$. Switching to B wins with probability $\frac{2}{3}$.

The guess "one door is gone, so the other two are fifty–fifty" fails because it ignores the information in Monty's choice: the event we saw was not equally likely under the two surviving hypotheses. Among the games in which the car is behind B, he opens C every time; among those in which it is behind our door A, only half the time. So the opening we saw is twice as probable under H_B as under H_A, and the two surviving posteriors carry that same factor of two.

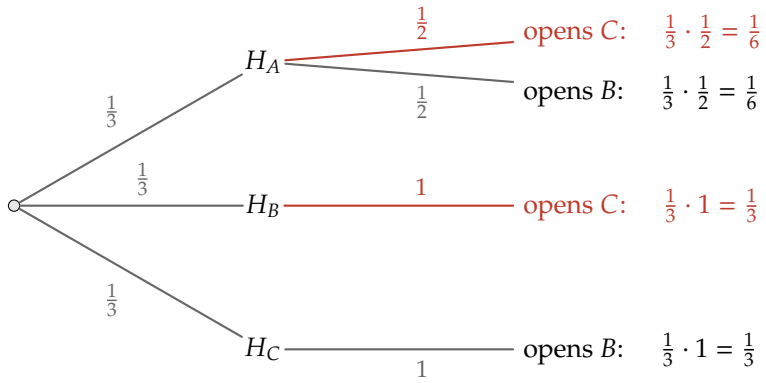


Figure 6.2: The Monty Hall tree when we have picked door A. The first split is the location of the car (the prior, $\frac{1}{3}$ each). The second split is Monty's choice under each location: a fair coin under H_A, a forced choice under H_B and H_C. The red leaves are the paths that end in the observation "Monty opened C"; their weights $\frac{1}{6}$ and $\frac{1}{3}$ are in the ratio 1 : 2, which is the posterior.

A simulation follows the same tree forwards. For each game it draws the car's location (the first split), then Monty's choice (the second split), then keeps the games that ended in "Monty opened C" and counts where the car was (Chapter 1 did exactly this). The simulation contains no formula for the posterior, but it must reproduce the three probabilities $\mathrm{P}(E \mid H_A) = \frac{1}{2}$, $\mathrm{P}(E \mid H_B) = 1$, $\mathrm{P}(E \mid H_C) = 0$ as frequencies. The last two are forced by Monty's rules. The first is produced by nothing but the random numbers that toss Monty's coin. If those numbers are wrong, the simulation computes the posterior of a different game.

Example: the same game with a bad coin. A simulation needs two random integers x per game. The first places the car: its remainder on division by 3 picks the door. The second tosses Monty's coin: its remainder on division by 2 picks heads or tails. Here is the heart of such a simulation.

Listing 6.1: One simulated game per pair of integers (`code/ch06/01_monty_generators.py`)

```
28     """ints has 2N entries; game k uses ints[2k] for the car and ints[2k+1] for the coin."""
```

```
    car = ints[0::2] % 3                          # 0 = A, 1 = B, 2 = C
    coin = ints[1::2] % 2                         # Monty's tie-break when the car is behind A
    # We picked A. Monty opens a door that is neither ours nor the car's.
    opened = np.where(car == 0, np.where(coin == 1, 2, 1),   # free choice: B or C
                      np.where(car == 1, 2, 1))              # forced choice
    lik = np.array([np.mean(opened[car == h] == 2) for h in range(3)])   # P(open C | H)
    sel = opened == 2
    post = np.array([np.mean(car[sel] == h) for h in range(3)])          # P(H | open C)
    switch = np.mean(car != 0)
    return lik, post, switch
```

Line 29 places the car, line 30 tosses the coin. Lines 32–33 are Monty's rules; line 34 measures $\mathrm{P}(E \mid H)$ as the fraction of games with the car behind each door in which he opened C; lines 35–36 select the games with "Monty opened C" and count where the car was. We feed it integers from two sources. The first is numpy's generator, through `rng_for`. The second is a textbook generator with a long history, the rule $x \mapsto (1664525\,x + 1013904223) \bmod 2^{32}$, which we will meet again in Section 6.2. Both play 300,000 games.

With numpy's integers the measured probabilities of E are 0.500, 1 and 0 under H_A, H_B, H_C, and among the games with "Monty opened C" the car was behind A in a fraction 0.333 and behind B in 0.667: the answer $\frac{1}{3}$, $\frac{2}{3}$ of (6.1). With the textbook generator started from the seed 2026, the coin always comes up the same way. Monty, facing a free choice, *never* opens C: the measured $\mathrm{P}(E \mid H_A)$ is 0.000, and among the games in which he opened C the car was behind B in a fraction 1.000. Started from the seed 2027, the coin always comes up the other way, Monty *always* opens C when he may, $\mathrm{P}(E \mid H_A)$ becomes 1.000, and the car was behind B in only 0.501 of the selected games. That is the fifty–fifty answer of the wrong intuition, now "confirmed" by a simulation (Figure 6.3).

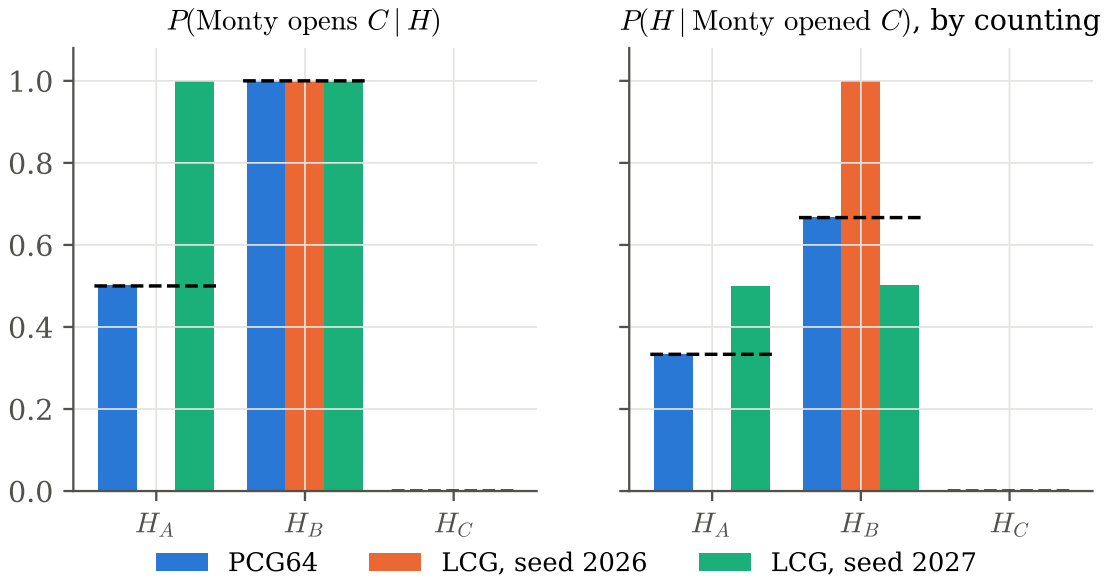


Figure 6.3: Monty Hall simulated with three sources of integers. Left: the measured probability that Monty opens C under each location of the car. With PCG64 (blue) it is $\frac{1}{2}$, 1, 0 (dashed). With the textbook generator the coin never varies, so under H_A Monty either never (orange) or always (green) opens C. Right: where the car was, among the games in which Monty opened C. The wrong probability under H_A produces the wrong posterior: 0 and 1, or $\frac{1}{2}$ and $\frac{1}{2}$, instead of $\frac{1}{3}$ and $\frac{2}{3}$.

This failure is dangerous because nothing gives it away. First, the fraction of all games that switching wins is 0.667 with numpy, 0.667 and 0.667 with the textbook generator: the headline number is right in all three runs, because it does not involve the coin. The damage is visible

only in the conditional question. Second, nothing crashed and nothing looked odd; the car's location, taken from the same generator, is perfectly uniform. The cause, derived in Section 6.2, is that the lowest bit of this generator alternates $0, 1, 0, 1, \ldots$, so every second integer, the one used for the coin, has the same lowest bit.

6.1.2 What we ask of a stream of random numbers

The example shows what a source of random numbers must do. A simulation turns numbers $u_1, u_2, \ldots$ into events, and every probability it reports is a frequency of those events. So the simulation is correct if the u_i behave, for the purpose at hand, like **independent draws from the uniform distribution** [every u_i equally likely to fall anywhere in $[0, 1)$, and knowing some of them tells nothing about the others] on $[0, 1)$, whose density is

$$g(u) = \begin{cases} 1, & 0 \le u < 1, \\ 0, & \text{otherwise} \end{cases} \tag{6.2}$$

[Cowan (3.1)].

Every Monte Carlo method assumes such numbers. Monte Carlo integration takes them for granted: to estimate $\int_a^b h(x)\,\mathrm{d}x$ we "generate $X_1, \ldots, X_N \sim \mathrm{Unif}(a, b)$" and average [Wasserman p. 404, (24.2)]. So does the idea of representing a distribution by a large sample of representative values [Kruschke §7.1]: the numbers x built from the u_i are treated as simulated measurements, and the probability of any region is estimated from how many of them fall in it. The payoff is greatest in many dimensions, where integrating a joint density over a complicated region by other methods may not be feasible [Cowan Ch. 3]. In practice we ask a stream of numbers for five things at once.

1. *Statistical quality:* no test that matters for our application can distinguish the numbers from independent uniform draws.
2. *A long period:* the sequence must not repeat within any computation we will ever run.
3. *Reproducibility:* running the program again gives the same numbers, so that a result can be checked and a bug can be found.
4. *Speed:* a large simulation consumes 10^9 to 10^{12} numbers.
5. *Independent streams:* many simulations running at once each need their own numbers, with no hidden relations between them.

The obvious source of randomness is physics. Cowan mentions tossing a coin [Cowan §3.1]; real devices time radioactive decays, amplify the thermal (Johnson) noise of a resistor or the shot noise of a diode, or measure which output port a single photon takes at a beam splitter, where quantum mechanics itself guarantees that no hidden variable decides the outcome [Lambert §12.7, fn. 3]. The operating system of every computer collects such **physical entropy** [bits whose values nobody, including the computer itself, could have predicted] from the timing jitter of keystrokes, disk and network interrupts, and from hardware noise sources on the processor chip. Physical entropy fails requirements 3 to 5 of the list above. It cannot be replayed, so no result can be reproduced (requirement 3). It is far too slow to feed one large simulation, let alone many running at once (requirements 4 and 5). And its raw bits are biased and correlated, so they must be post-processed before use. We therefore use it for one thing only: to choose a starting point, the seed. Everything else comes from an algorithm.

Working definition 6.1 (pseudorandom number generator)

A **pseudorandom number generator** (PRNG) is a finite set S of **states** [all the bit patterns its memory can hold], a **transition function** $f : S \to S$, an **output function** g from states to numbers, and a **seed** [the state s_0 we start from, chosen by us]. It produces

$$s_{n+1} = f(s_n), \qquad u_n = g(s_n), \qquad n = 1, 2, \ldots$$

The sequence $u_1, u_2, \ldots$ is completely determined by s_0. It is called **pseudorandom** [Cowan §3.1] because it is accepted as random only in the operational sense of requirement 1. *Operational test:* apply statistical tests of the hypothesis "independent and uniform" (Section 6.6), and above all check that the application gives the same answer with a second, unrelated generator.

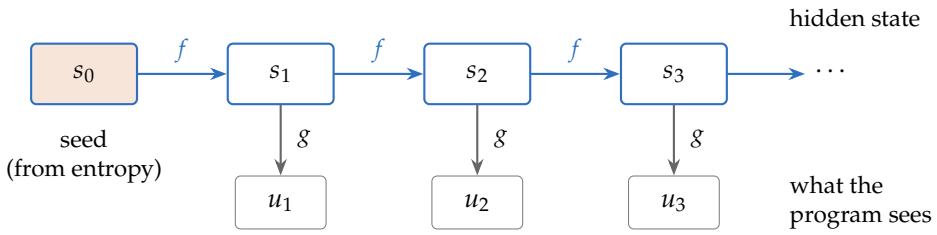


Figure 6.4: A pseudorandom number generator as a deterministic dynamical system. The state (top row) is updated by the same transition f at every step; the program sees only the outputs $g(s_n)$ (bottom row). The seed, the only input from outside, is the orange box.

A generator is thus a discrete dynamical system, like the Poincaré map of a chaotic pendulum. Its apparent randomness comes from the same place as chaos: nearby states are sent far apart, so after a few steps the output looks unrelated to the input [Lambert §12.7]. Kruschke's summary is exact: PRNGs "are not actually random; they are in fact deterministic. But the properties of the sequences they generate mimic the properties of random processes," and every sampling method relies heavily on how well they do so [Kruschke p. 74, fn. 4].

6.1.3 A finite machine must repeat

A generator has finitely many states, so it must eventually come back to a state it has visited, and from then on it repeats. Here is why. The state space S is finite: a generator with b bits of state has $|S| = 2^b$ states. Take the first $|S| + 1$ states $s_0, s_1, \ldots, s_{|S|}$. There are more of them than there are distinct states, so two must coincide (the **pigeonhole principle** [$n + 1$ letters in n pigeonholes put two letters in one hole]): $s_i = s_j$ for some $i < j \le |S|$. From there on the sequence repeats:

$$s_{j+k} \overset{(1)}{=} f^k(s_j) \overset{(2)}{=} f^k(s_i) \overset{(3)}{=} s_{i+k} \qquad \text{for every } k \ge 0. \tag{6.3}$$

(1) s_{j+k} is obtained from s_j by applying f k times; f^k means f composed k times.
(2) $s_j = s_i$, and a function gives the same output for the same input.
(3) The same statement read from s_i.

So the sequence of states, and with it the sequence of outputs, is **eventually periodic**: after a **tail** of i steps it runs round a cycle whose length P, the **period**, is at most $j - i \le |S|$. Can the tail be avoided? Yes, if f is a bijection [every state has exactly one predecessor]: applying f^{-1} to $s_i = s_j$ gives $s_{i-1} = s_{j-1}$, and i applications give $s_0 = s_{j-i}$, so the seed itself lies on the cycle. Good generators are designed to be bijections with a single cycle through (almost) all 2^b states.

Example: von Neumann's middle-square rule. The first proposal for a computer generator (von Neumann, in the 1940s) was to square a four-digit number and keep the middle four digits of the eight-digit square. Starting from 2916,

$$2916^2 = 08\,5030\,56 \;\to\; 5030, \qquad 5030^2 = 25\,3009\,00 \;\to\; 3009,$$
$$3009^2 = 09\,0540\,81 \;\to\; 0540, \qquad 0540^2 = 00\,2916\,00 \;\to\; 2916,$$

where the square is padded with leading zeros to eight digits. After four steps we are back at the seed: a period of 4 out of 10^4 possible states. This is not bad luck. Running the rule from all 10^4 seeds, every sequence repeats a value within 111 steps (43.7 on average), and 20% of the seeds end stuck at 0, since $0^2 = 0$. The middle-square map is not a bijection (the squares of x and of numbers close to x share middle digits), so many states funnel into a few short cycles. So a "chaotic-looking" rule is not enough; we need one whose cycles we can prove something about. Linear generators (Section 6.2) are such rules.

Example: what the period does to a Monte Carlo estimate. A generator that has gone round its cycle adds no new information, yet the error bar of a simulation keeps shrinking as if it did. To see this, estimate π from pairs (u_1, u_2) of uniform numbers. The pair falls inside the quarter disc $u_1^2 + u_2^2 < 1$ with probability $p = \pi/4$ (the area of a quarter of the unit disc), so $\hat{\pi} = 4\times$(fraction of pairs inside). The usual standard error of a fraction, $\sqrt{p(1-p)/N}$, gives the error bar $4\sqrt{\hat{p}(1-\hat{p})/N}$ that a script would report (it is derived with the error of a Monte Carlo average in Section 6.11.2); here $\hat{p}$ is the fraction of pairs inside and N the number of pairs. Now take the pairs from the tiny generator $x \mapsto (1229\,x + 1) \bmod 2048$ of Lambert's exercise [Lambert Problem 12.7], whose period is 2048 numbers, i.e. 1024 pairs. Figure 6.5 shows what happens. Up to about a thousand pairs the estimate behaves like any other. After one period, each new pair is one we have already seen, so the estimate stops improving: with 10^6 pairs it is 3.0899, off by -0.0517. The reported error bar knows nothing of this and keeps shrinking as $1/\sqrt{N}$, down to 0.0017, so the script reports an estimate that is wrong by thirty of its own standard errors. With PCG64 the same 10^6 pairs give 3.1389 ± 0.0016, within two standard errors of π, as it should be.

Watch out: The error bar cannot see the generator

A standard error $s/\sqrt{N}$ is computed under the assumption that the N draws are independent. If the generator repeats, or if its draws are correlated, the formula still returns a small number. Nothing in the output of a simulation warns us; only knowledge of the generator, or a rerun with an unrelated generator, does.

A modern generator never runs out of period: at 10^9 numbers per second, PCG64's period of $2^{128} \approx 3.4 \times 10^{38}$ would take 10^{22} years. What remains is requirement 1: making the numbers statistically indistinguishable from independent uniform draws in every way that could matter, and finding out when they are not.

6.2 Linear congruential generators

The generator that broke Monty's coin in Section 6.1.1 is a *linear congruential generator*, the oldest family still in use and, until the 1990s, the default in most scientific libraries. Its simple update, one multiplication and one addition per step, lets us derive the period, examine how quickly individual bits repeat, and locate successive outputs geometrically. These properties also matter for a more recent generator, PCG64, whose internal state evolves by a linear congruential rule.

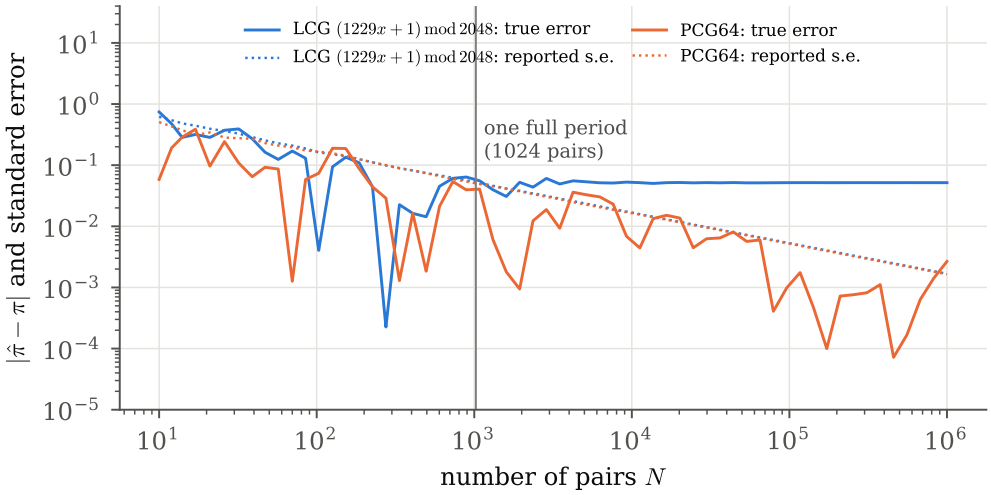


Figure 6.5: Estimating π from N pairs of uniform numbers. Solid: the actual error $|\hat{\pi} - \pi|$; dotted: the standard error the script reports. With PCG64 (orange) the actual error scatters around the reported one, both falling as $1/\sqrt{N}$. With a generator of period 2048 (blue) the actual error freezes at 0.05 once the sequence has gone round its cycle (vertical line), while the reported error keeps falling.

6.2.1 The rule and its closed form

Pick three integers, a **modulus** m, a **multiplier** a and an **increment** c, with $0 < a < m$ and $0 \le c < m$. From a seed x_0 in $\{0, 1, \ldots, m-1\}$ generate

$$x_{n+1} = (a\,x_n + c) \bmod m, \qquad u_n = \frac{x_n}{m}, \tag{6.4}$$

where $y \bmod m$ [the remainder of the integer y on division by m, always in $0, \ldots, m-1$] keeps the state in range [Cowan (3.2)–(3.3)], [Lambert Problem 12.7]. The state is one integer, the transition is $x \mapsto (ax + c) \bmod m$, and the output divides by m. With $c = 0$ the generator is called **multiplicative**; this is the form Cowan presents, and there x_n is never 0 (a zero state would stay zero), so $u_n = x_n/m$ lies strictly between 0 and 1 [Cowan §3.1]. Implementing it takes four lines:

Listing 6.2: A linear congruential generator (`code/ch06/lib_prng.py`)

```
23 def lcg_ints(a: int, c: int, m: int, seed: int, n: int) -> np.ndarray:
24     """Return x_1, ..., x_n of the LCG (the seed x_0 itself is not returned)."""
25     out = np.empty(n, dtype=np.int64 if m <= 2**62 else object)
26     x = seed % m
27     for i in range(n):
28         x = (a * x + c) % m
29         out[i] = x
30     return out
```

Python integers never overflow, so line 28 is exactly (6.4). In C or Fortran one chooses $m = 2^{32}$ or 2^{64}, and the reduction mod$m$ is then free: it is what the hardware does anyway when an unsigned product overflows.

Where is the state after n steps? Applying the rule n times answers this with a closed form that we will use repeatedly. For $a \neq 1$,

$$x_n \overset{(1)}{=} a\,x_{n-1} + c \overset{(2)}{=} a^2 x_{n-2} + ac + c \overset{(3)}{=} a^n x_0 + c\,(a^{n-1} + \cdots + a + 1)$$

$$\overset{(4)}{=} a^n x_0 + c\,\frac{a^n - 1}{a - 1} \pmod{m}. \tag{6.5}$$

(1) The rule (6.4); the notation " $\pmod m$" at the end means every equality in the chain holds up to multiples of m. Reducing after each step or once at the end gives the same remainder, because remainders of sums and products depend only on the remainders of the terms.
(2) Substitute the rule for x_{n-1}.
(3) Repeat n times; each step multiplies what is there by a and adds c.
(4) Geometric sum. The fraction $(a^n - 1)/(a - 1)$ is an integer (it equals the sum on the previous line), so this is an integer identity before we reduce.

In particular, for a multiplicative generator $x_n = a^n x_0 \bmod m$: the whole sequence is powers of a.

Example: Lambert's toy generator. Take $a = 2$, $c = 3$, $m = 10$ and the seed $x_0 = 5$ [Lambert Problem 12.7.2]:

$$5 \;\to\; (2 \cdot 5 + 3) \bmod 10 = 3 \;\to\; 9 \;\to\; (21 \bmod 10) = 1 \;\to\; 5 \;\to\; 3 \;\to \cdots$$

The sequence comes back to the seed after 4 steps, although there are 10 states: the period is 4. Figure 6.6 shows the transition of every state. The seed 7 is a fixed point ($2 \cdot 7 + 3 = 17 \to 7$), and the seeds $0, 2, 4, 6, 8$ each lead into the cycle or the fixed point after one step, so they have a tail. The rule is not a bijection: 3 has two predecessors, 0 and 5, because $2x$ and $2(x + 5)$ differ by 10. Whenever a shares a factor with m, two states that differ by $m/\gcd(a, m)$ are sent to the same place, and the tails of Eq. (6.3) appear.

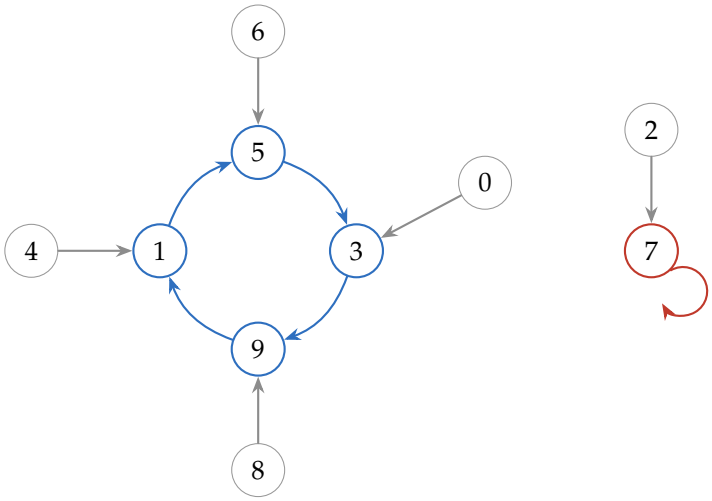


Figure 6.6: Every state of $x \mapsto (2x + 3) \bmod 10$ and where it goes. The seeds 1, 3, 5, 9 lie on a cycle of length 4 (blue); the seed 7 is a fixed point (red); the even seeds are not on any cycle and fall into one after a single step (grey). Because 2 and 10 share the factor 2, two states always share each image.

6.2.2 How long can the period be?

Multiplicative generators with a prime modulus. Let $c = 0$ and m prime, and start from any $x_0 \neq 0$. When does the sequence come back to its seed? By (6.5) $x_n = a^n x_0 \bmod m$, so it returns

exactly when

$$x_n = x_0 \overset{(1)}{\iff} a^n x_0 \equiv x_0 \pmod m \overset{(2)}{\iff} (a^n - 1)\, x_0 \equiv 0 \pmod m \overset{(3)}{\iff} a^n \equiv 1 \pmod m. \tag{6.6}$$

(1) The closed form (6.5) with $c = 0$; "$\equiv \pmod m$" means "differs by a multiple of m".
(2) Move x_0 to the left and factor.
(3) m is prime and does not divide x_0 (because $0 < x_0 < m$), so it must divide $a^n - 1$. A prime that divides a product divides one of the factors.

The period is therefore the smallest $n \geq 1$ with $a^n \equiv 1 \pmod m$, the **order** of a modulo m, and it is the same for every seed. How large can it be? Fermat's little theorem says $a^{m-1} \equiv 1 \pmod m$ for prime m. Call the order P and write $m - 1 = qP + r$ with $0 \leq r < P$. Then $a^r = a^{m-1}(a^P)^{-q} \equiv 1$, and since P is the smallest positive power that gives 1, r must be 0. So the order divides $m - 1$, and the period is at most $m - 1$ (all states except 0). It equals $m - 1$ exactly when a is a **primitive root** [its powers $a, a^2, \ldots, a^{m-1}$ run through all nonzero remainders].

Example: the two multipliers modulo 7. With $m = 7$ and $a = 3$, starting from 1: $1 \to 3 \to 9 \bmod 7 = 2 \to 6 \to 18 \bmod 7 = 4 \to 12 \bmod 7 = 5 \to 15 \bmod 7 = 1$. The period is $6 = m - 1$; 3 is a primitive root. With $a = 2$: $1 \to 2 \to 4 \to 8 \bmod 7 = 1$, period 3, because $2^3 = 8 \equiv 1$. From the seed 3 the same multiplier gives $3 \to 6 \to 5 \to 3$, a different cycle of the same length: the six nonzero states split into two cycles of three.

Cowan's example generator is of this kind: $m = 2147483399$ is prime and $a = 40692$ has period $m - 1 \approx 2 \times 10^9$ [Cowan §3.1], [150]. In the 1980s that was respectable. Today a single core produces 10^9 numbers per second, so the whole period is used up in about

$$\frac{2.1 \times 10^9 \text{ numbers}}{10^9 \text{ numbers/s}} \approx 2 \text{ s},$$

and a simulation that runs for an hour would go round the cycle about seventeen hundred times. L'Ecuyer's remedy in the same paper was to combine two such generators with different prime moduli, so that the period becomes of the order of the product of the two periods (about 2×10^{18}); this idea of a small cheap transition with a very large period runs through all modern designs. Cowan's example of a much longer period is the RANMAR generator, about 10^{43} [Cowan §3.1].

Full period with an increment. With $c \neq 0$ the state 0 is allowed, so all m states could in principle form one cycle. Which choices of a, c and m achieve this? A classical theorem gives the exact answer.

Key idea: Full period of $x \mapsto (ax + c) \bmod m$ (Hull and Dobell)

The generator has period m for every seed if and only if

1. c and m have no common factor, $\gcd(c, m) = 1$;
2. $a - 1$ is divisible by every prime that divides m;
3. $a - 1$ is divisible by 4 if m is divisible by 4.

Why the first condition is needed can be seen in three lines. Suppose $d = \gcd(c, m) > 1$ and start from $x_0 = 0$. By (6.5), $x_n = c\,(a^{n-1} + \cdots + 1) \bmod m$. Every term is a multiple of d, and reducing modulo m (itself a multiple of d) keeps it a multiple of d. So only the m/d multiples

of d are ever visited, and the period is at most $m/d < m$. The second and third conditions force $a \equiv 1$ modulo every prime that divides m (and modulo 4 when 4 divides m). The map then behaves enough like $x \mapsto x + c$, which steps through every remainder in turn, that it too visits every state before returning. Their proof is elementary number theory, and nothing later depends on it.

Example: a full cycle of sixteen. For $m = 16 = 2^4$ take $a = 5$, $c = 3$. Then $\gcd(3, 16) = 1$, and $a - 1 = 4$ is divisible by 2 (the only prime in 16) and by 4. From $x_0 = 0$ the sequence is

$$3,\ 2,\ 13,\ 4,\ 7,\ 6,\ 1,\ 8,\ 11,\ 10,\ 5,\ 12,\ 15,\ 14,\ 9,\ 0,$$

all sixteen states once each. Lambert's generator $x \mapsto (1229x + 1) \bmod 2048$ with $2048 = 2^{11}$, also satisfies the conditions ($c = 1$; $1228 = 4 \cdot 307$), and indeed its period is 2048. So does his second, $x \mapsto (1597x + 51749) \bmod 244944$ [Lambert Problem 12.7.6]: here $244944 = 2^4 \cdot 3^7 \cdot 7$, 51749 is divisible by none of $2, 3, 7$, and $1596 = 2^2 \cdot 3 \cdot 7 \cdot 19$ is divisible by 2, 3, 7 and 4, and the period comes out as 244944. Lambert's question "what is the maximum period?" [Lambert Problem 12.7.1] has the answer m, reached exactly under these three conditions.

6.2.3 The low bits of a power-of-two modulus

Look again at the sixteen numbers above and write down only whether each is odd or even: $1, 0, 1, 0, \ldots$. The lowest bit alternates. This is not an accident of the example. Suppose $m = 2^k$ and look at the lowest j bits of the state, i.e. at $x_n \bmod 2^j$ with $j \le k$. Then

$$x_{n+1} \bmod 2^j \overset{(1)}{=} \big((a x_n + c) \bmod 2^k\big) \bmod 2^j \overset{(2)}{=} (a x_n + c) \bmod 2^j \overset{(3)}{=} \big(a\,(x_n \bmod 2^j) + c\big) \bmod 2^j. \tag{6.7}$$

(1) The rule (6.4) with $m = 2^k$, then reduction modulo 2^j.
(2) 2^j divides 2^k, so removing multiples of 2^k does not change the remainder modulo 2^j.
(3) Replacing x_n by its remainder changes $a x_n$ by a multiple of 2^j.

The lowest j bits therefore form an LCG of their own, with modulus 2^j and no input from the higher bits. It has only 2^j states, so its period is at most 2^j: the lowest bit has period at most 2, the lowest two bits at most 4, and so on. For $j = 1$, a odd and c odd (as the full-period conditions require), $x_{n+1} \equiv x_n + 1 \pmod 2$: the parity flips at every step.

Example: the bits of a 32-bit generator. Figure 6.7 shows the 32 bits of the first 64 outputs of $x \mapsto (1664525x + 1013904223) \bmod 2^{32}$, the generator of the Monty example. The measured periods of bits 0, 1, 2, 3 are 2, 4, 8, 16, and bit 11 has period 4096: bit j has period exactly 2^{j+1}. The Monty simulation of Section 6.1.1 drew two integers per game and took the coin from the second, $x \bmod 2$, which is bit 0 of every second output. Since bit 0 alternates, every second output has the same parity, and the coin never changes. The cure is never to use the low bits of such a generator. Take the *high* bits instead, e.g. $\lfloor 2u_n \rfloor$ with $u_n = x_n/2^{32}$, whose period is the full 2^{32}.

Watch out: The remainder trick

Writing `x % n` to get an integer in $0, \ldots, n-1$ uses the *lowest* bits of x. For a power-of-two linear congruential generator those are the worst bits it has. It also has a bias whenever n does not divide the number of possible values of x. Library functions such as `rng.integers(n)` avoid both problems.

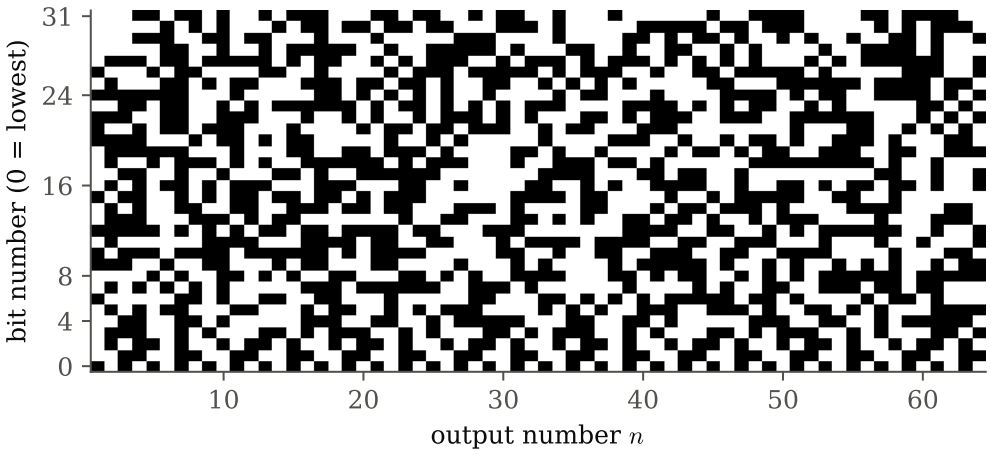


Figure 6.7: The bits of the first 64 outputs of the 32-bit linear congruential generator used in the Monty example (black = 1). The lowest rows are periodic with periods 2, 4, 8, . . . ; only the upper rows look random. The lowest row alternates, which is why taking x mod 2 from every second output gave Monty a coin that never changes.

A full period says nothing about how the numbers are arranged within it. The tiny generator $x \mapsto (1229x + 1) \bmod 2048$ visits every one of its 2048 states, yet the left panel of Figure 6.8, which plots each output against the next, shows that consecutive pairs lie on six parallel lines (one of them very short) [Lambert Problem 12.7.5]. Lambert's larger generator and RANDU look fine in the same plot. The lines come from a lattice that every linear congruential generator carries (Section 6.3); for RANDU that lattice shows up in three dimensions.

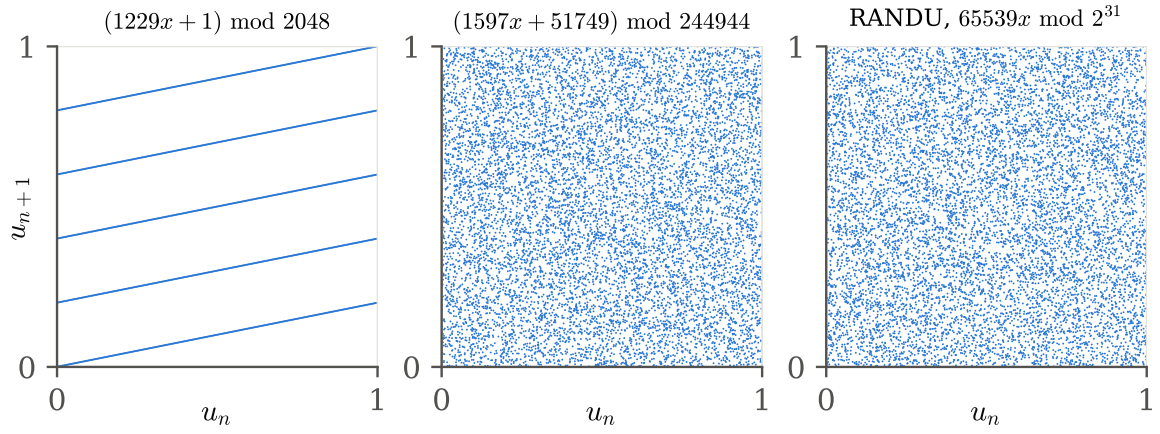


Figure 6.8: Consecutive pairs (u_n, u_{n+1}) from three linear congruential generators. Left: all 2048 pairs of $x \mapsto (1229x + 1) \bmod 2048$ lie on a few parallel lines. Middle and right: 10^4 pairs from Lambert's second generator and from RANDU, which show no structure at this resolution.

6.3 Random numbers fall mainly in the planes

Many simulations use the uniform numbers in groups. A point in a detector volume needs three coordinates; a photon's direction needs two angles and its energy a third number; a rejection-sampling step (taken up later in the chapter) needs a candidate and a test value. Group consecutive outputs of a linear congruential generator into t-tuples $(u_n, u_{n+1}, \ldots, u_{n+t-1})$ and plot them as points in the unit cube $[0, 1)^t$. For independent uniform numbers the points would

fill the cube evenly. In 1968 Marsaglia showed that for *every* multiplicative linear congruential generator they do something else: they lie on a family of parallel hyperplanes, and for some generators on very few of them [151]. The six lines in the left panel of Figure 6.8 were the two-dimensional case. So three questions follow. Where exactly do the t-tuples lie? How many planes are there? And how do we measure how bad that is?

6.3.1 Why the points form a lattice

The points are far from independent because a whole t-tuple is fixed by one number, the state at its start. Take a multiplicative generator, $x_{n+1} = a x_n \bmod m$ (Eq. (6.4) with $c = 0$), and its outputs $u_n = x_n/m$. By the closed form (6.5), $x_{n+j} = a^j x_n \bmod m$, i.e. $a^j x_n$ minus some integer multiple $k_j m$ of m. Dividing by m, the t-tuple that starts at x_n is

$$\boldsymbol{u} \equiv (u_n, u_{n+1}, \ldots, u_{n+t-1}) \overset{(1)}{=} \frac{x_n}{m}\,(1, a, a^2, \ldots, a^{t-1}) - (0, k_1, k_2, \ldots, k_{t-1}). \tag{6.8}$$

(1) $u_{n+j} = x_{n+j}/m = (a^j x_n - k_j m)/m = a^j x_n/m - k_j$ for $j = 1, \ldots, t-1$, and $u_n = x_n/m$.

Every tuple is therefore an integer multiple of the single vector $\boldsymbol{v} = (1, a, \ldots, a^{t-1})/m$ plus a vector of integers. The set of all such combinations $\{x\,\boldsymbol{v} + \boldsymbol{z} : \; x \in \mathbb{Z},\; \boldsymbol{z} \in \mathbb{Z}^t\}$ is a **lattice** [a regular, crystal-like grid of points: all integer combinations of a few fixed vectors]. The generator's tuples are the points of this lattice that fall in the unit cube (at most $m-1$ of them, one for each state). In crystal language: the points of a generator are the atoms of a crystal, and a crystal is organised in lattice planes.

Where do the planes come from? A family of parallel planes is described by a normal vector $\boldsymbol{h}$: the planes are $\boldsymbol{h}\cdot\boldsymbol{u} = k$ for $k = 0, \pm 1, \pm 2, \ldots$. So we ask which integer vectors $\boldsymbol{h} = (h_1, \ldots, h_t)$ make $\boldsymbol{h}\cdot\boldsymbol{u}$ an integer for every tuple. The answer is every $\boldsymbol{h}$ with

$$h_1 + a\,h_2 + a^2 h_3 + \cdots + a^{t-1} h_t \equiv 0 \pmod{m}. \tag{6.9}$$

To check it, compute $\boldsymbol{h}\cdot\boldsymbol{u}$ for any tuple:

$$\boldsymbol{h}\cdot\boldsymbol{u} \overset{(1)}{=} \frac{x_n}{m}\left(h_1 + a h_2 + \cdots + a^{t-1} h_t\right) - \sum_{j \ge 1} h_{j+1} k_j \overset{(2)}{=} x_n\, q - \sum_{j \ge 1} h_{j+1} k_j \;\in\; \mathbb{Z}, \tag{6.10}$$

(1) Dot $\boldsymbol{h}$ into (6.8).
(2) By (6.9) the bracket is $q\,m$ for some integer q; what is left is a sum of products of integers.

so every tuple lies on one of the parallel hyperplanes $\boldsymbol{h}\cdot\boldsymbol{u} = k$, $k \in \mathbb{Z}$. The integer vectors that satisfy (6.9) form a lattice of their own, the **dual lattice** [the integer vectors $\boldsymbol{h}$ with $\boldsymbol{h}\cdot\boldsymbol{u}$ an integer for every point of the generator], and each nonzero $\boldsymbol{h}$ in it defines one family of planes. Two facts about such a family follow from elementary geometry.

- *Spacing.* The planes $\boldsymbol{h}\cdot\boldsymbol{u} = k$ and $\boldsymbol{h}\cdot\boldsymbol{u} = k+1$ are a distance $1/|\boldsymbol{h}|$ apart, with $|\boldsymbol{h}| = \sqrt{h_1^2 + \cdots + h_t^2}$. (A plane $\boldsymbol{h}\cdot\boldsymbol{u} = k$ lies at distance $k/|\boldsymbol{h}|$ from the origin, along the unit normal $\boldsymbol{h}/|\boldsymbol{h}|$.)
- *Number.* Inside the unit cube, $\boldsymbol{h}\cdot\boldsymbol{u}$ lies strictly between $\sum_{h_i<0} h_i$ and $\sum_{h_i>0} h_i$, an open interval of length $|h_1| + \cdots + |h_t|$. It therefore contains at most $|h_1| + \cdots + |h_t|$ integers, and the points lie on at most that many planes. (When the endpoints are integers that the points never reach, as for a multiplicative generator, the interval contains one fewer, $|h_1| + \cdots + |h_t| - 1$.)

So a *short* dual vector is bad news: it means few planes, far apart, with nothing in between. For an increment $c \neq 0$ the tuples are the same lattice shifted by a constant vector, so the planes are the same, merely displaced.

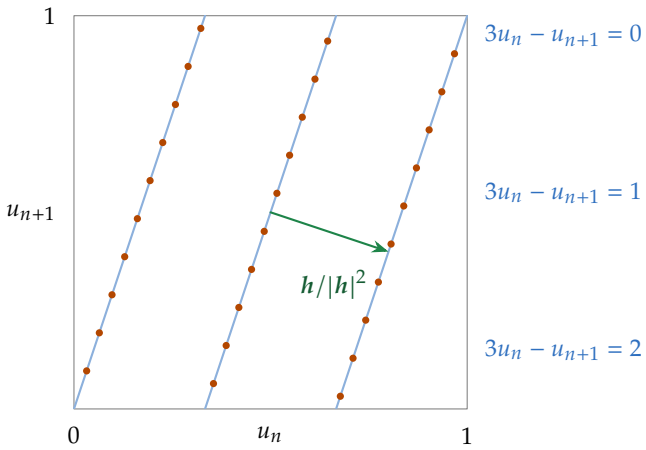


Figure 6.9: All 30 pairs (u_n, u_{n+1}) of the toy generator $x \mapsto 3x \bmod 31$ (orange dots). The dual vector $\boldsymbol{h} = (3, -1)$ satisfies $3 - 3 \cdot 1 \equiv 0$, so every pair has $3u_n - u_{n+1}$ equal to 0, 1 or 2 and lies on one of three parallel lines (blue). The green arrow joins two neighbouring lines along their common normal; its length is the gap $1/|\boldsymbol{h}| = 1/\sqrt{10}$.

Example: the toy generator modulo 31. For $x \mapsto 3x \bmod 31$ in two dimensions, condition (6.9) reads $h_1 + 3h_2 \equiv 0 \pmod{31}$. The vector $\boldsymbol{h} = (3, -1)$ satisfies it ($3 - 3 = 0$), so every pair has $3u_n - u_{n+1} \in \mathbb{Z}$. Inside the square $3u_n - u_{n+1}$ lies between -1 and 3, so it is 0, 1 or 2 (the value -1 would need $u_{n+1} = 1$ and $u_n = 0$, which cannot occur): three lines, a distance $1/\sqrt{3^2 + 1^2} = 0.32$ apart (Figure 6.9). No shorter dual vector exists: the only other candidates with $|\boldsymbol{h}| \le \sqrt{10}$ are $(\pm 1, 0)$, $(0, \pm 1)$, $(\pm 1, \pm 1)$, $(\pm 2, \pm 1)$, $(\pm 1, \pm 2)$ and so on, and none gives $h_1 + 3h_2$ divisible by 31.

Example: Lambert's generator. For $x \mapsto (1229x + 1) \bmod 2048$ the vector $\boldsymbol{h} = (1, -5)$ works: $1 - 5 \cdot 1229 = -6144 = -3 \cdot 2048$. So $u_n - 5u_{n+1}$ takes only integer values (shifted by the constant $-5/2048$ that the increment contributes), at most $|1| + |-5| = 6$ of them, and the pairs lie on six lines a distance $1/\sqrt{26} = 0.20$ apart. That is the left panel of Figure 6.8. A good generator with 2048 states could have spread its pairs over lines a few hundredths apart.

6.3.2 RANDU's fifteen planes

RANDU was the generator supplied with IBM's System/360 scientific library in the 1960s, and it was copied into many other systems. It is multiplicative, with

$$x_{n+1} = 65539\, x_n \bmod 2^{31}, \qquad 65539 = 2^{16} + 3,$$

seeded with an odd x_0. The multiplier was chosen because multiplying by $2^{16} + 3$ is a shift and two additions, fast on the hardware of the time. That choice is its downfall. To see why, ask how three consecutive outputs are related. Two steps multiply the state by a^2, $x_{n+2} = a^2 x_n \bmod 2^{31}$, so we need a^2 modulo 2^{31}:

$$\begin{aligned} a^2 &\overset{(1)}{=} (2^{16} + 3)^2 \overset{(2)}{=} 2^{32} + 6 \cdot 2^{16} + 9 \overset{(3)}{=} 6 \cdot 2^{16} + 9 \pmod{2^{31}} \\ &\overset{(4)}{=} 6\,(2^{16} + 3) - 18 + 9 \overset{(5)}{=} 6a - 9 \pmod{2^{31}}. \end{aligned} \tag{6.11}$$

(1) $a = 65539 = 2^{16} + 3$.
(2) $(p + q)^2 = p^2 + 2pq + q^2$ with $p = 2^{16}$, $q = 3$.
(3) $2^{32} = 2 \cdot 2^{31}$ is a multiple of the modulus.
(4) Write $6 \cdot 2^{16} = 6\,(2^{16} + 3) - 18$.

(5) Collect: $-18 + 9 = -9$, and $2^{16} + 3 = a$.

So a^2 is a small combination of a and 1. Multiplying (6.11) by x_n gives $x_{n+2} = a^2 x_n \equiv 6a x_n - 9x_n \equiv 6x_{n+1} - 9x_n$, that is, every output is fixed by the two before it:

$$x_{n+2} - 6x_{n+1} + 9x_n \equiv 0 \pmod{2^{31}} \qquad \Longrightarrow \qquad 9u_n - 6u_{n+1} + u_{n+2} \in \mathbb{Z}. \tag{6.12}$$

This is (6.10) with the dual vector $\boldsymbol{h} = (9, -6, 1)$. Inside the cube $9u_n - 6u_{n+1} + u_{n+2}$ lies strictly between -6 and 10, so it is one of the fifteen integers $-5, -4, \ldots, 9$. *All* triples of RANDU lie on fifteen parallel planes, a distance $1/\sqrt{81 + 36 + 1} = 1/\sqrt{118} \approx 0.092$ apart. A small ball of diameter just under 0.09 can sit between two neighbouring planes and contain no point at all. Knuth's verdict, quoted ever since, was that the very name RANDU is enough to bring dismay into the eyes and stomachs of many computer scientists.

The script `04_randu_planes.py` generates 30000 consecutive triples from the seed 1, computes $9u_n - 6u_{n+1} + u_{n+2}$ for each, and checks that it is an exact integer; it takes exactly 15 values, from -5 to 9. The planes are invisible from a generic viewpoint, but looking along them (viewing direction $(1, 1.5, 0)$, which is perpendicular to the normal $(9, -6, 1)$) they appear edge-on as fifteen lines (Figure 6.10). This is Marsaglia's classic picture [151].

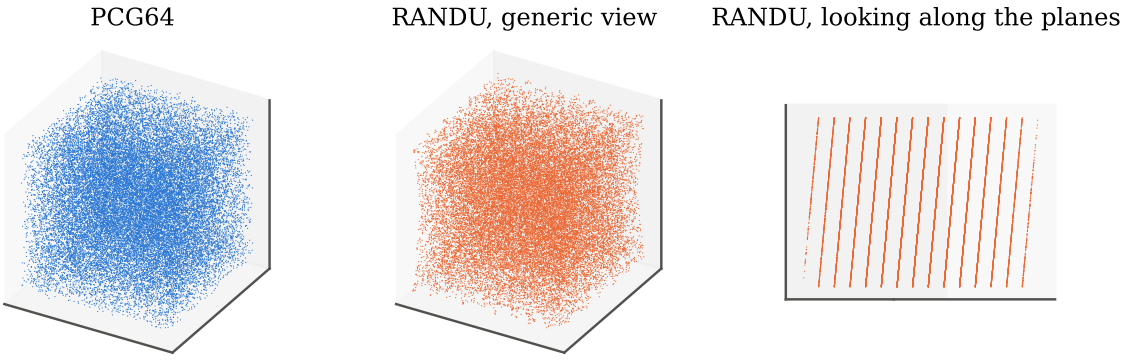


Figure 6.10: Thirty thousand consecutive triples (u_n, u_{n+1}, u_{n+2}) in the unit cube. Left: from PCG64. Middle: from RANDU, seen from the same generic angle; nothing looks wrong. Right: RANDU seen along the direction $(1, 1.5, 0)$, which lies in its planes $9u_n - 6u_{n+1} + u_{n+2} = k$; the planes appear edge-on, fifteen of them.

What does this do to a physics simulation? Anything that uses three consecutive numbers as a point in space samples only fifteen slices of the space. A detector acceptance computed by throwing points into a cube, a random walk in three dimensions, or a three-body phase-space generator would all be biased, and the bias does not shrink with N: it is a property of the generator, not a statistical fluctuation.

6.3.3 The spectral test

How do we measure how bad a generator's lattice is in t dimensions? By the widest gap between its planes, because a wide gap is a slab of the cube that no tuple ever visits. The gap of a family is $1/|\boldsymbol{h}|$, so the widest gap belongs to the shortest nonzero dual vector. Its length

$$\nu_t = \min\bigl\{|\boldsymbol{h}| : \ \boldsymbol{h} \in \mathbb{Z}^t, \ \boldsymbol{h} \neq \boldsymbol{0}, \ h_1 + a h_2 + \cdots + a^{t-1} h_t \equiv 0 \pmod{m}\bigr\}, \qquad d_t = \frac{1}{\nu_t}, \tag{6.13}$$

gives the largest distance d_t between adjacent hyperplanes that hold all the t-tuples. Computing ν_t for $t = 2, \ldots, 6$ or 8 is the **spectral test** (the name comes from its origin in Fourier analysis: $\boldsymbol{h}$

is a wave vector in which the empirical density of the points has a spike, the lattice Bragg peak of the crystal). A small d_t means that even the most widely spaced family of planes is dense, so the points come close to filling the cube.

Working definition 6.2 (spectral test)

For a linear congruential generator with multiplier a and modulus m, and a dimension t: find the shortest nonzero integer vector $\boldsymbol{h}$ with $h_1 + a h_2 + \cdots + a^{t-1} h_t \equiv 0 \pmod m$. Its length is ν_t; the t-tuples lie on hyperplanes $\boldsymbol{h} \cdot \boldsymbol{u} \in \mathbb{Z}$ a distance $d_t = 1/\nu_t$ apart. *Operational test:* compare d_t with the smallest gap any lattice with m points per unit volume could have (Eq. (6.15)); a generator whose d_t is many times larger fails.

Finding the shortest vector of a lattice is a classic problem, and in low dimension it is easy: the script `05_spectral_test.py` writes down a basis of the dual lattice, namely $(m, 0, \ldots, 0)$ and $(-a^{j-1} \bmod m)\,\boldsymbol{e}_1 + \boldsymbol{e}_j$ for $j = 2, \ldots, t$ (any solution of (6.9) is an integer combination of these), reduces it with the Lenstra–Lenstra–Lovász (LLL) algorithm [repeated Gram–Schmidt-like steps that make the basis vectors short and nearly orthogonal], and then searches small integer combinations of the reduced vectors for the shortest one. For RANDU in $t = 3$ it finds $\boldsymbol{h} = (9, -6, 1)$ with $\nu_3 = 10.86$, confirming (6.12). The same calculation for three classic generators is plotted in Figure 6.11.

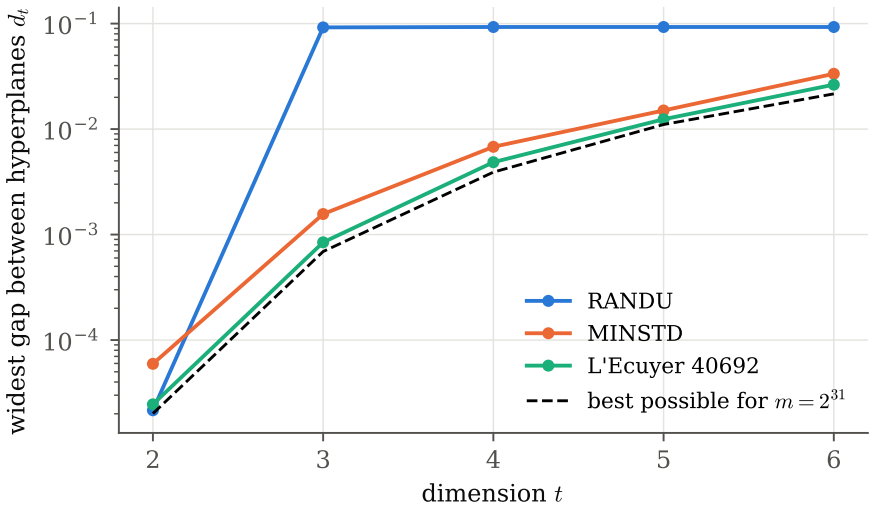


Figure 6.11: The spectral test for three multiplicative generators with $m \approx 2^{31}$: the widest gap d_t between the hyperplanes that hold the t-tuples. The dashed line is the smallest gap any lattice with 2^{31} points per unit volume can have. RANDU is near-ideal in two dimensions and then jumps to $d_3 \approx 0.09$, where it stays. MINSTD ($a = 16807$, $m = 2^{31} - 1$) and L'Ecuyer's $a = 40692$ stay within a factor of about three of the ideal.

How small can the gap be? Even the best multiplier leaves gaps of a certain size, and Minkowski's theorem tells us how large. A lattice with one point per state has m points per unit volume: the volume of its unit cell is $1/m$, and its dual lattice has unit-cell volume m (the reciprocal lattice of a crystal has the inverse cell volume, the same fact from solid-state physics). A lattice cannot have arbitrarily long shortest vectors at fixed cell volume. **Minkowski's theorem** [a convex region, symmetric about the origin, with volume larger than 2^t times the cell volume must contain a nonzero lattice point] makes this precise. Apply it to the **cross-polytope**

$\{\boldsymbol{h} : |h_1| + \cdots + |h_t| \le r\}$, the t-dimensional octahedron:

$$\operatorname{vol}\{|h_1| + \cdots + |h_t| \le r\} \overset{(1)}{=} 2^t \frac{r^t}{t!} \overset{(2)}{\ge} 2^t\, m \quad \Longleftrightarrow \quad r \ge (t!\, m)^{1/t}. \tag{6.14}$$

(1) The octahedron is 2^t copies (one per sign pattern) of the simplex $\{h_i \ge 0,\ \sum h_i \le r\}$, whose volume is $r^t/t!$ (for $t = 2$, a triangle of area $r^2/2$; for $t = 3$, a tetrahedron of volume $r^3/6$).
(2) Minkowski's condition, with the dual cell volume m.

So for $r = (t!\, m)^{1/t}$ the octahedron contains a nonzero dual vector $\boldsymbol{h}$ with $|h_1| + \cdots + |h_t| \le r$, and by the "number" bullet above, the t-tuples lie on at most $(t!\, m)^{1/t}$ hyperplanes. This is Marsaglia's theorem [151]: *every* multiplicative generator with modulus m puts its t-tuples on at most $(t!\, m)^{1/t}$ parallel hyperplanes. For $m = 2^{32}$ the bound is 2954 planes in three dimensions and only 42 in ten. No choice of multiplier can avoid it; a good choice only avoids being much worse. The sharpest statement of the same kind uses the Hermite constants γ_t of lattice theory [the largest possible value of (shortest vector)^2/(cell volume)$^{2/t}$ over all lattices in t dimensions] and gives the smallest possible gap,

$$d_t \ \ge\ \frac{1}{\gamma_t^{1/2}\, m^{1/t}}, \qquad \gamma_2 = \left(\tfrac{4}{3}\right)^{1/2},\ \gamma_3 = 2^{1/3},\ \gamma_4 = 2^{1/2},\ \ldots, \tag{6.15}$$

the dashed line in Figure 6.11: for $m = 2^{31}$ and $t = 3$ it is 6.905×10^{-4}, a hundred times smaller than RANDU's gap.

Example: two generators that pass. MINSTD, $x \mapsto 16807\, x \bmod (2^{31} - 1)$, the "minimal standard" of the 1990s, has $d_2 = 5.950 \times 10^{-5}$ and $d_3 = 0.001565$, and L'Ecuyer's $a = 40692$ from Cowan's example has $d_2 = 2.457 \times 10^{-5}$ [Cowan §3.1]. Both stay close to (6.15) in every dimension plotted. But in high dimension even the best possible gap grows, because $m^{1/t}$ shrinks as t grows: at $t = 6$ MINSTD has $d_6 = 0.03343$, and the six-tuples of any 2^{31}-state generator lie on at most $(6! \cdot 2^{31})^{1/6} \approx 108$ planes. A simulation that consumes numbers in groups of six or more (and most do) sees this structure. The modulus has to be much larger, which is exactly what the 128-bit linear congruential generator inside PCG64 provides (Section 6.4); tables of multipliers with good lattice structure, found by running the spectral test over many candidates, are given in [152].

Watch out: Two-dimensional plots certify nothing

RANDU's pair plot (Figure 6.8, right) and its d_2 are close to ideal. Its defect appears only in three dimensions. A generator has to be examined in every dimension in which the application uses it.

6.4 Modern generators: a bigger state and a better output function

The linear congruential generators of Sections 6.2 and 6.3 fail in two ways. Their state is small, so the period is short and, by Marsaglia's bound, the tuples lie on few hyperplanes. And their output is the state itself (divided by m), so every regularity of the state, the short periods of the low bits and the lattice, shows up directly in the numbers. Modern generators repair both. They keep the picture of Figure 6.4 (a state s, a transition f, an output g) but choose a large state with a transition that is cheap and provably has a huge period, and an output function that hides whatever structure the transition leaves behind. For three widely used generators we ask the same three questions: what is the state, how is it updated, and what does the output function do to it? The three are xorshift, the Mersenne Twister (the default of Python's `random` module and of numpy's older `RandomState`) and PCG64 (numpy's default today).

6.4.1 Xorshift: a linear map on bits

Marsaglia's xorshift generators [153] keep a single 64-bit word x and update it with three operations of the form "shift x by a few bits and combine the result with x by **exclusive or**" [bitwise addition without carry, $0 \oplus 0 = 1 \oplus 1 = 0$, $0 \oplus 1 = 1 \oplus 0 = 1$]:

Listing 6.3: Xorshift64 with shifts 13, 7, 17 (`code/ch06/lib_prng.py`)

```
88 def xorshift64(seed: int, n: int) -> list[int]:
89     x, out = seed & M64, []
90     assert x != 0, "the all-zero state maps to itself"
91     for _ in range(n):
92         x ^= (x << 13) & M64
93         x ^= x >> 7
94         x ^= (x << 17) & M64
95         out.append(x)
96     return out
```

Line 92 shifts x left by 13 bits (dropping bits that fall off the top) and XORs it into x; line 93 does the same with a right shift by 7; line 94 with a left shift by 17.

Example: one step by hand. Start from $x = 1$, a single bit at position 0. Line 92: $1 \ll 13 = 2^{13} = 8192$, and $1 \oplus 8192 = 8193$ (bits 0 and 13). Line 93: $8193 \gg 7 = 64$ (bit 13 moves to bit 6, bit 0 falls off), and $8193 \oplus 64 = 8257$ (bits $0, 6, 13$). Line 94: $8257 \ll 17$ has bits $17, 23, 30$, none of which overlaps bits $0, 6, 13$, so the XOR is a sum: $8257 + 8257 \cdot 2^{17} = 1082269761$ (bits $0, 6, 13, 17, 23, 30$). One bit has become six; after a few more steps every bit of the state depends on many bits of the seed.

Why the period is $2^{64} - 1$. How can we know the period without stepping through 10^{19} states? Each of the three lines is linear in the bits, so one step is a matrix, and the period is the smallest power of that matrix that gives the identity. To see the matrix, write the state as a column vector $\boldsymbol{x} \in \{0,1\}^{64}$ of its bits and do arithmetic modulo 2, where XOR is addition. A left shift by k is a matrix L^k (L has ones just below the diagonal), a right shift is R^k, and the three lines of the update are the matrices $\mathbb{1} + L^{13}$, $\mathbb{1} + R^7$, $\mathbb{1} + L^{17}$. So one step is a fixed 64×64 matrix over the two-element field,

$$\boldsymbol{x}_{n+1} = T\,\boldsymbol{x}_n, \qquad T = (\mathbb{1} + L^{17})(\mathbb{1} + R^7)(\mathbb{1} + L^{13}) \pmod 2, \tag{6.16}$$

and the n-th state is $T^n \boldsymbol{x}_0$. The zero state is fixed ($T\mathbf{0} = \mathbf{0}$), so the largest possible period is $2^{64} - 1$, a single cycle through all nonzero states. That happens exactly when the smallest $n \geq 1$ with $T^n = \mathbb{1}$ is $N = 2^{64} - 1$. The script `06_modern_generators.py` builds T column by column (column j is the image of the state with only bit j set), checks $T\boldsymbol{x} = \text{xorshift}(\boldsymbol{x})$ for random $\boldsymbol{x}$, and then checks the period with eight matrix powers (T^N and the seven powers $T^{N/p}$). The logic is

$$\begin{aligned} T^N = \mathbb{1} \text{ and } T^{N/p} \neq \mathbb{1} \text{ for every prime } p \mid N &\overset{(1)}{\Longrightarrow} \text{order of } T \text{ divides } N \text{ but no } N/p \\ &\overset{(2)}{\Longrightarrow} \text{order of } T = N. \end{aligned} \tag{6.17}$$

(1) If $T^n = \mathbb{1}$ and $T^N = \mathbb{1}$, then $T^{\gcd(n,N)} = \mathbb{1}$, so the smallest such n (the order) divides N.

(2) A proper divisor d of N divides N/p for at least one prime p in the factorisation of N, and then $T^{N/p} = (T^d)^{N/(pd)} = \mathbb{1}$, which was excluded.

Here $2^{64} - 1 = 3 \cdot 5 \cdot 17 \cdot 257 \cdot 641 \cdot 65537 \cdot 6700417$, and each matrix power takes 64 squarings by repeated squaring. Both conditions hold, so xorshift64 with shifts $(13, 7, 17)$ has period $2^{64} - 1$: no search through 10^{19} states was needed.

The linearity that proves the period is also xorshift's weakness. Since $T^n(x_0 \oplus y_0) = T^n x_0 \oplus T^n y_0$, the stream from the seed $s_1 \oplus s_2$ is the XOR of the streams from s_1 and s_2; the script checks this for 50 outputs. Any statistic that is linear over bits, such as the rank of a matrix built from output bits or the length of the shortest linear recurrence that reproduces a bit sequence, detects the generator. Modern variants (xorshift128+, xoshiro) add a nonlinear output step, an integer addition or multiplication, for this reason.

6.4.2 The Mersenne Twister

The Mersenne Twister MT19937 [154] is a much larger linear generator over bits. Its state is 624 words of 32 bits. The transition ("twist") replaces each word by the XOR of the word 397 places ahead with a bit-mixed combination of the top bit of the word itself and the lower 31 bits of its neighbour:

Listing 6.4: MT19937: twist (transition) and tempering (output) (`code/ch06/lib_prng.py`)

```
112     def _twist(self):
113         mt = self.mt
114         for k in range(624):
115             y = (mt[k] & 0x80000000) | (mt[(k + 1) % 624] & 0x7FFFFFFF)
116             mt[k] = mt[(k + 397) % 624] ^ (y >> 1) ^ (0x9908B0DF if y & 1 else 0)
117         self.i = 0
118 
119     def next32(self) -> int:
120         if self.i >= 624:
121             self._twist()
122         y = self.mt[self.i]
123         self.i += 1
124         y ^= y >> 11                        # tempering: an invertible scramble
125         y ^= (y << 7) & 0x9D2C5680
126         y ^= (y << 15) & 0xEFC60000
127         y ^= y >> 18
128         return y
```

Lines 115–116 are the recurrence $x_{k+624} = x_{k+397} \oplus (x_k^{\text{upper}} \mid x_{k+1}^{\text{lower}})A$, where the multiplication by the matrix A is "shift right by one, and XOR the constant `0x9908B0DF` if the bit shifted out was 1". Only $32 \cdot 624 - 31 = 19937$ bits of the state take part (the lower 31 bits of one word are never used), and the constants are chosen so that the transition, a 19937×19937 matrix over the bits, has order $2^{19937} - 1$. That number is a **Mersenne prime** [a prime of the form $2^p - 1$], which gives the generator its name and makes the period check (6.17) a single condition, since N has no prime factor other than itself. The period, about 10^{6001}, can never be exhausted.

The output function, **tempering** (lines 124–127), applies four more shift-and-XOR steps to the word before returning it. Each step is invertible, so it does not change the period; it redistributes bits so that the top bits of the outputs, which a program uses first, are well mixed. The designers proved that MT19937 is **623-dimensionally equidistributed** to 32-bit accuracy [over a full period, every sequence of 623 consecutive 32-bit outputs occurs equally often, except that the all-zero sequence occurs once less]. No linear congruential generator comes close to this in high dimension. The script reproduces numpy's MT19937 output exactly: 10,000 of the first 10,000 outputs agree, starting from the conventional seed 5489.

Two weaknesses remain, and both come from linearity. First, the generator is *predictable*. Tempering is invertible, so each output reveals one full word of the state; 624 consecutive outputs reveal all of it. The script takes 624 outputs from the middle of numpy's stream, undoes the tempering (the `untemper` function, a few lines), and predicts the next 1000 outputs: 1,000 of them are right. For physics this does not matter, since a simulation is not an adversary; for cryptography it disqualifies the generator. Second, it fails statistical tests aimed at linearity.

In the TestU01 battery (Section 6.6) MT19937 passes everything except two linear-complexity tests [155], which measure how long a linear recurrence is needed to reproduce a sequence of output bits; for MT19937 it is 19937, far shorter than for a random sequence of the same length. It is also slow to recover from a state with very few one-bits, and its 2.5 kilobytes of state are large for a GPU thread.

6.4.3 PCG64: a linear congruential generator with a permuted output

O'Neill's insight was that the linear congruential generator is a fine transition function if the modulus is large, and that its problems all come from using the state as the output [156]. PCG ("permuted congruential generator") keeps a 128-bit linear congruential state and passes it through a small nonlinear permutation. This is PCG64, the default bit generator of numpy's `default_rng` and of `rng_for`, written out in full:

Listing 6.5: PCG64: 128-bit LCG state, XSL-RR output (`code/ch06/lib_prng.py`)

```
141 class PCG64:
142     """Bit-for-bit the same as numpy.random.PCG64 when given its (state, inc)."""
143
144     def __init__(self, state: int, inc: int):
145         self.state, self.inc = state & M128, inc & M128   # inc is odd
146
147     def next64(self) -> int:
148         self.state = (self.state * PCG_MULT + self.inc) & M128     # transition: an LCG mod
    2^128
149         s = self.state
150         return rotr64(((s >> 64) ^ s) & M64, s >> 122)           # output: xor-fold, then
    rotate
```

Line 148 is the transition: $s \mapsto (as + c) \bmod 2^{128}$, with the 128-bit multiplier `PCG_MULT` and an odd increment c (`inc`). By the Hull–Dobell conditions of Section 6.2 (c odd, $a - 1$ divisible by 4) the period is 2^{128} for every increment, and different increments give different sequences, so the increment selects a **stream**. Line 150 is the output function **XSL-RR** ("xor-shift low, random rotation", Figure 6.12), in two steps:

1. *Fold:* $(s \gg 64) \oplus s$, keeping the low 64 bits, XORs the high half of the state onto the low half. Each output bit i now combines state bit i (a poor, low bit) with state bit $i + 64$ (a good, high bit).
2. *Rotate:* rotate the 64-bit result right by $r = s \gg 122$, the value of the top six state bits, a number between 0 and 63.

The rotation is what hides the bad bits. The top six bits are the best bits of the state (by (6.7) their period is the full 2^{128}), and they choose, at every step, which bit of the folded word ends up in which position of the output. A fixed output position therefore sees a different state bit from step to step, and the short periods of the low bits (Figure 6.7) and the lattice of Section 6.3 are scrambled away. The multiplication and addition in line 148 are cheap (a few machine instructions on 64-bit hardware), and the state is only 256 bits including the increment.

The class `PCG64` is not an imitation: given numpy's internal state (`like_numpy`, lines 152–155 of the library), it produces numpy's bits. Of the first 100,000 outputs of `numpy.random.PCG64`, 100,000 agree exactly. Every random number in this book has come out of the ten lines above.

PCG64 passes all the tests of TestU01's largest battery [156]. One caveat was found later: two PCG64 streams whose increments and starting states are related in a particular way produce correlated outputs, because the fold-and-rotate output mixes too little for such pairs. Seeding through numpy's SeedSequence (Section 6.5) chooses state and increment as unrelated hashes

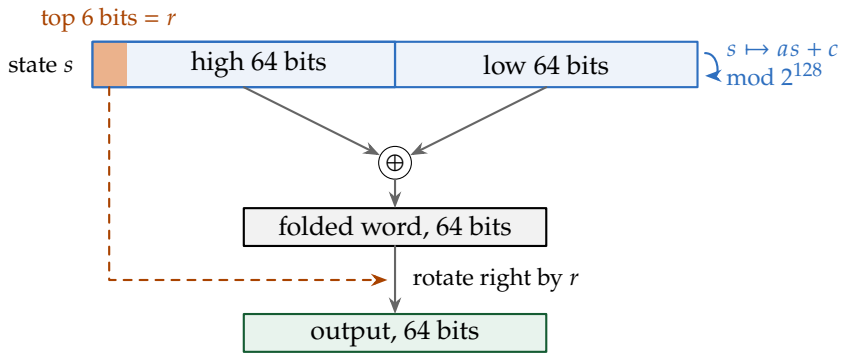


Figure 6.12: The output function of PCG64. The 128-bit state is updated by a linear congruential step (blue loop). The output XORs the high half of the state onto the low half, then rotates the 64-bit result by an amount r read from the top six state bits (orange), the bits with the longest period. The rotation moves every output bit around from step to step, so the short periods of the low state bits never reach a fixed output position.

and avoids it; for simulations with enormous numbers of parallel streams numpy also offers `PCG64DXSM`, the same state with a stronger, multiplicative output function.

Example: what the permutation does, at toy size. The effect is easiest to see on a generator small enough to plot completely. Take a 16-bit state with the classic multiplier $a = 25173$ and increment $c = 13849$ (full period 2^{16} by Hull–Dobell), and produce an 8-bit output in two ways: by keeping the top 8 bits of the state (a truncated LCG, the usual practice before PCG), or by PCG's 16 → 8 bit output XSH-RR, which XORs the state with itself shifted right by 5, keeps bits 5 to 12, and rotates them by the value of the top 3 bits. Figure 6.13 plots each output against the next for the whole period. The truncated LCG shows rows of the lattice; the permuted output does not.

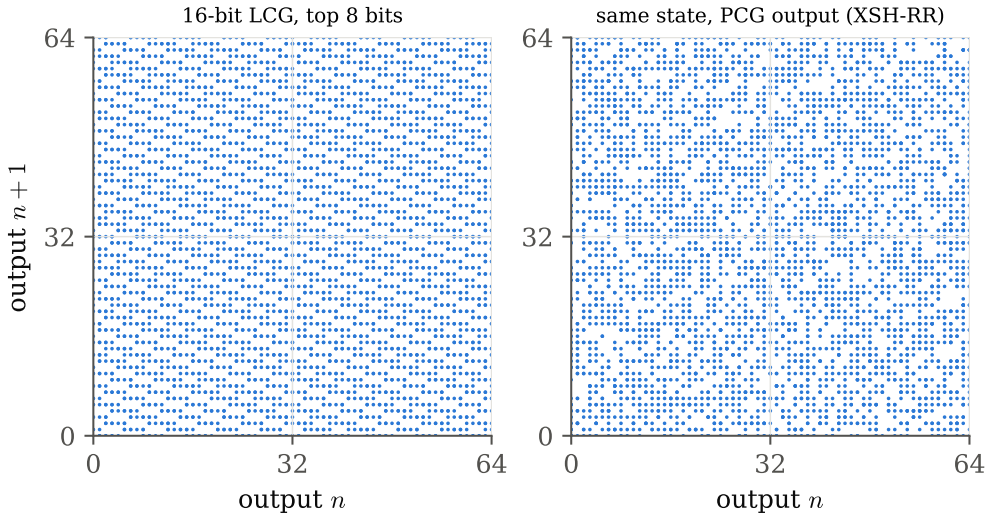


Figure 6.13: A 16-bit linear congruential state turned into 8-bit outputs in two ways; each dot is a pair of consecutive outputs (a corner of the full 256 × 256 plot is shown). Left: keeping the top 8 bits; the pairs fall on regular rows, the lattice of Section 6.3. Right: PCG's xorshift-and-rotate output applied to the same state; the regular rows are gone.

Counting makes the same point. There are $K = 256^2 = 65536$ possible pairs of 8-bit outputs, and one period gives $n = 65536$ consecutive pairs. How many distinct pairs would a truly

random sequence of that length produce? Take one particular pair. A random sequence hits it at least once with probability

$$1-\left(1-\frac{1}{K}\right)^{n} \overset{(1)}{=} 1-\exp\left[n\ln\left(1-\frac{1}{K}\right)\right] \overset{(2)}{=} 1-\exp\left[-\frac{n}{K}+O\left(\frac{n}{K^2}\right)\right] \overset{(3)}{=} 1-e^{-1}+O(K^{-1}) \approx 0.632. \tag{6.18}$$

(1) Each of the n pairs misses the given pair with probability $1-1/K$, independently; $y^n = e^{n\ln y}$.
(2) $\ln(1-\epsilon) = -\epsilon - \epsilon^2/2 - \dots$ with $\epsilon = 1/K$.
(3) $n = K$.

Multiplying by the K possible pairs, about $K(1-e^{-1}) \approx 41{,}428$ distinct pairs are expected, with a standard deviation of about 80 (the variance of the number of empty cells is approximately $Ke^{-1}(1-2e^{-1})$). The permuted output produces 41,488 distinct pairs (63.3%), as expected. The truncated LCG produces 40,960 (62.5%), about six standard deviations too few: its pairs are too regular to spread as randomly as random pairs do.

6.4.4 Counter-based generators, and a summary

A different design removes the transition function altogether. A **counter-based generator** uses the state $s_n = n$, a plain counter, and puts all the work into the output: $u_n = g_k(n)$, where g_k is a bijective scramble of 64 or 128 bits controlled by a key k, built from a few rounds of a simplified block cipher [157]. Jumping ahead by any amount is free (just set n), and every key gives an independent stream, which suits GPUs where each of thousands of threads needs its own numbers. Numpy provides one of them as `Philox`.

Taxonomy: Generators and what is inside them

generator	state	transition / output	period	known weakness
RANDU	31 bits	$65539x \bmod 2^{31}$ / $x/2^{31}$	2^{29}	triples on 15 planes
MINSTD	31 bits	$16807x \bmod (2^{31}-1)$ / x/m	$2^{31}-2$	short period, lattice for $t \gtrsim 5$
xorshift64	64 bits	three shift-XORs / the state	$2^{64}-1$	linear over bits
MT19937	19937 bits	linear twist / tempering	$2^{19937}-1$	linear over bits, predictable
PCG64	128 + 128 bits	LCG mod 2^{128} / XSL-RR	2^{128} per stream	streams with specially related seeds can be correlated
Philox	counter + key	$n \mapsto n+1$ / keyed scramble	2^{256}	none known for simulation

Every modern design splits the work the same way. The transition only needs to guarantee a long period, which a linear rule, over integers or over bits, does with a proof. The output function supplies the statistical quality, by mixing the good and bad parts of the state nonlinearly. Whether the result is good enough is decided not by the design but by statistical tests (Section 6.6).

6.5 Seeds, streams and reproducible parallel simulations

A realistic Monte Carlo study is not one run but thousands. To estimate the covariance of a power-spectrum estimator in Chapter 8 we will simulate a thousand skies; a detector study runs

the same event simulation on hundreds of cluster nodes at once. Each job needs its own stream of random numbers. Three things must be true of these streams. Each must be *reproducible*: if job 417 produced a strange map, we must be able to regenerate exactly that map to find out why. They must be *different*: two jobs with the same stream would produce the same sky twice and silently halve our sample. And they must be *unrelated*: no simple relation between the numbers of different jobs, or the thousand skies are not a thousand independent draws. So the question is: how do we seed many streams so that all three hold? The helper `rng_for`, which makes every random stream used here, is one answer, and we follow it down to the bits.

6.5.1 Why "seed = job number" can go wrong

The obvious recipe is to seed job j with the integer j. It is safe only if the generator forgets where it started. Nearby seeds are nearby states, and some transitions keep nearby states related forever.

Example: RANDU from the seeds 1 and 3. RANDU (Section 6.3) is multiplicative: by (6.5), $x_n = a^n x_0 \bmod 2^{31}$. Start one stream from $x_0 = 1$ and another from $x'_0 = 3$. Then

$$x'_n \overset{(1)}{=} a^n \cdot 3 \bmod 2^{31} \overset{(2)}{=} 3\,x_n \bmod 2^{31}, \qquad \text{so} \qquad u'_n \overset{(3)}{=} 3u_n \bmod 1. \tag{6.19}$$

(1) The closed form with $c = 0$ and seed 3.
(2) $a^n \cdot 3 = 3 \cdot a^n$, and $a^n \equiv x_n$ for the stream from the seed 1; reducing before or after multiplying by 3 gives the same remainder.
(3) Divide by 2^{31}: $u'_n = x'_n / 2^{31}$ is the fractional part of $3u_n$.

So the second stream is a fixed function of the first. Each stream on its own is as good (or as bad) as RANDU, but plotted against each other they lie on three lines (Figure 6.14, left); the script confirms $u'_n = 3u_n \bmod 1$ to the last bit. For any multiplicative generator, the seeds x_0 and kx_0 give streams related by multiplication by k.

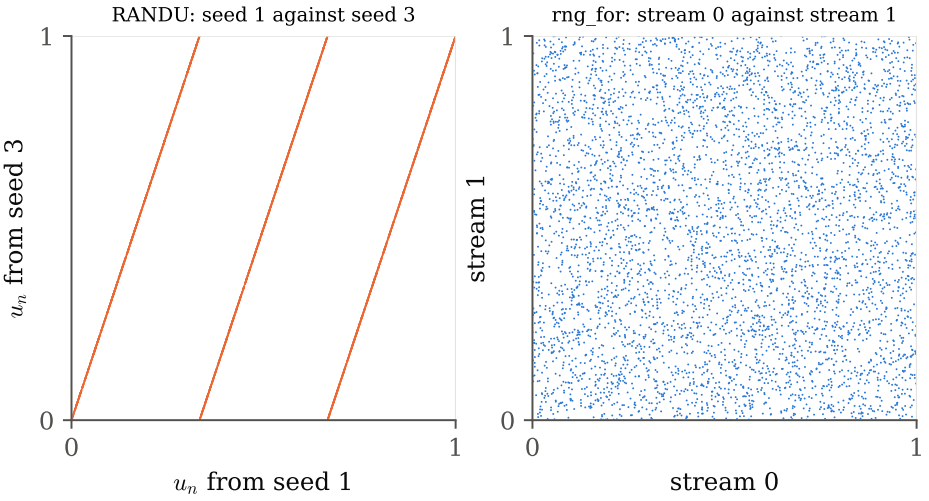


Figure 6.14: Two streams plotted against each other, 5000 numbers each. Left: RANDU started from the seeds 1 and 3; the second stream is the fractional part of three times the first, so the points lie on three lines. Right: streams 0 and 1 of `rng_for` for the same script; no relation is visible.

Example: random 32-bit seeds collide. The opposite recipe, a random seed per job, has its own trap if the seed is small. Many older interfaces accept a 32-bit integer, so there are only

$K = 2^{32} \approx 4.3 \times 10^9$ seeds. How likely is it that two of n jobs draw the same one? This is the birthday problem, with K seeds in place of 365 days. The chance that n independently chosen seeds are all different is

$$\mathrm{P}(\text{all } n \text{ seeds different}) \overset{(1)}{=} \prod_{k=0}^{n-1}\Big(1-\frac{k}{K}\Big) \overset{(2)}{=} \exp\Big[\sum_{k=0}^{n-1}\ln\Big(1-\frac{k}{K}\Big)\Big] \overset{(3)}{\approx} \exp\Big[-\sum_{k=0}^{n-1}\frac{k}{K}\Big] \overset{(4)}{=} \exp\Big[-\frac{n(n-1)}{2K}\Big]. \tag{6.20}$$

(1) The $(k+1)$-th seed must avoid the k values already taken, which has probability $1 - k/K$, independently of the earlier draws.
(2) A product is the exponential of the sum of the logarithms.
(3) $\ln(1-\epsilon) \approx -\epsilon$ for $\epsilon = k/K \ll 1$.
(4) $\sum_{k=0}^{n-1} k = n(n-1)/2$.

With $n = 10^3$ jobs a collision has probability 0.0001, with 10^4 jobs 0.012, and with 10^5 jobs 0.688: more likely than not, two of a hundred thousand simulations are identical. The exact product and the approximation agree to three digits. With 128-bit seeds and the same 10^5 jobs the probability is 1.469×10^{-29}. Seeds must be long.

6.5.2 SeedSequence: hashing a seed into a state

Numpy solves both problems by never putting a user's seed directly into the generator's state. A **SeedSequence** takes any collection of nonnegative integers as **entropy** [here, the numbers that identify the job: a base seed and a stream index] and mixes them with a hash function into a pool of 128 bits. Every input bit affects every pool bit, so the entropies $[s, 0]$ and $[s, 1]$, which differ in one bit, give pools that look unrelated. From the pool it then generates as many state words as the generator needs, again through a hash. The two problems of the examples disappear: nearby inputs no longer give nearby states, and the state is drawn from 2^{128} or more possibilities. The method `spawn(n)` makes n child sequences by appending the child's index to the entropy, so that a job can in turn seed its own sub-jobs, and the resulting tree of streams is reproducible from the single root.

The helper `rng_for` of `code/common.py` is three lines:

Listing 6.6: Every random stream in the book is made here (`code/common.py`)

```
122 def rng_for(chapter: str, name: str, stream: int = 0) -> np.random.Generator:
123     """Reproducible PCG64 generator seeded from the script identity (+ optional stream
        index)."""
124     key = f"{chapter}/{name}" + (f"/salt{SEED_SALT}" if SEED_SALT else "")
125     seed = int.from_bytes(hashlib.sha256(key.encode()).digest()[:8], "little")
126     return np.random.Generator(np.random.PCG64(np.random.SeedSequence([seed, stream])))
```

Line 124 builds the script's name, e.g. `"ch06/07_seeds_streams"` (the optional salt is the testing switch of Section 0.6; in the book's own runs it is empty). Line 125 turns that name into a 64-bit integer: the first eight bytes of its SHA-256 hash. For this script it is 16002383964210800950. The name, not a number typed by hand, is the seed, so two scripts can never share a stream by accident, and renaming a script is the only way to change its random numbers. Line 126 feeds [seed, stream] into a SeedSequence and hands the result to PCG64. The script `07_seeds_streams.py` follows the path down to the bits (Figure 6.15):

Listing 6.7: rng_for traced by hand: SeedSequence words to the PCG64 state (`code/ch06/07_seeds_streams.py`)

```
36 ss = np.random.SeedSequence([seed, 0])
37 w = ss.generate_state(4, np.uint64)                    # four 64-bit words of mixed entropy
```

```
38 initstate = (int(w[0]) << 64) | int(w[1])
39 initseq = (int(w[2]) << 64) | int(w[3])
40 inc = ((initseq << 1) | 1) & M128                          # the increment must be odd
41 st = 0
42 st = (st * 0x2360ED051FC65DA44385DF649FCCF645 + inc) & M128   # PCG's seeding: step, add,
       step
43 st = (st + initstate) & M128
44 st = (st * 0x2360ED051FC65DA44385DF649FCCF645 + inc) & M128
```

Line 37 asks the SeedSequence for four 64-bit words. The first two become a 128-bit starting value (line 38), the last two the stream selector (line 39), from which line 40 makes the odd increment $c = 2 \times \text{selector} + 1$ of the linear congruential step of Section 6.4. Lines 41–44 are PCG's seeding rule: one step from zero, add the starting value, one more step. The state and increment obtained this way are identical to the ones inside numpy's generator, and so are the next 1000 outputs of the `PCG64` class of Section 6.4. So nothing in the random numbers is hidden: every operation from the name of a script to its first output is now explicit.

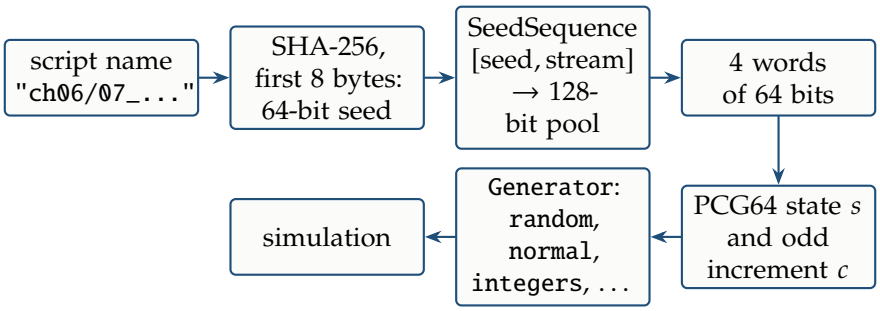


Figure 6.15: What `rng_for` does. The script's name is hashed to a 64-bit seed; the seed and a stream index are mixed by a SeedSequence into a 128-bit pool, which generates four 64-bit words; these fix the 128-bit state and the odd increment of PCG64. The `Generator` object then turns the 64-bit outputs into numbers of any distribution.

Example: eight streams of one script. `rng_for("ch06", "07_seeds_streams", stream=k)` for $k = 0, \ldots, 7$ gives eight streams. From 10^6 numbers of each, the script computes all 28 correlation coefficients r between pairs of streams. Under independence each r is approximately Gaussian with mean 0 and standard deviation $1/\sqrt{N} = 0.0010$ (derived in Section 6.6), so the 28 values of $r\sqrt{N}$ should look like 28 draws from $\mathcal{N}(0,1)$. Their standard deviation is 1.10, the largest $|r|$ is 0.0032, and a Kolmogorov–Smirnov test of $\mathcal{N}(0,1)$ (Chapter 4) gives $p = 0.28$: no sign of any relation. The right panel of Figure 6.14 shows two of the streams against each other.

How likely is it that two streams overlap? Suppose a generator has a single cycle of length P and each of n jobs starts at an independent random point of it and consumes L numbers. Two jobs overlap if their starting points are within L of each other. Fix where the first job starts; the second overlaps it if it starts in a window of $2L$ positions (L before and L after), out of P. So one pair overlaps with probability about $2L/P$. There are $n(n-1)/2$ pairs, so $\text{P(any overlap)} \lesssim n^2 L/P$. For $n = 10^4$ jobs of $L = 10^{12}$ numbers each (a large simulation campaign) and $P = 2^{128} \approx 3.4 \times 10^{38}$, this is $10^{20}/3.4 \times 10^{38} \approx 3 \times 10^{-19}$. With PCG64 the streams also have different increments and are therefore different sequences altogether. Numpy additionally offers `jumped()`, which advances a PCG64 state by about $(\phi - 1)\,2^{128}$ steps (ϕ the golden ratio) as if that many numbers had been drawn, for the rare case in which non-overlapping blocks of one single stream are wanted.

Key idea: Seeding many simulations

Seed once, with a long seed (from physical entropy, or from a hash of a name). Derive every stream from it by a SeedSequence (a stream index, or `spawn`), never by adding job numbers to a seed. Record the root seed next to the results: it is all that is needed to regenerate any simulation exactly.

6.5.3 From 64 bits to a floating-point number

The generator produces integers; a simulation wants $u \in [0, 1)$. A double-precision number has a 53-bit significand, so numpy keeps the top 53 of the 64 output bits (the best ones) and scales:

$$u = \left\lfloor x/2^{11} \right\rfloor \cdot 2^{-53} = \frac{k}{2^{53}}, \qquad k \in \{0, 1, \ldots, 2^{53} - 1\} \text{ equally likely.} \tag{6.21}$$

The script checks that numpy's `random()` is exactly this, and the function `to_double` of the library implements it. The values form a grid of spacing $2^{-53} \approx 1.1 \times 10^{-16}$. The value $u = 1$ never occurs. The value $u = 0$ occurs, with probability 2^{-53}: rare, but at 10^9 draws per second it happens about once every

$$\frac{2^{53}}{10^9 \text{ s}^{-1}} \approx 9 \times 10^6 \text{ s} \approx 104 \text{ days}$$

of computing, which a large campaign easily exceeds. Cowan's generator excluded both endpoints [Cowan (3.3)]; numpy's includes 0.

Watch out: $\ln u$ with $u = 0$

The recipe $x = -\tau \ln u$ for exponential decay times (inverse-transform sampling, Section 6.8) returns $+\infty$ when $u = 0$. Use $x = -\tau \ln(1 - u)$, whose argument lies in $(0, 1]$, or the library's `rng.exponential`, which handles this.

6.6 Testing a generator: randomness as a null hypothesis

A dice manufacturer who wants to know whether its dice are fair does not inspect the dice; it rolls them many times and checks that the results behave like those of ideal, unbiased dice [Lambert §12.2]. A generator is tested the same way. We cannot prove that a deterministic sequence is random, since it is not; what we can do is state the property we need as a hypothesis and look for evidence against it. This is the logic of Chapter 4, applied to a stream of numbers.

- The null hypothesis H_0: the numbers $u_1, \ldots, u_N$ are independent draws from the uniform distribution (6.2). As in Chapter 4, we take this as the null because it is the case whose distribution we know: for truly random uniform numbers we can compute the distribution of any statistic built from them.
- A test statistic $T(u_1, \ldots, u_N)$ that is sensitive to one particular way in which H_0 could fail: unequal frequencies, correlation between neighbours, too many or too few ups and downs, structure in triples.
- The p-value: how often truly random numbers would give a value of T as far out as the one we got, or further in the same direction.

A generator "fails" a test when the p-value is tiny, and keeps being tiny when the test is repeated with fresh numbers. So two questions follow. Which statistics catch which defects? And why must a good generator pass all of them, while no finite set of them can prove it good?

6.6.1 Four tests, each derived

1. Equal frequencies: the χ^2 test of a histogram. The simplest defect is that some values come up more often than others. To look for it, divide $[0,1)$ into k equal bins and count the numbers O_j that fall into bin j. Under H_0 each number lands in bin j with probability $1/k$, independently, so the expected count is $E = N/k$ and the counts are multinomial. The statistic

$$X^2 = \sum_{j=1}^{k} \frac{(O_j - E)^2}{E} \tag{6.22}$$

is small when the counts are close to their expectation. What value should we expect under H_0? Its mean follows in one line:

$$\mathbb{E}[X^2] \overset{(1)}{=} \sum_{j=1}^{k} \frac{\operatorname{Var}(O_j)}{E} \overset{(2)}{=} k\,\frac{N\frac{1}{k}\left(1-\frac{1}{k}\right)}{N/k} \overset{(3)}{=} k-1. \tag{6.23}$$

(1) $\mathbb{E}[(O_j - E)^2] = \operatorname{Var}(O_j)$ because $E = \mathbb{E}[O_j]$.
(2) Each O_j is binomial with N trials and success probability $1/k$, so $\operatorname{Var}(O_j) = N \cdot \frac{1}{k}(1-\frac{1}{k})$ (Chapter 2); all k terms are equal.
(3) Cancel N/k.

The mean is $k-1$, not k, because the counts obey one constraint, $\sum_j O_j = N$: only $k-1$ of them are free. For large E each $(O_j - E)/\sqrt{E}$ is nearly Gaussian, and X^2 is then distributed as χ^2 with $k-1$ degrees of freedom, the goodness-of-fit result of Chapter 4. We use $k = 64$ bins, so H_0 predicts $X^2 \approx 63 \pm 11$ (the χ^2_ν distribution has variance 2ν).

2. Neighbours: the serial correlation. A generator can have perfect frequencies and still let each number influence the next. The lag-one **serial correlation** is the correlation coefficient of the pairs (u_i, u_{i+1}),

$$r = \frac{\sum_{i=1}^{N-1}(u_i - \bar{u})(u_{i+1} - \bar{u})}{\sum_{i=1}^{N}(u_i - \bar{u})^2},$$

with $\bar{u}$ the mean of the u_i. Under H_0 the true correlation is 0, so the question is how far from 0 the measured r strays by chance alone. Its spread is easy to find if we replace $\bar{u}$ by its expectation $\frac{1}{2}$ (the error this makes is of relative order $1/N$). Write $v_i = u_i - \frac{1}{2}$, which has mean 0 and variance $\frac{1}{12}$ (the variance of a uniform variable on an interval of length 1). Then for the numerator $S = \sum_i v_i v_{i+1}$,

$$\begin{aligned}\operatorname{Var}(S) &\overset{(1)}{=} \sum_i \mathbb{E}[v_i^2 v_{i+1}^2] + 2\sum_i \mathbb{E}[v_i v_{i+1}\, v_{i+1} v_{i+2}] + (\text{terms with no shared index}) \\ &\overset{(2)}{=} (N-1)\,\mathbb{E}[v^2]^2 + 2\sum_i \mathbb{E}[v_i]\,\mathbb{E}[v_{i+1}^2]\,\mathbb{E}[v_{i+2}] + 0 \overset{(3)}{=} \frac{N-1}{144},\end{aligned} \tag{6.24}$$

(1) $\mathbb{E}[S] = \sum_i \mathbb{E}[v_i]\mathbb{E}[v_{i+1}] = 0$, so $\operatorname{Var}(S) = \mathbb{E}[S^2]$. Expanding the square gives products of two terms $v_i v_{i+1}$ and $v_j v_{j+1}$: the diagonal ($j = i$), neighbours that share one index ($j = i \pm 1$), and pairs with no index in common.
(2) Independence factorises every expectation. The diagonal gives $\mathbb{E}[v_i^2]\,\mathbb{E}[v_{i+1}^2] = (\frac{1}{12})^2$. A shared-index term contains the unshared v_i and v_{i+2} once each, and $\mathbb{E}[v] = 0$; a term with no shared index contains four factors of which at least one appears once, so it vanishes too.
(3) $(\frac{1}{12})^2 = \frac{1}{144}$.

The denominator is close to its mean $N/12$, so $\operatorname{Var}(r) \approx \frac{N/144}{(N/12)^2} = \frac{1}{N}$. The statistic $z = r\sqrt{N}$ is therefore approximately standard normal under H_0 (the central limit theorem applied to S), and $|z| > 3$ has probability 0.3%.

3. Ups and downs: the runs test. Write + when a number is larger than the one before it and − when it is smaller. A **run** is a maximal block of equal signs (Figure 6.16). A sequence that oscillates too regularly has too many runs; one that drifts has too few. How many runs should independent numbers give? Count the sign changes: the number of runs is $R = 1+$ (number of sign changes), and a sign change happens at u_i exactly when u_i is the largest or the smallest of the three numbers u_{i-1}, u_i, u_{i+1}. Under H_0 the six orderings of three independent continuous numbers are equally likely, and in four of them the middle one is the largest or the smallest. So

$$\mathbb{E}[R] \overset{(1)}{=} 1 + \sum_{i=2}^{N-1} \mathrm{P}(u_i \text{ is a turning point}) \overset{(2)}{=} 1 + (N-2)\,\frac{4}{6} \overset{(3)}{=} \frac{2N-1}{3}. \tag{6.25}$$

(1) R is 1 plus a sum of indicator variables, one for each interior position, and the expectation of an indicator is its probability.
(2) Four of the six equally likely orderings.
(3) $1 + \frac{2N-4}{3} = \frac{2N-1}{3}$.

The variance, $(16N - 29)/90$, comes from the same counting done for pairs of positions (neighbouring turning points are not independent, which is why it is not simply $N \cdot \frac{2}{3} \cdot \frac{1}{3}$). Then $z = (R - \mathbb{E}[R])/\sqrt{\operatorname{Var} R}$ is approximately standard normal under H_0.

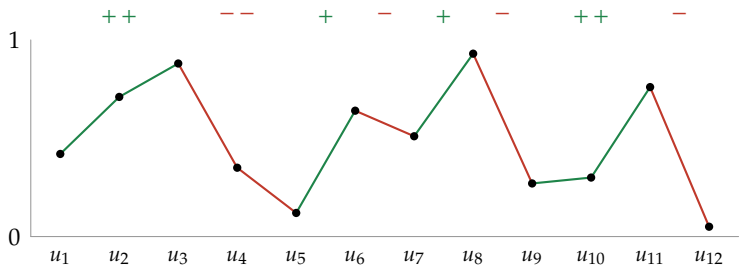


Figure 6.16: Runs up and down in twelve numbers. Rising steps are green, falling steps red. The signs of the steps form eight runs (++, −−, +, −, +, −, ++, −), against $\mathbb{E}[R] = (2 \cdot 12 - 1)/3 \approx 7.7$ for independent uniform numbers.

4. Triples: the χ^2 test in three dimensions. A generator can pass every test on single numbers and on neighbouring pairs and still fail in three dimensions, as RANDU does (Section 6.3). To look there, group the numbers into non-overlapping triples, treat each as a point in the unit cube, divide the cube into $20^3 = 8000$ cells and apply (6.22) to the cell counts: X^2 should be 7999 ± 126. A generator whose triples lie on a few planes leaves whole cells empty, and X^2 becomes huge.

6.6.2 Four generators under the four tests

The script `08_randomness_tests.py` applies the four tests to $N = 999999$ numbers from each of four generators: PCG64, MT19937, RANDU and the small generator $x \mapsto (1229x + 1) \bmod 2048$ of Section 6.2. Table 6.1 lists the statistics and the p-values (two-sided for the normal statistics, the upper tail for X^2).

	histogram X^2 ($\nu = 63$)	serial z	runs z	triples X^2 ($\nu = 7999$)
PCG64	69.93 ($p = 0.256$)	−0.69 (0.493)	1.84 (0.066)	7748 (0.977)
MT19937	69.53 (0.267)	−0.40 (0.693)	0.16 (0.873)	8144 (0.126)
RANDU	46.66 (0.939)	−0.49 (0.624)	1.02 (0.306)	2.473×10^5 ($< 10^{-300}$)
LCG, $m = 2048$	0.02259 (1.000)	197.09 ($< 10^{-300}$)	−156.69 ($< 10^{-300}$)	4.870×10^6 ($< 10^{-300}$)

Table 6.1: Four tests of H_0 (independent and uniform) on 999999 numbers from each generator. Entries are the statistic and, in brackets, its p-value.

Reading the table. PCG64 and MT19937 pass everything: every statistic is where H_0 puts it, and the p-values are unremarkable numbers between 0 and 1. RANDU passes the first three tests just as well. Its frequencies are uniform, its neighbours uncorrelated, its runs normal, which is why it was used for years. The triple test destroys it: $X^2 = 2.473 \times 10^5$ where 7999 ± 126 was expected, because its triples occupy only fifteen planes of the cube (Figure 6.10). The small generator fails in a way that deserves a second look. Its serial correlation is $z = 197.09$, i.e. $r \approx 0.2$: its pairs lie on the six lines of slope $\frac{1}{5}$ in Figure 6.8, so u_{n+1} tends to rise with u_n with slope about $\frac{1}{5}$. Its runs and triples fail as badly. And its histogram is *too good*: $X^2 = 0.02259$ where 63 ± 11 was expected. Over each full period the generator visits every one of its 2048 states exactly once, so each of the 64 bins gets exactly 32 numbers per period; the counts have none of the multinomial scatter that independent numbers would have. The upper-tail p-value is 1, but the lower tail tells the story: under H_0 an X^2 this small has probability 9.9×10^{-97}. A fit that is too good is as much evidence against H_0 as one that is too bad.

The p-value is itself a random variable, and under H_0 it is uniformly distributed on $[0, 1]$. The reason is that the p-value is the distribution function of the statistic evaluated at the statistic, and a continuous random variable pushed through its own distribution function is uniform (the fact behind inverse-transform sampling, Section 6.8.1). So a good generator gives p-values spread evenly over $[0, 1]$, and a p-value of 0.01 appears once in a hundred tests. The left panel of Figure 6.17 checks this for 400 histogram tests on fresh batches of 10^4 PCG64 numbers; a Kolmogorov–Smirnov test of their uniformity gives $p = 0.36$. The right panel shows where the triple statistics of the table fall relative to the χ^2_{7999} density.

Watch out: One small p-value is not a failure

Run a hundred tests on a perfect generator and about one will give $p < 0.01$. A defect shows itself as a p-value that is small again and again with fresh numbers, or that becomes smaller as N grows (for a real defect the statistic grows with N). Conversely, passing a test says only that this particular defect, at this sample size, is absent.

6.6.3 Batteries of tests

Serious testing uses hundreds of statistics, each designed around a way in which real generators have failed: gaps between repeated values (birthday spacings), matrix ranks and linear complexity of the output bits (the tests that catch xorshift and the Mersenne Twister, Section 6.4), random walks, the collision counts of many balls in many cells, and so on. The standard collection is TestU01 [155], organised in three batteries of increasing size and running time: SmallCrush (minutes), Crush and BigCrush (hours, more than a hundred test statistics on about 2^{38} numbers). Its authors ran it on the generators in common use in 2007: the power-of-two linear congruential generators and RANDU fail the smaller batteries, MT19937 passes BigCrush except for the two linear-complexity tests, and several generators then shipping in popular

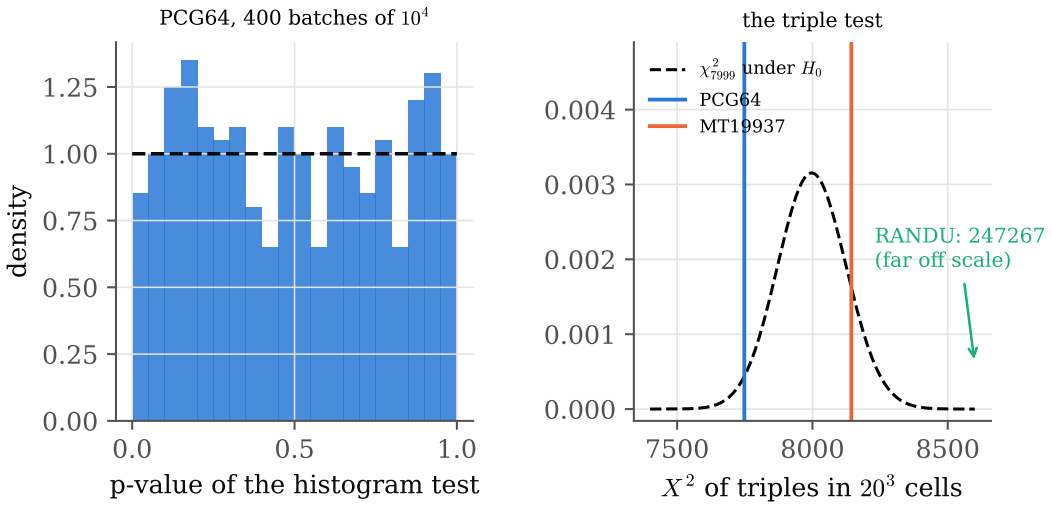


Figure 6.17: Left: the p-values of 400 histogram tests on independent batches of PCG64 numbers are spread uniformly over $[0, 1]$, as they must be under H_0. Right: the triple statistic X^2 for PCG64 and MT19937 (vertical lines) sits inside its χ^2_{7999} density under H_0 (dashed); RANDU's value, about thirty times larger, is far off the scale.

software failed badly. PCG64 passes BigCrush [156].

What no battery can do is certify a generator for *every* application. Each test is a single statistic, and a deterministic sequence always has some statistic that exposes it (for instance "replay the generator from the seed and compare"). The practical rules that follow are simple: use a modern generator that passes BigCrush; never use the low bits of a power-of-two linear congruential generator; and for a result that matters, repeat the simulation with a second generator of an unrelated design (numpy offers PCG64, MT19937, Philox and SFC64 behind the same interface) and check that the answer does not change. That last check would have caught the Monty Hall failure of Section 6.1.1 at once.

6.7 When randomness is not the point: quasi-random numbers

Return to the oldest use of random numbers in this chapter: an integral $I = \int h(x)\,\mathrm{d}x$ over the unit cube $[0, 1]^d$, estimated by the average of h at N uniform points [Wasserman p. 404, (24.2)]. The error of that average falls as $\sigma/\sqrt{N}$ (derived in Section 6.11.2), with σ the standard deviation of $h(X)$ for a uniform point X. For an integral, independence costs us accuracy. Independent points clump by chance and leave gaps by chance, and the $1/\sqrt{N}$ is exactly the size of those chance fluctuations. For an integral we do not need points that are unpredictable. We need points that cover the cube evenly. A regular grid does that, but in d dimensions it needs n^d points for n per axis, and it cannot be extended one point at a time. So the question is: are there point sets that can be extended one point at a time and still fill the cube more evenly than chance? And if so, how much accuracy does that buy?

6.7.1 Van der Corput and Sobol sequences

In one dimension there is a simple construction. Write $n = 1, 2, 3, \ldots$ in binary and mirror the digits about the binary point: $n = (b_k \ldots b_2 b_1)_2$ becomes $\phi(n) = (0.b_1 b_2 \ldots b_k)_2$, the **radical**

inverse of n. For example,

$$\phi(6) \overset{(1)}{=} \phi\big((110)_2\big) \overset{(2)}{=} (0.011)_2 \overset{(3)}{=} 0 \cdot \tfrac{1}{2} + 1 \cdot \tfrac{1}{4} + 1 \cdot \tfrac{1}{8} = \tfrac{3}{8}. \tag{6.26}$$

(1) $6 = 4 + 2 = (110)_2$.
(2) Read the digits from the right, after the binary point.
(3) The i-th digit after the point is worth 2^{-i}.

The **van der Corput sequence** $\phi(1), \phi(2), \ldots$ begins $\frac{1}{2}, \frac{1}{4}, \frac{3}{4}, \frac{1}{8}, \frac{5}{8}, \frac{3}{8}, \frac{7}{8}, \frac{1}{16}, \ldots$ (the script `09_sobol.py` prints the first eight). Every new point lands in a largest gap left by the earlier ones (Figure 6.18): after 2^m points, each interval $[j/2^m, (j+1)/2^m)$ contains exactly one of them. Why does mirroring the digits do this? The last binary digit of n alternates fastest, and it becomes the first digit of $\phi(n)$, so consecutive points jump between the two halves of the interval; the next digit chooses the quarter, and so on.

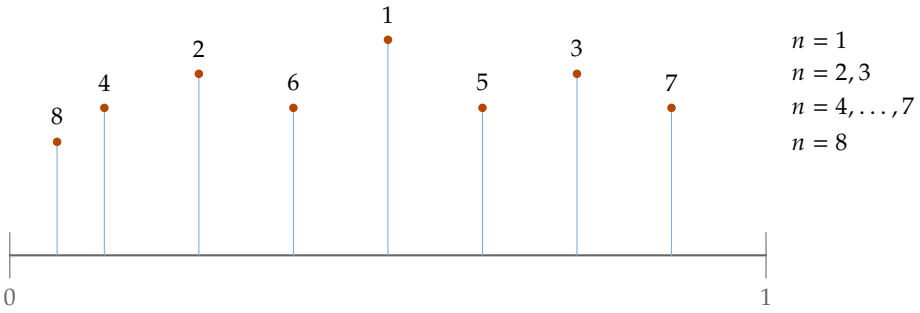


Figure 6.18: The first eight points of the van der Corput sequence, drawn at a height that marks the generation they belong to. The first point halves the interval, the next two halve the halves, the next four halve the quarters. After 2^m points every interval of length 2^{-m} holds exactly one point.

Sobol generalised this to d dimensions [158]. Coordinate j of the n-th point is built from the binary digits $b_1, b_2, \ldots$ of n as $x_{n,j} = b_1 v_{j,1} \oplus b_2 v_{j,2} \oplus \cdots$, an XOR of fixed binary fractions $v_{j,i}$, the **direction numbers** of coordinate j. For the first coordinate $v_{1,i} = 2^{-i}$, which is the van der Corput sequence; for the others the direction numbers are derived from different primitive polynomials over the bits, chosen so that the coordinates do not line up. The result is a **digital net**: the first 2^m points of a two-dimensional Sobol sequence have exactly one point in every rectangle $[a2^{-i}, (a+1)2^{-i}) \times [b2^{i-m}, (b+1)2^{i-m})$, for every $i = 0, \ldots, m$. With $m = 8$ and $i = 4$, every cell of a 16×16 grid contains exactly one of the first 256 points. Independent points do not do this: the chance that a given cell is empty is $(1 - \frac{1}{256})^{256} \approx e^{-1}$, so about 94 of the 256 cells should be empty. In Figure 6.19, 256 PCG64 points leave 92 cells empty and 256 Sobol points leave 0.

6.7.2 Why even points integrate better

Why should evenly spread points give a smaller error? To answer this we first need a number for how evenly N points cover the cube. That number is the **star discrepancy**

$$D_N^* = \sup_{\boldsymbol{y} \in [0,1]^d} \left| \frac{\#\{i : \boldsymbol{x}_i \in [\boldsymbol{0}, \boldsymbol{y})\}}{N} - y_1 y_2 \cdots y_d \right|,$$

the worst difference, over all boxes with one corner at the origin, between the fraction of points in the box and the volume of the box. In one dimension it has a simple meaning: $D_N^* = \sup_y |F_N(y) - y|$, the largest gap between the **empirical distribution function** $F_N(y)$ [the fraction of the points below y] and the uniform distribution function y, which is the

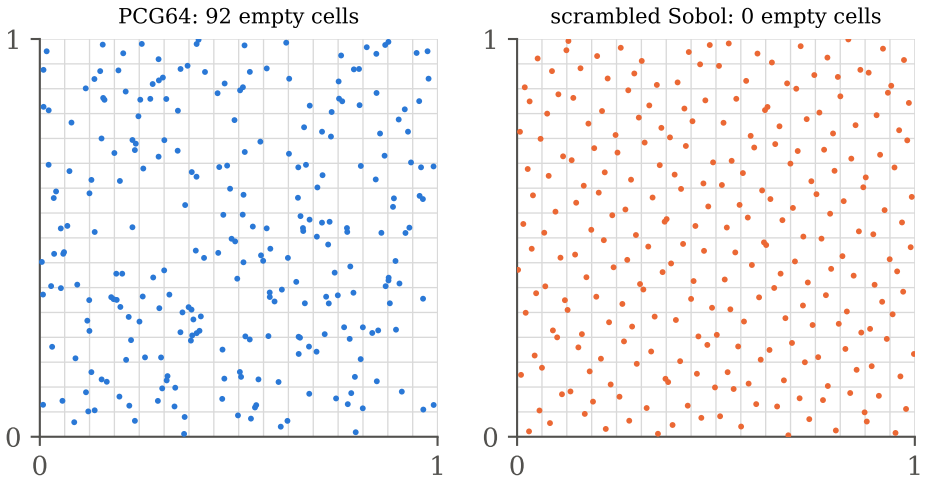


Figure 6.19: 256 points in the unit square with a 16×16 grid. Left: independent uniform points from PCG64, which clump and leave about a third of the cells empty. Right: the first 256 points of a scrambled Sobol sequence, exactly one per cell.

Kolmogorov–Smirnov statistic of Chapter 4. The answer to our question is that the integration error is bounded by the discrepancy. In one dimension, for a function h with a continuous derivative,

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N} h(x_i) - \int_0^1 h(x)\,\mathrm{d}x &\overset{(1)}{=} \int_0^1 h(x)\,\mathrm{d}\big[F_N(x) - x\big] \\
&\overset{(2)}{=} \Big[h(x)\big(F_N(x)-x\big)\Big]_0^1 - \int_0^1 \big(F_N(x)-x\big)\,h'(x)\,\mathrm{d}x \\
&\overset{(3)}{=} -\int_0^1 \big(F_N(x)-x\big)\,h'(x)\,\mathrm{d}x, \\
\left|\frac{1}{N}\sum_{i=1}^{N} h(x_i) - \int_0^1 h\,\mathrm{d}x\right| &\overset{(4)}{\le} \sup_x |F_N(x)-x| \int_0^1 |h'(x)|\,\mathrm{d}x \overset{(5)}{=} D_N^*\, V(h). \qquad (6.27)
\end{aligned}$$

(1) F_N jumps by $1/N$ at each point, so integrating against $\mathrm{d}F_N$ gives the average $\frac{1}{N}\sum h(x_i)$; integrating against $\mathrm{d}x$ gives the integral (a Stieltjes integral is just this bookkeeping).

(2) Integration by parts.

(3) At $x = 0$, $F_N(0) = 0$ (no point lies below 0); at $x = 1$, $F_N(1) = 1$. Both boundary terms vanish.

(4) Bound the integrand by the largest value of $|F_N - x|$ times $|h'|$.

(5) $V(h) = \int_0^1 |h'|\,\mathrm{d}x$ is the total variation of h, how much it goes up and down.

The error is a product of two factors, one belonging to the integrand (its variation) and one belonging to the points (their discrepancy). In d dimensions the same statement, with a suitable multidimensional variation $V_{\mathrm{HK}}(h)$, is the **Koksma–Hlawka inequality**, $|\text{error}| \le V_{\mathrm{HK}}(h)\, D_N^*$.

Key idea: Discrepancy against chance

For independent uniform points the discrepancy, like any Kolmogorov–Smirnov statistic, shrinks only as $N^{-1/2}$. For a Sobol sequence it shrinks as $(\ln N)^d/N$. For an integrand of bounded variation the Monte Carlo error $\propto N^{-1/2}$ is therefore replaced by a quasi-Monte Carlo error of order $(\ln N)^d/N$, which for moderate d and practical N behaves close to N^{-1}.

Example: a smooth integrand and one with a jump, in five dimensions. Take $d = 5$ and two integrands with known integrals: the smooth product $h(\boldsymbol{x}) = \prod_{i=1}^{5} \frac{\pi}{2} \sin \pi x_i$, whose integral is 1 (each factor integrates to $\frac{\pi}{2} \cdot \frac{2}{\pi} = 1$), and the indicator of the ball of radius $\frac{1}{2}$ about the centre of the cube, whose integral is the ball's volume $\pi^{5/2}(\frac{1}{2})^5/\Gamma(\frac{7}{2}) = 0.1645$. For $N = 2^4, \ldots, 2^{16}$ the script estimates both with PCG64 points and with scrambled Sobol points, repeats everything 40 times with independent randomisations, and plots the root-mean-square error (Figure 6.20). With PCG64 both errors fall with slope -0.51 on the log–log plot, the $N^{-1/2}$ law. With Sobol points the smooth integrand's error falls with slope -1.08, and at $N = 2^{16}$ it is 135 times smaller than the Monte Carlo error; to get the same accuracy by Monte Carlo would need about $135^2 \approx 2 \times 10^4$ times more points. For the ball the slope is only -0.57 and the gain a factor 2.5. The indicator of a ball has a jump across a curved surface, and its Hardy–Krause variation is infinite for $d \geq 2$, so the Koksma–Hlawka bound gives nothing; the points still help a little, because the part of the cube away from the surface is integrated well.

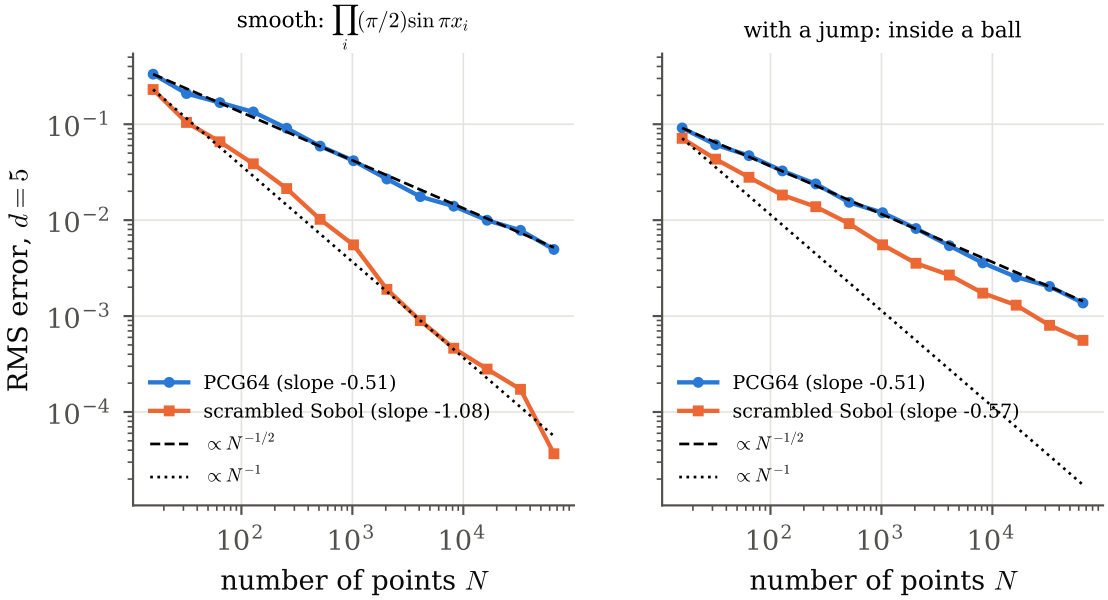


Figure 6.20: Root-mean-square error of the estimate of a five-dimensional integral against the number of points, over 40 independent randomisations. Left: a smooth integrand; the Sobol error (orange) falls about as N^{-1} (dotted), the Monte Carlo error (blue) as $N^{-1/2}$ (dashed). Right: the indicator of a ball, which has a jump; the Sobol points gain only a constant factor.

Without randomisation a quasi-random estimate has no error bar. An unscrambled Sobol sequence is deterministic, and its error, though small, cannot be estimated from the points themselves: the N points are not independent, so the formula $s/\sqrt{N}$ does not apply. **Scrambling** applies random permutations to the binary digits of every coordinate in a way that keeps the net property of Figure 6.19 but makes each individual point uniform on the cube. The estimate is then unbiased, and R independent scramblings give R independent estimates whose scatter is an honest error bar. This is how the RMS errors in Figure 6.20 were measured.

Quasi-random points are not random numbers. Sobol points are built to be as dependent as possible, each one filling the gaps left by the others. Used where independence matters (Monty's coin, the steps of a random walk, the proposals and acceptance tests of MCMC in Chapter 9) they give wrong answers. They replace a generator only in the estimation of an integral, where the dimension d is the number of uniform numbers used per sample.

How this is used in real research: Random numbers in research code

Generators. Research code in Python uses `numpy.random.default_rng` (PCG64) or, for massively parallel work, `PCG64DXSM` or `Philox`. Detector simulation with Geant4 uses the MixMax generator of the CLHEP library by default; Pythia 8 uses its own implementation of the Marsaglia–Zaman–Tsang generator; lattice field theory codes traditionally use Lüscher's RANLUX, whose "luxury level" trades speed for proven decorrelation. The old warnings (Sections 6.2 and 6.3) concern code that still calls C's `rand()` or a home-made linear congruential generator.
Seeds. A simulation campaign has one root seed, recorded with the outputs; every job derives its stream from it with `SeedSequence.spawn` or a job index, as `rng_for` does. Cosmological simulation suites store the seed of every realisation, so that any single CMB map or N-body initial condition can be regenerated exactly when it turns out to be an outlier.
Verification and research. Every numerical experiment with generators so far *verified* something we could also compute: the Monty Hall posterior, π, integrals with known values, lattice planes derived by hand. That is the right way to validate a generator. In research the same machinery becomes the instrument: the covariance of a masked, noisy power-spectrum estimator in Chapter 8 has no formula, and its only definition is a histogram over simulations. There the generator's quality can no longer be checked against an answer, and the rules above (a modern generator, long seeds, SeedSequence streams, a rerun with a second generator) are what make the result trustworthy. Quasi-Monte Carlo is used where the integrand is smooth and moderately low-dimensional: option pricing in finance, radiative-transfer and rendering integrals, and phase-space integration of smooth matrix elements, for which scrambled Sobol points are available as `scipy.stats.qmc.Sobol`.

Every simulation rests on one object: a stream of numbers that must behave like independent uniform draws on $[0, 1)$. Such a stream comes from a deterministic state machine, so its quality is a property we test, and its seeds and streams are choices we record. Uniform numbers can then be turned into draws from any distribution (Section 6.8), and averages over those draws into estimates with an error we can quote (Section 6.11.2).

Recap

- A simulation reproduces probabilities as frequencies, so it is only as good as its random numbers: a coin taken from the lowest bit of a linear congruential generator changed Monty's $\mathrm{P}(E \mid H_A)$ from $\frac{1}{2}$ to 0 or 1, and with it the posterior.
- A pseudorandom number generator is a deterministic state machine (state, transition, output, seed). Finite state forces a period; after it the estimate stops improving while the reported error bar keeps shrinking.
- Linear congruential generators $x \mapsto (ax + c) \bmod m$: period m under the Hull–Dobell conditions, the order of a for prime m; the low j bits of a power-of-two modulus have period 2^j.
- Their t-tuples lie on hyperplanes $\boldsymbol{h} \cdot \boldsymbol{u} \in \mathbb{Z}$ for every dual vector $\boldsymbol{h}$; RANDU's triples lie on 15 planes ($9u_n - 6u_{n+1} + u_{n+2} \in \mathbb{Z}$). The spectral test measures the widest gap $d_t = 1/\nu_t$; Marsaglia's bound allows at most $(t!\, m)^{1/t}$ planes.
- Modern generators separate a cheap long-period transition (xorshift and MT19937, linear over bits; the 128-bit LCG of PCG64) from a mixing output function (tempering; XSL-RR). Our ten-line PCG64 reproduces numpy's bits.
- Seeds: long, hashed through a SeedSequence, one root recorded; streams by index or

spawn. Numpy's doubles are $k/2^{53}$ and can be 0.

- Testing: H_0 = independent and uniform. Histogram χ^2, serial correlation ($\operatorname{Var} r \approx 1/N$), runs ($\mathbb{E}\,R = (2N-1)/3$), triples; RANDU passes three and fails the fourth; too good a fit also rejects H_0.
- Quasi-random (Sobol) points fill space evenly; for smooth integrands the error falls close to N^{-1} instead of $N^{-1/2}$ (Koksma–Hlawka), but they are not random numbers.

6.8 Uniform numbers into any distribution: the inverse CDF

A muon stops in a block of scintillator and decays. We want to simulate a million such decays, each with its own decay time, to see how a detector with a 100 ns dead time distorts the measured lifetime.

What we have. The pseudorandom generators of the first part of this chapter give us, on demand, numbers $u_1, u_2, \ldots$ that behave like independent draws from the uniform distribution on $[0, 1)$.

What we need. Decay times are not uniform. They are exponential, with density $e^{-t/\tau}/\tau$, where $\tau = 2.197\,\mu$s is the muon lifetime. Other simulations need other shapes: cosmic-ray energies follow a steep power law, and the number of photons in a CCD pixel is Poisson. So every simulation in physics starts with the same small question.

The question. Given a supply of uniform numbers U, which function $x(U)$ turns them into numbers that follow a prescribed distribution?

To answer it we first ask the reverse question: if we feed a random variable into its own cumulative distribution function, how is the result distributed? The answer to that turns into a one-line recipe.

6.8.1 Why $F(X)$ is uniform

Recall the **cumulative distribution function** (CDF) of a random variable X, $F(x) = \mathrm{P}(X \le x)$ [the probability that X comes out at or below x. It climbs from 0 to 1], and for a continuous X it is the integral of the density, $F(x) = \int_{-\infty}^{x} p(x')\,\mathrm{d}x'$ (Chapter 1). Suppose for the moment that F is continuous and strictly increasing, so that it has an ordinary inverse F^{-1}: for every $u \in (0,1)$ there is exactly one x with $F(x) = u$. That x is the u-quantile, the value below which a fraction u of the probability lies.

Now feed X into its own CDF and ask how the new number $F(X)$ is distributed. For any $u \in (0,1)$,

$$\mathrm{P}\big(F(X) \le u\big) \overset{(1)}{=} \mathrm{P}\big(X \le F^{-1}(u)\big) \overset{(2)}{=} F\big(F^{-1}(u)\big) \overset{(3)}{=} u. \tag{6.28}$$

(1) F is strictly increasing, so $F(X) \le u$ happens exactly when $X \le F^{-1}(u)$: applying the increasing function F^{-1} to both sides of an inequality keeps its direction.
(2) This is the definition of F, evaluated at the point $F^{-1}(u)$.
(3) F^{-1} undoes F.

So $F(X)$ is uniform on $[0,1]$, whatever the distribution of X: its CDF is $\mathrm{P}(F(X) \le u) = u$, and that is the CDF of the uniform distribution. This is the **probability integral transform** [CasellaBerger Thm 2.1.10], already met in Chapter 1 and behind the uniform distribution of

p-values under the null hypothesis in Chapter 4. In words: $F(X)$ is the fraction of the distribution lying below the value we drew, and every fraction between 0 and 1 is equally likely.

Read backwards, the same fact is a recipe. Take U uniform and set $X = F^{-1}(U)$. Then for every x,

$$\mathrm{P}\big(F^{-1}(U) \le x\big) \overset{(1)}{=} \mathrm{P}\big(U \le F(x)\big) \overset{(2)}{=} F(x), \tag{6.29}$$

(1) Apply the increasing function F to both sides of $F^{-1}(U) \le x$.
(2) U is uniform, so $\mathrm{P}(U \le c) = c$ for any $c \in [0,1]$, here $c = F(x)$.

so $F^{-1}(U)$ has exactly the CDF F we wanted.

The same step can be read as a matching condition. Each value u is sent to one value x, so the probability that U lands in a small strip $[u, u + \mathrm{d}u]$ must equal the probability that X lands in the matching interval $[x, x + \mathrm{d}x]$. The first is $\mathrm{d}u$, because U is uniform; the second is $p(x)\,\mathrm{d}x$. Setting them equal gives $\mathrm{d}u = p(x)\,\mathrm{d}x$, which integrates to $u = F(x)$ [Cowan (3.4)].

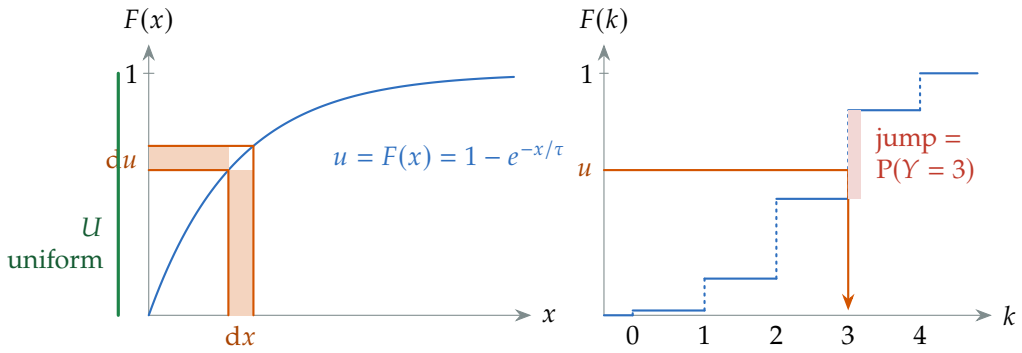


Figure 6.21: The inverse-CDF method. Left: a uniform number u on the vertical axis is carried across to the curve $u = F(x)$ (an exponential CDF, drawn for $\tau = 1$) and down to $x = F^{-1}(u)$. A strip of height $\mathrm{d}u$ lands on an interval of width $\mathrm{d}x = \mathrm{d}u/p(x)$, narrow where the curve is steep, so values pile up where the density is high. Right: for a discrete variable the CDF is a staircase, and u selects the step it falls under; the chance of selecting k is the height of the jump at k.

Figure 6.21 (left) shows the same thing as a picture. The uniform number lives on the vertical axis, where every strip of height $\mathrm{d}u$ is equally likely. Where F is steep, the density $p = F'$ is large, and a strip of fixed height maps onto a narrow interval of x: many draws crowd into a small range. Where F is flat, the same strip is stretched over a wide interval, and the draws are sparse.

Working definition 6.3 (inverse-transform sampling)

To draw X with CDF F: draw $U \sim \mathrm{Uniform}(0,1)$ and return $X = F^{-1}(U)$ [the U-quantile of the target, the value below which a fraction U of the probability lies]. For a CDF with jumps or flat stretches use the **generalised inverse** $F^{-1}(u) = \min\{x : F(x) \ge u\}$. *Operational test:* it needs a closed form (or a fast numerical solution) of $F(x) = u$ for x.

6.8.2 Continuous examples: decay times and power laws

Example: muon decay times. A muon at rest decays with lifetime τ, so its decay time t has density $p(t) = e^{-t/\tau}/\tau$ for $t \ge 0$ (the waiting time of a Poisson process, Chapter 2). Solve $F(t) = u$:

$$u = F(t) \overset{(1)}{=} \int_0^t \frac{e^{-t'/\tau}}{\tau}\,\mathrm{d}t' \overset{(2)}{=} 1 - e^{-t/\tau} \quad\Longrightarrow\quad t = F^{-1}(u) \overset{(3)}{=} -\tau \ln(1-u). \tag{6.30}$$

(1) The CDF is the integral of the density from the lowest possible time, 0.
(2) The antiderivative of $e^{-t'/\tau}/\tau$ is $-e^{-t'/\tau}$.
(3) Rearrange to $e^{-t/\tau} = 1 - u$ and take the natural logarithm of both sides.

The same formula appears as [Cowan (3.5)–(3.7)] and as [CasellaBerger Ex. 5.6.3]. Since $1 - U$ is uniform whenever U is, $t = -\tau \ln U$ works equally well.

Does it work? With $N = 100000$ uniforms, the script below gives a sample mean $\bar{t} = 2.196\,\mu$s, against the true mean $\tau = 2.197\,\mu$s. Its standard error, the sample standard deviation divided by $\sqrt{N}$ (Chapter 3), is $0.006967\,\mu$s. The sample standard deviation itself is $2.203\,\mu$s; for an exponential it should equal the mean, and it does. The histogram lies on the density (Figure 6.22, left).

Example: a cosmic-ray power law. Above about 10 GeV the flux of cosmic rays falls roughly as $E^{-\gamma}$ with spectral index $\gamma \approx 2.7$ [the power with which the number of particles per unit energy drops]. To simulate primaries between $E_{\min}$ and $E_{\max}$ we normalise the density on that range and invert its CDF. Write $\gamma \neq 1$ and abbreviate $a = E_{\min}^{1-\gamma}$ and $b = E_{\max}^{1-\gamma}$ (with $\gamma > 1$, $a > b$):

$$p(E) \overset{(1)}{=} \frac{(\gamma-1)\,E^{-\gamma}}{a-b}, \qquad F(E) \overset{(2)}{=} \int_{E_{\min}}^{E} \frac{(\gamma-1)E'^{-\gamma}}{a-b}\,\mathrm{d}E' \overset{(3)}{=} \frac{a - E^{1-\gamma}}{a-b},$$
$$E = F^{-1}(u) \overset{(4)}{=} \left[a - u\,(a-b)\right]^{1/(1-\gamma)}. \tag{6.31}$$

(1) The normalisation: $\int_{E_{\min}}^{E_{\max}} E^{-\gamma}\mathrm{d}E = \left[E^{1-\gamma}/(1-\gamma)\right]_{E_{\min}}^{E_{\max}} = (a-b)/(\gamma-1)$, and p is $E^{-\gamma}$ divided by this.
(2) The CDF is the integral of the density from the lower end of the range.
(3) The same antiderivative, now between $E_{\min}$ and E: $(\gamma-1)\,[E_{\min}^{1-\gamma} - E^{1-\gamma}]/(\gamma-1)$ divided by $a-b$.
(4) Set $F(E) = u$, so $E^{1-\gamma} = a - u(a-b)$, and raise both sides to the power $1/(1-\gamma)$.

Check the ends: $u = 0$ gives $E = a^{1/(1-\gamma)} = E_{\min}$ and $u = 1$ gives $E_{\max}$.

Now the numbers, with $\gamma = 2.7$, $E_{\min} = 10$ GeV and $E_{\max} = 10^6$ GeV. The exact fraction of primaries above 1 TeV is $1 - F(10^3\,\text{GeV}) = 3.981 \times 10^{-4}$. Among 1000000 simulated energies, 394 are above 1 TeV, a fraction 3.940×10^{-4}. The expected count is 398, and a Poisson count of 398 fluctuates by about $\sqrt{398} \approx 20$, so the simulation agrees within one standard deviation. The largest of the million energies is only $4.31e+04$ GeV: a steep spectrum rarely reaches its upper cut-off. That is why air-shower arrays, which hunt the highest energies, must cover square kilometres. On logarithmic axes the histogram is a straight line of slope -2.7 (Figure 6.22, middle).

Example: isotropic directions. A source that emits isotropically sends equal numbers of particles into equal solid angles. Which two numbers should we draw, and from what? Write the solid-angle element in the polar angle θ and the azimuth ϕ: $\mathrm{d}\Omega = \sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi = \mathrm{d}(\cos\theta)\,\mathrm{d}\phi$ up to a sign, because $\mathrm{d}(\cos\theta) = -\sin\theta\,\mathrm{d}\theta$. Equal solid angles are therefore equal areas in the variables $(\cos\theta, \phi)$. So "uniform in solid angle" means that $\cos\theta$ is uniform on $[-1, 1]$ and ϕ is uniform on $[0, 2\pi)$, independently. Both inverse CDFs are straight lines: $\cos\theta = 2u_1 - 1$ and $\phi = 2\pi u_2$. The common mistake is to draw θ uniform on $[0, \pi]$. That piles directions up at the poles, because a step in θ near a pole sweeps out less solid angle than the same step near the equator. We will use these isotropic directions in Section 6.11.4 to compute the acceptance of a detector.

6.8.3 Discrete outcomes by cumulative sums

A discrete random variable Y taking values $y_1 < y_2 < \dots$ has a staircase CDF (Figure 6.21, right). The jump at y_{i+1} has height $F(y_{i+1}) - F(y_i) = \mathrm{P}(Y = y_{i+1})$. The generalised inverse of the recipe box above says: draw U and return the value y_{i+1} whose step contains it, that is, the one with $F(y_i) \le U < F(y_{i+1})$ (with the conventions $y_0 = -\infty$ and $F(y_0) = 0$). Why does this give each value with the right probability? One line:

$$\mathrm{P}\big(F(y_i) \le U < F(y_{i+1})\big) \overset{(1)}{=} F(y_{i+1}) - F(y_i) \overset{(2)}{=} \mathrm{P}(Y = y_{i+1}). \tag{6.32}$$

(1) For a uniform U the probability of landing in an interval is its length.
(2) The CDF jumps at y_{i+1} by exactly the probability of that value.

This is [CasellaBerger (5.6.7)]. In words: lay the probabilities $\mathrm{P}(Y = y_1), \mathrm{P}(Y = y_2), \dots$ end to end along $[0, 1]$ like segments of a ruler, throw U at the ruler, and report the segment it hits.

Example: a binomial table. For $Y \sim \mathrm{Binomial}(4, 5/8)$, the probabilities are $\binom{4}{k}(5/8)^k(3/8)^{4-k}$. For example $\mathrm{P}(Y = 0) = (3/8)^4 = 81/4096$ and $\mathrm{P}(Y = 1) = 4 \cdot (5/8)(3/8)^3 = 540/4096$. The cumulative sums are $F(0) = 0.01978$, $F(1) = 0.1516$, $F(2) = 0.4812$, $F(3) = 0.8474$, $F(4) = 1$, which rounds to Casella and Berger's table 0.020, 0.152, 0.481, 0.847 [CasellaBerger (5.6.8)]. So $U = 0.6$ gives $Y = 3$, because $0.481 \le 0.6 < 0.847$. Over 100000 draws, the largest difference between an observed frequency and its exact probability is 0.001509. That is the size we should expect: a frequency estimated from N draws scatters by its binomial standard error $\sqrt{p(1-p)/N}$, which is about 0.0015 for a probability near $p = 1/3$.

Example: Poisson counts. A PMT pixel that sees on average $\lambda = 3.5$ photons per exposure gives Poisson(3.5) counts. The range is infinite, but the cumulative sum reaches 1 to double precision well before $k = 30$, so a table of 30 entries suffices. Over 100000 draws, the sample mean is 3.499 and the sample variance 3.546, both close to λ, as a Poisson variable requires. For large tables one searches by bisection, or starts the search near the mean [CasellaBerger p. 203].

6.8.4 The code

Listing 6.8: Decay times, power-law energies and discrete counts by the inverse CDF (`code/ch06/21_inverse_cdf.py`)

```
# ---------------------------------------------------------------- 1. exponential decay times
tau = 2.197                                     # muon lifetime in microseconds
u = rng.random(N)
t = -tau * np.log(1.0 - u)                      # inverse CDF of 1 - exp(-t/tau)
out.update(SixBInvTmean=t.mean(), SixBInvTstd=t.std(ddof=1),
           SixBInvTse=t.std(ddof=1) / np.sqrt(N))

# ---------------------------------------------------------------- 2. power-law energies
g, Emin, Emax = 2.7, 10.0, 1.0e6
a, b = Emin ** (1 - g), Emax ** (1 - g)
E = (a - rng.random(N * 10) * (a - b)) ** (1.0 / (1 - g))   # 10^6 energies
frac_TeV = np.mean(E > 1000.0)
exact_TeV = (1000.0 ** (1 - g) - b) / (a - b)
out.update(SixBInvNE=E.size, SixBInvFracTeV=frac_TeV, SixBInvExactTeV=exact_TeV,
           SixBInvCountTeV=int(np.sum(E > 1000.0)), SixBInvEmax=E.max())

# ---------------------------------------------------------------- 3. discrete via cumulative sums
```

```
k4 = np.arange(5)
p4 = stats.binom.pmf(k4, 4, 5 / 8)
cum4 = np.cumsum(p4)                                  # 0.020, 0.152, 0.481, 0.847, 1
y4 = np.searchsorted(cum4, rng.random(N), side="right")     # first k with U < F(k)
freq4 = np.bincount(y4, minlength=5) / N

lam = 3.5
kp = np.arange(0, 30)
cump = np.cumsum(stats.poisson.pmf(kp, lam))
yp = np.searchsorted(cump, rng.random(N), side="right")
freqp = np.bincount(yp, minlength=kp.size)[: kp.size] / N
```

Line 32 is (6.30): `rng.random` returns numbers in $[0, 1)$, so $1 - u$ is never zero and the logarithm is always finite (with $-\tau \ln u$ the rare draw $u = 0$ would give infinity). Line 39 is (6.31), applied to a million uniforms at once. Lines 48–49 are the discrete recipe: `np.cumsum` builds the ruler of cumulative probabilities, and `np.searchsorted(cum, u, side="right")` returns, for each u, the index of the first entry of the ruler that is strictly larger than u, which is the k with $F(k-1) \le u < F(k)$. Lines 52–56 do the same for the Poisson table.

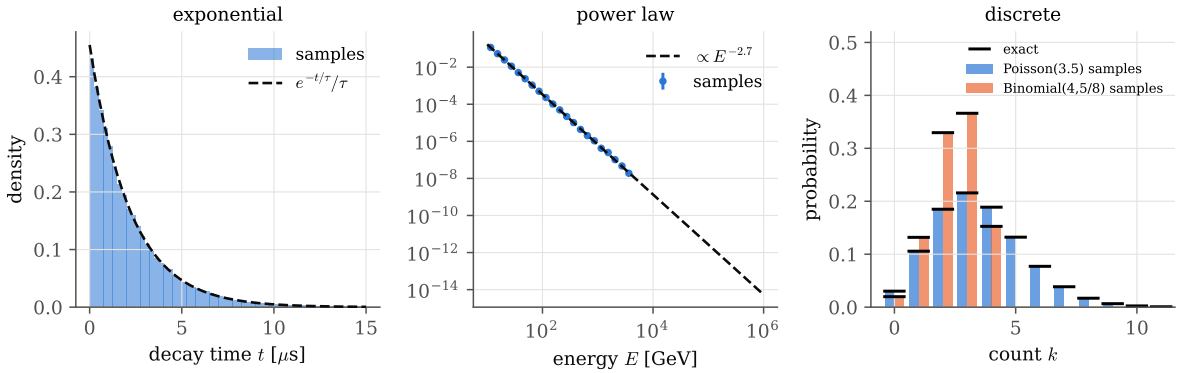


Figure 6.22: Samples made by the inverse-CDF method, against the exact distributions (black dashed). Left: 10^5 muon decay times. Middle: 10^6 energies from an $E^{-2.7}$ spectrum between 10 GeV and 10^6 GeV, on logarithmic axes, with $\sqrt{n}$ error bars on bins holding at least ten events; no simulated energy comes near the upper cut-off. Right: frequencies of Poisson and binomial counts made by cumulative sums, with the exact probabilities as black ticks.

Watch out: Inverting the wrong thing

The recipe needs the inverse of the *cumulative* distribution, not of the density: $x = p^{-1}(u)$ produces nonsense. And a formula for F is not enough: we must be able to solve $F(x) = u$ for x. The Gaussian CDF Φ has no closed-form inverse, which is why the next section needs a different trick.

The inverse-CDF method has one more property that the later methods lack: it is *monotone*. Nearby values of u give nearby values of x, and the x's come out in the same order as the u's. So any structure put into the uniform numbers, for instance the evenly spread quasi-random points of this chapter's first part or the strata of Section 6.15, passes straight through to the target distribution. Casella and Berger call it a *direct* method: one uniform in, one draw out, with nothing thrown away [CasellaBerger §5.6.1].

6.9 Gaussian numbers: Box–Muller and the Cholesky factor

Gaussian draws enter simulations of amplifier noise, gravitational-wave detector noise, calorimeter energy resolution and the primordial fluctuations of the CMB. In the first three cases they model measurement fluctuations, while in the last they model the signal itself. So we need a fast, exact way to turn uniform numbers into standard normal numbers $Z \sim \mathcal{N}(0,1)$, and then into correlated Gaussian vectors with any covariance matrix we like.

The inverse-CDF method of Section 6.8 stalls here, because it needs us to solve $F(x) = u$ for x. The Gaussian CDF,

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\,\mathrm{d}t,$$

has no closed form, and neither does its inverse. A second route also fails. Sums of logarithms of uniforms build χ^2, gamma and beta variables, for example

$$-2\sum_{j=1}^{\nu} \ln U_j \sim \chi^2_{2\nu},$$

since each $-2\ln U_j$ is an exponential of mean 2, which is a χ^2_2 variable. But these sums only produce an *even* number of degrees of freedom. So they never give $\chi^2_1 = Z^2$, and hence never a single normal [CasellaBerger (5.6.5)]. The way out is to stop asking for one Gaussian and ask for two.

6.9.1 Two Gaussians are easier than one

The question. Which pair of functions of two independent uniforms (U_1, U_2) gives two independent standard normals (X, Y)?

Work backwards from the answer. Two independent standard normals have joint density

$$p(x,y) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \cdot \frac{1}{\sqrt{2\pi}} e^{-y^2/2} = \frac{1}{2\pi}\, e^{-(x^2+y^2)/2}.$$

This depends only on the distance $r = \sqrt{x^2+y^2}$ from the origin, so it is rotationally symmetric, like the Maxwell velocity distribution in two dimensions. A round density is simplest in round coordinates, so switch to polar coordinates $x = r\cos\theta$, $y = r\sin\theta$. There the area element is $\mathrm{d}x\,\mathrm{d}y = r\,\mathrm{d}r\,\mathrm{d}\theta$ (the Jacobian of polar coordinates, Figure 6.23). The probability in a small patch must be the same whichever coordinates we use to describe it:

$$p_{R,\Theta}(r,\theta)\,\mathrm{d}r\,\mathrm{d}\theta \overset{(1)}{=} p(x,y)\,\mathrm{d}x\,\mathrm{d}y \overset{(2)}{=} \frac{1}{2\pi} e^{-r^2/2}\, r\,\mathrm{d}r\,\mathrm{d}\theta \overset{(3)}{=} \underbrace{\frac{1}{2\pi}\,\mathrm{d}\theta}_{\Theta}\ \underbrace{r\, e^{-r^2/2}\,\mathrm{d}r}_{R}. \tag{6.33}$$

(1) Probability is conserved under a change of variables: the patch $[r, r+\mathrm{d}r] \times [\theta, \theta+\mathrm{d}\theta]$ and its image in the plane hold the same probability.

(2) Insert the density, which is $e^{-r^2/2}/2\pi$, and the polar area element $r\,\mathrm{d}r\,\mathrm{d}\theta$.

(3) The expression is a product of a function of θ alone and a function of r alone, so R and Θ are independent. Θ is uniform on $[0, 2\pi)$ with density $1/2\pi$, and R has density $r\,e^{-r^2/2}$ on $r \geq 0$ (the Rayleigh distribution of Chapter 2).

So the problem has split into two easy pieces: make a uniform angle, and make an independent radius with density $r\,e^{-r^2/2}$.

The angle. $\Theta = 2\pi U_2$.

The radius. It is easier to make its square, $S = R^2$. Since $\mathrm{d}s = 2r\,\mathrm{d}r$, the density of S is $p_S(s)\,\mathrm{d}s = r\,e^{-r^2/2}\mathrm{d}r = \frac{1}{2}e^{-s/2}\mathrm{d}s$, an exponential with mean 2. (It is χ^2_2, the sum of two squared standard normals.) An exponential we already know how to make: (6.30) with $\tau = 2$ gives $S = -2\ln U_1$, and $R = \sqrt{S}$.

Putting the pieces together gives the **Box–Muller transform** [159], [CasellaBerger Ex. 5.6.4]:

$$R = \sqrt{-2\ln U_1}, \qquad \Theta = 2\pi U_2, \qquad X = R\cos\Theta, \qquad Y = R\sin\Theta. \tag{6.34}$$

We built (6.34) piece by piece, working backwards. Now check it forwards: start from (U_1, U_2), whose density is 1 on the unit square, and compute the density of (X, Y). First invert the map: $U_1 = e^{-(X^2+Y^2)/2}$ and $U_2 = \theta(X,Y)/2\pi$, with θ the polar angle of the point (X, Y). For a one-to-one change of variables, the new density is the old density times the absolute value of the Jacobian determinant of the inverse map:

$$\begin{aligned} p(x,y) &\overset{(1)}{=} 1 \cdot \left|\det\begin{pmatrix} \partial u_1/\partial x & \partial u_1/\partial y \\ \partial u_2/\partial x & \partial u_2/\partial y \end{pmatrix}\right| \overset{(2)}{=} \left|\det\begin{pmatrix} -x\,u_1 & -y\,u_1 \\ -\dfrac{y}{2\pi r^2} & \dfrac{x}{2\pi r^2} \end{pmatrix}\right| \\ &\overset{(3)}{=} \left|-\frac{u_1\,(x^2+y^2)}{2\pi r^2}\right| \overset{(4)}{=} \frac{1}{2\pi}e^{-(x^2+y^2)/2} = \frac{e^{-x^2/2}}{\sqrt{2\pi}}\cdot\frac{e^{-y^2/2}}{\sqrt{2\pi}}. \end{aligned} \tag{6.35}$$

(1) The multivariate change-of-variables rule of Chapter 1, with input density 1.

(2) Differentiate $u_1 = e^{-(x^2+y^2)/2}$, which gives $-x\,u_1$ and $-y\,u_1$. For the angle, $\theta = \arctan(y/x)$ has $\partial\theta/\partial x = -y/r^2$ and $\partial\theta/\partial y = x/r^2$, and $u_2 = \theta/2\pi$.

(3) The determinant is $(-xu_1)(x/2\pi r^2) - (-yu_1)(-y/2\pi r^2)$.

(4) $x^2 + y^2 = r^2$ cancels, and $u_1 = e^{-r^2/2}$.

The result is the product of two standard normal densities, so X and Y are independent $\mathcal{N}(0,1)$: every pair of uniforms gives two Gaussians, exactly, with no approximation and nothing thrown away.

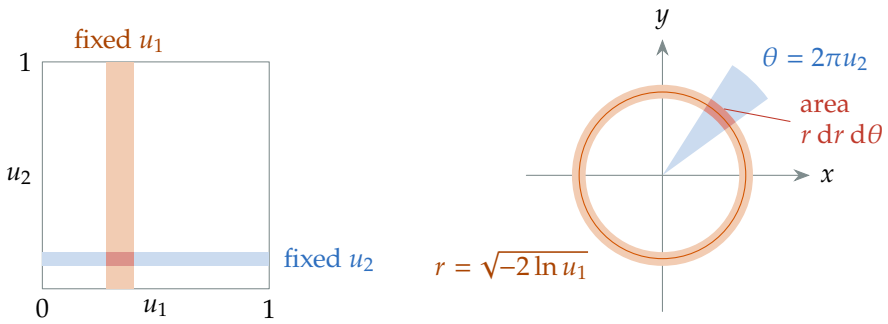


Figure 6.23: The Box–Muller map. A vertical strip of the unit square (fixed u_1) becomes a thin ring of radius $\sqrt{-2\ln u_1}$; a horizontal strip (fixed u_2) becomes a wedge at angle $2\pi u_2$, which reaches every radius (it is cut at the edge of the plot). Their overlap, the small red rectangle, is carried to the red patch of area $r\,\mathrm{d}r\,\mathrm{d}\theta$. Small u_1 gives large radii, but the logarithm makes those rare, which produces the Gaussian tails.

Example: checking a hundred thousand pairs. The script below applies (6.34) to $N = 100000$ pairs of uniforms. If the outputs are independent standard normals, five things should hold, and we check each against its expected scatter.

- Means near 0: they are 3.503×10^{-4} and 9.210×10^{-4}, with standard error $1/\sqrt{N} \approx 0.003$.
- Variances near 1: they are 1.006 and 1.001, with standard error $\sqrt{2/N} \approx 0.0045$.
- No correlation: the sample correlation is 0.003211.
- $X^2 + Y^2$ exponential with mean 2: its mean is 2.007.
- Gaussian tails: the fourth moment $\mathbb{E}[X^4]$ is 3.054, against 3 for a Gaussian.

A Kolmogorov–Smirnov test of X against $\mathcal{N}(0,1)$ (Chapter 4) gives $p = 0.9435$. Figure 6.24 shows the input square, the output cloud and the histograms.

Listing 6.9: Box–Muller and its polar variant (`code/ch06/22_box_muller.py`)

```
28 def box_muller(u1, u2):
29     r = np.sqrt(-2.0 * np.log(u1))              # radius: R^2 = -2 ln U1 is exponential,
       mean 2
30     th = 2.0 * np.pi * u2                       # angle: uniform on [0, 2 pi)
31     return r * np.cos(th), r * np.sin(th)
32
33
34 u1 = 1.0 - rng.random(N)                       # in (0, 1], so log(u1) is finite
35 u2 = rng.random(N)
36 x, y = box_muller(u1, u2)
37
38 # polar method: points uniform in the square, keep those inside the unit disc
39 v = 2.0 * rng.random((N, 2)) - 1.0
40 s = np.sum(v**2, axis=1)
41 keep = (s < 1.0) & (s > 0.0)
42 vk, sk = v[keep], s[keep]
43 fac = np.sqrt(-2.0 * np.log(sk) / sk)
44 xp = vk[:, 0] * fac
```

Line 29 is the radius, line 30 the angle, line 31 the two outputs. Line 34 uses $1 - u$ so that u_1 lies in $(0, 1]$ and $\ln u_1$ is finite.

Example: the polar method. Lines 38–44 implement a variant due to Marsaglia and Bray [160] that avoids computing a sine and a cosine. Box–Muller needs two things: a uniform number for the radius, and the cosine and sine of a uniform angle. The polar method gets both from one random point in the unit disc.

The recipe. Draw (V_1, V_2) uniform in the square $[-1, 1]^2$, and keep the pair only if $S = V_1^2 + V_2^2 < 1$, that is, only if the point lies in the unit disc. A kept point is uniform in the disc.

Why it works. Its angle is uniform, so $V_1/\sqrt{S}$ and $V_2/\sqrt{S}$ are the cosine and sine of a uniform angle. Its S is uniform on $(0, 1)$, because the fraction of the disc within radius $\sqrt{s}$ is $\pi s/\pi = s$; and S is independent of the angle. So S can play the role of U_1, and $X = V_1\sqrt{-2\ln S/S}$, $Y = V_2\sqrt{-2\ln S/S}$ are again independent normals.

The price. Some pairs are thrown away. The kept fraction is the area of the disc over the area of the square, $\pi/4 = 0.785$, and the script finds 0.7847. Throwing candidates away to shape a distribution is the whole idea of the next section.

Watch out: Twelve uniforms minus six

An old shortcut makes a "Gaussian" as $U_1 + \cdots + U_{12} - 6$, which has mean 0 and variance $12 \times \frac{1}{12} = 1$ and is close to normal by the central limit theorem (Chapter 2). It can never exceed 6 in magnitude, and its tails are wrong long before that. A simulation of the background tail beyond 5σ, which is exactly where a discovery claim lives, would be wrong. Use an exact method.

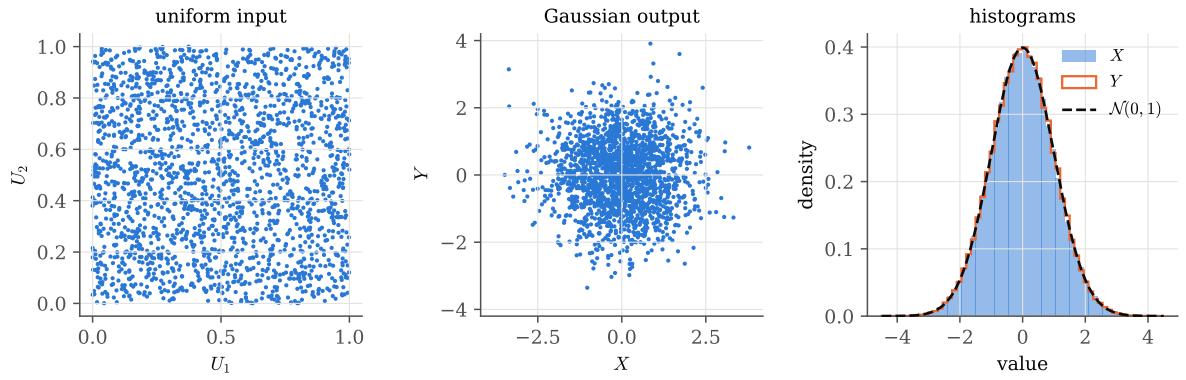


Figure 6.24: Box–Muller in action. Left: 2000 pairs of uniforms. Middle: the same pairs after the map (6.34), a round Gaussian cloud. Right: the histograms of all 10^5 values of X and of Y lie on the standard normal density (black dashed).

6.9.2 Correlated Gaussian vectors: the Cholesky factor

Box–Muller gives independent standard normals. Physics needs correlated ones: the pixel values of a CMB map, the strain in two detectors driven by a common environmental noise, the residuals of a fit with a non-diagonal covariance matrix (Chapter 3). The target is a Gaussian vector $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$ of dimension n with a prescribed mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}$.

The question. Which linear map turns a vector $\boldsymbol{z}$ of n independent standard normals, whose covariance is the identity $\mathbb{1}$, into a vector with covariance $\mathbf{C}$?

Try $\boldsymbol{X} = \boldsymbol{\mu} + \mathbf{L}\boldsymbol{z}$ for some $n \times n$ matrix $\mathbf{L}$. The transformation rule for covariance matrices under a linear map, $\mathrm{Cov}(\mathbf{A}\boldsymbol{z}) = \mathbf{A}\,\mathrm{Cov}(\boldsymbol{z})\mathbf{A}^{\mathsf{T}}$ (Chapter 1), gives

$$\mathrm{Cov}(\boldsymbol{X}) \overset{(1)}{=} \mathrm{Cov}(\mathbf{L}\boldsymbol{z}) \overset{(2)}{=} \mathbf{L}\,\mathrm{Cov}(\boldsymbol{z})\,\mathbf{L}^{\mathsf{T}} \overset{(3)}{=} \mathbf{L}\,\mathbb{1}\,\mathbf{L}^{\mathsf{T}} = \mathbf{L}\mathbf{L}^{\mathsf{T}}. \tag{6.36}$$

(1) Adding the constant $\boldsymbol{\mu}$ shifts the vector but not its fluctuations.
(2) The covariance of a linear transformation: $\mathbb{E}[(\mathbf{L}\boldsymbol{z})(\mathbf{L}\boldsymbol{z})^{\mathsf{T}}] = \mathbf{L}\,\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^{\mathsf{T}}]\,\mathbf{L}^{\mathsf{T}}$, since $\boldsymbol{z}$ has mean zero.
(3) The components of $\boldsymbol{z}$ are independent with unit variance.

So any $\mathbf{L}$ with $\mathbf{L}\mathbf{L}^{\mathsf{T}} = \mathbf{C}$ gives the right covariance. The result is also Gaussian, because a linear combination of Gaussians is Gaussian (Chapter 2). For a symmetric positive definite $\mathbf{C}$ there is exactly one lower-triangular such $\mathbf{L}$ with positive diagonal, the **Cholesky factor** [a "square root" of $\mathbf{C}$ with zeros above the diagonal]. It costs about $n^3/3$ operations to compute.

Example: two correlated variables. For $n = 2$ with standard deviations σ_x, σ_y and correlation coefficient ρ, write $\mathbf{L} = \left(\begin{smallmatrix} l_{11} & 0 \\ l_{21} & l_{22} \end{smallmatrix}\right)$ and match $\mathbf{L}\mathbf{L}^{\mathsf{T}}$ to $\mathbf{C}$ entry by entry:

$$\begin{pmatrix} l_{11}^2 & l_{11}l_{21} \\ l_{11}l_{21} & l_{21}^2 + l_{22}^2 \end{pmatrix} \overset{(1)}{=} \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix} \Longrightarrow\; l_{11} \overset{(2)}{=} \sigma_x, \quad l_{21} \overset{(3)}{=} \rho\sigma_y, \quad l_{22} \overset{(4)}{=} \sigma_y\sqrt{1-\rho^2}. \tag{6.37}$$

(1) Multiply out $\mathbf{L}\mathbf{L}^{\mathsf{T}}$ and require it to equal the 2×2 covariance matrix.
(2) The top-left entry, taking the positive root.
(3) The off-diagonal entry, divided by $l_{11} = \sigma_x$.
(4) The bottom-right entry: $l_{22}^2 = \sigma_y^2 - \rho^2\sigma_y^2$.

So $x = \mu_x + \sigma_x z_1$ and $y = \mu_y + \rho\sigma_y z_1 + \sigma_y\sqrt{1-\rho^2}\, z_2$. In this construction y inherits a fraction ρ of the fluctuation z_1 of x, in units of its own width σ_y, and adds fresh noise z_2 of its own (the same example as Chapter 2). For $\sigma_x = 1$, $\sigma_y = 2$, $\rho = 0.8$ this gives $l_{21} = 1.6$ and $l_{22} = 1.2$. From 50 000 draws the sample covariances are 1.006, 1.609 and 4.011, against 1, 1.6 and 4, and the sample correlation is 0.8009.

Example: a correlated field on a line. A random field is a random function of position. On a grid of n points it is simply a random vector of n values. "Statistically homogeneous" means that the covariance of the values at two points depends only on their separation r. Take $n = 200$ points on $[0, 20]$ and the exponential correlation $C(r) = e^{-|r|/\ell}$ with correlation length $\ell = 2$, so $\mathbf{C}_{ij} = e^{-|x_i - x_j|/\ell}$. The script factors this 200×200 matrix and draws $M = 4000$ fields $\boldsymbol{f} = \mathbf{L}\boldsymbol{z}$.

Listing 6.10: A correlated one-dimensional Gaussian field by the Cholesky factor (`code/ch06/23_cholesky_field.py`)

```
39 n, Lbox, ell, M = 200, 20.0, 2.0, 4000
40 xg = np.linspace(0, Lbox, n)
41 C = np.exp(-np.abs(xg[:, None] - xg[None, :]) / ell)
42 L = np.linalg.cholesky(C)
43 Z = rng.standard_normal((M, n))
44 F = Z @ L.T                                  # M realisations, each row one field
45 Fbad = Z @ C.T                               # the pitfall: multiply by C instead
       of L
```

Line 41 builds the covariance matrix from the separations of all pairs of points, line 42 factors it, and line 44 multiplies a whole stack of white-noise vectors by $\mathbf{L}^{\mathsf{T}}$ from the right, which is the row-vector form of $\mathbf{L}\boldsymbol{z}$.

What comes out? The realisations (Figure 6.25, left) wander smoothly over distances of order ℓ, while the white noise they are made from jumps from point to point. The correlation function estimated from the realisations follows $e^{-r/\ell}$: at $r = \ell$ it is 0.3608, against $e^{-1} = 0.368$. The sample covariance matrix has 40 000 entries. The largest deviation of any of them from $\mathbf{C}$ is 0.06748. A single entry has standard error $\sqrt{2/M} = 0.02236$, so this is about three standard errors, which is what the largest of 40 000 roughly Gaussian errors should be.

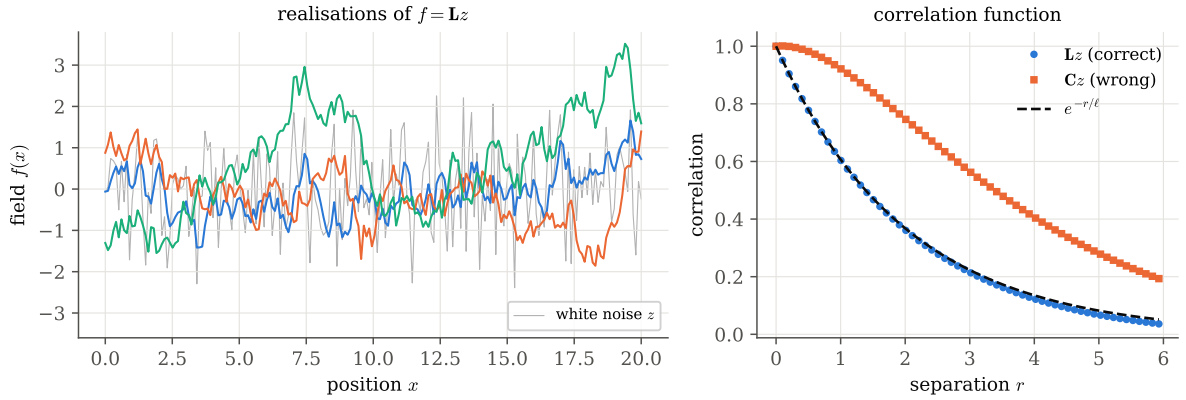


Figure 6.25: A correlated Gaussian field made from white noise. Left: three realisations of $\boldsymbol{f} = \mathbf{L}\boldsymbol{z}$ with correlation length $\ell = 2$, and one of the white-noise vectors $\boldsymbol{z}$ (grey) they are built from. Right: the correlation function averaged over 4000 realisations follows $e^{-r/\ell}$ (blue). Multiplying by $\mathbf{C}$ instead of its Cholesky factor gives correlations that are far too long (orange).

The orange points of Figure 6.25 show a tempting mistake: multiplying the white noise by $\mathbf{C}$ itself. By (6.36) with $\mathbf{L}$ replaced by $\mathbf{C}$, the result has covariance $\mathbf{C}\mathbf{C}^{\mathsf{T}} = \mathbf{C}^2$, not $\mathbf{C}$; its correlation at $r = \ell$ is 0.746 instead of 0.368.

Key idea: Colouring white noise

Any Gaussian vector with covariance $\mathbf{C}$ is a linear map of white noise: $\boldsymbol{X} = \boldsymbol{\mu} + \mathbf{L}\boldsymbol{z}$ with $\mathbf{L}\mathbf{L}^{\mathsf{T}} = \mathbf{C}$ and $\boldsymbol{z}$ a vector of independent standard normals made by Box–Muller. Conversely, $\mathbf{L}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$ is white again, which is how correlated residuals are "whitened" before a χ^2 is formed.

The Cholesky factor costs $n^3/3$ operations and n^2 memory. For a CMB map with 5×10^7 pixels that is hopeless. A statistically isotropic field offers a way round. Expand it in spherical harmonics; its harmonic coefficients $a_{\ell m}$ are then uncorrelated, with variances given by the power spectrum C_ℓ. So in the harmonic basis the covariance is diagonal, and its "Cholesky factor" is simply $\sqrt{C_\ell}$ [one independent Gaussian per mode, scaled by the square root of the power spectrum]. Simulating a CMB sky is therefore Box–Muller followed by a spherical-harmonic transform, which is how Chapter 7 and Chapter 8 make maps. Cholesky remains the tool for small, dense covariance matrices: the correlated residuals of a fit, a few correlated nuisance parameters, or a masked patch of sky.

6.10 Rejection sampling: throwing points under a curve

In the reaction $e^+e^- \to \mu^+\mu^-$ at energies well below the Z pole, the muons come out with the angular distribution

$$f(x) = \tfrac{3}{8}\,(1 + x^2), \qquad x = \cos\theta \in [-1, 1], \tag{6.38}$$

where θ is the angle between the outgoing μ^- and the incoming e^- [Cowan (3.8)]. To simulate events we need draws of x from this density. The inverse-CDF method of Section 6.8 sets the CDF equal to a uniform number u and solves for x. Here that means solving the cubic $F(x) = \frac{1}{8}(x^3 + 3x + 4) = u$ for every event. That can still be done. Add the forward–backward asymmetry, or multiply by a detector efficiency, and the CDF can no longer be inverted by hand. The posterior densities of Chapter 5 are worse still: we can compute their value at any point, but we know nothing about their CDF. So we need a method that uses only the *values* of f.

The question. If all we can do is evaluate $f(x)$, and draw from some simpler density g, how do we produce draws from f?

6.10.1 Points uniform under a curve

The key fact is geometric. Draw the graph of a density f and consider the region under it, $A_f = \{(x, y) : 0 < y < f(x)\}$. Its area is $\int f\,\mathrm{d}x = 1$. Now pick a point (X, Y) uniformly in that region, so its joint density is 1 on A_f and 0 elsewhere. The density of its horizontal coordinate alone is the marginal (Chapter 1), obtained by integrating out the vertical coordinate:

$$p_X(x) \overset{(1)}{=} \int p_{X,Y}(x, y)\,\mathrm{d}y \overset{(2)}{=} \int_0^{f(x)} 1\,\mathrm{d}y \overset{(3)}{=} f(x). \tag{6.39}$$

(1) The marginal density of X averages out the nuisance coordinate Y.
(2) The joint density is 1 for $0 < y < f(x)$ and zero otherwise.
(3) The length of the vertical slice of A_f at x is the height of the curve.

So *the x-coordinates of points spread uniformly under the graph of f are draws from f*. More sampled points fall where the curve is high, in proportion to its height. The same holds if we only know f up to a constant factor $c > 0$. The region under cf has area c, so a uniform point in it has joint

density $1/c$, and the x-marginal is $cf(x)/c = f(x)$ again. So we never need the normalising constant of f.

How do we get points uniform under f? Doing it directly is as hard as drawing from f itself. Instead, we make points uniform under a *larger* curve that we know how to sample, and keep only those that happen to fall under f. The kept points cover the region under f evenly, because every part of it was covered evenly before we threw anything away. So they are uniform under f. This is von Neumann's acceptance–rejection method [Cowan §3.3].

6.10.2 The algorithm and why it works

Choose an **envelope** density g [a density we can already sample, for instance by the inverse CDF] and a constant M with $f(x) \le Mg(x)$ for every x. Then repeat:

1. draw a candidate $V \sim g$ and, independently, $U \sim \text{Uniform}(0, 1)$;
2. if $U < f(V)/\big(Mg(V)\big)$, accept and return V; otherwise reject and go back to step 1.

In the picture of points under a curve (Section 6.10.1), step 1 puts the point $(V,\, U\,Mg(V))$ uniformly under the curve Mg: its horizontal position V has density g, and its height $U\,Mg(V)$ is uniform between 0 and $Mg(V)$. Step 2 keeps the point if it also lies under f, because $U\,Mg(V) < f(V)$ is the same condition as $U < f(V)/Mg(V)$.

The question. Why are the kept values exact draws from f, and not something only close to it? A kept value is a candidate V for which the event "accept" happened. So what we need is the CDF of V given acceptance, $\mathrm{P}(V \le y \mid \text{accept})$. By the definition of conditional probability this is $\mathrm{P}(V \le y,\ \text{accept})/\mathrm{P}(\text{accept})$, so we need two probabilities [CasellaBerger Thm 5.6.8].

First, how often is a candidate accepted?

$$\mathrm{P}(\text{accept}) \overset{(1)}{=} \int g(v)\ \mathrm{P}\Big(U < \frac{f(v)}{Mg(v)}\Big)\,\mathrm{d}v \overset{(2)}{=} \int g(v)\,\frac{f(v)}{Mg(v)}\,\mathrm{d}v \overset{(3)}{=} \frac{1}{M}\int f(v)\,\mathrm{d}v \overset{(4)}{=} \frac{1}{M}. \tag{6.40}$$

(1) Law of total probability over the value v of the candidate, which has density g (Chapter 1).
(2) U is uniform and $f(v)/Mg(v) \le 1$ by the choice of M, so the probability that U lies below it is the number itself.
(3) The g's cancel.
(4) f is a normalised density.

Second, how often is a candidate both accepted and below a value y? It is the same integral, stopped at y. *Then divide*, which gives the CDF we want:

$$\mathrm{P}(V \le y,\ \text{accept}) \overset{(1)}{=} \int_{-\infty}^{y} g(v)\,\frac{f(v)}{Mg(v)}\,\mathrm{d}v \overset{(2)}{=} \frac{F(y)}{M}, \qquad \mathrm{P}(V \le y \mid \text{accept}) \overset{(3)}{=} \frac{F(y)/M}{1/M} = F(y). \tag{6.41}$$

(1) As in (6.40), but only candidates below y count.
(2) The g's cancel, and $\int_{-\infty}^{y} f = F(y)$.
(3) The definition of conditional probability (Chapter 1), with (6.40) in the denominator.

What it means. The accepted values have CDF F, so they are exact draws from f, with no approximation. In the language of Chapter 1, accepting is *conditioning*. We generate candidates from a larger sample space and keep only the outcomes in the event "the point fell under f". Keeping only that event turns the distribution g of the candidates into the distribution f of the survivors.

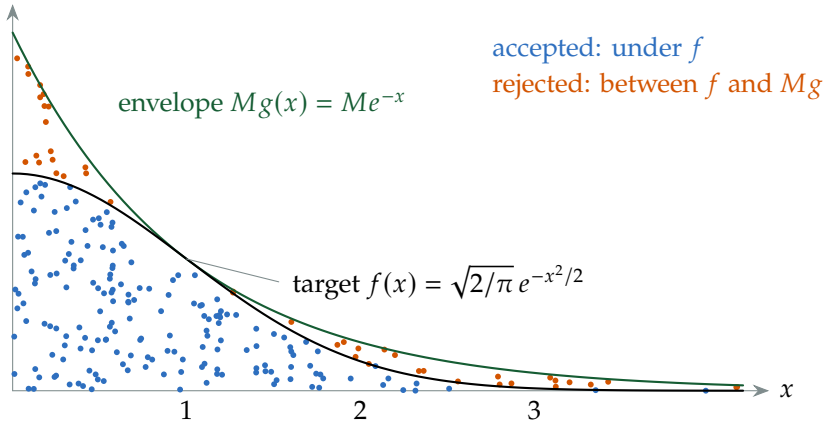


Figure 6.26: Rejection sampling as throwing points under a curve. Each candidate x is drawn from the exponential density g and given a height uniform between 0 and $Mg(x)$ (green), so the points are uniform under the green curve. The points that fall under the target f (black) are kept; their x-coordinates are distributed as f. The fraction kept is the area under f divided by the area under Mg, which is $1/M$. The two curves touch at $x = 1$, where f/g is largest.

The cost. Each candidate is accepted independently with probability $1/M$, by (6.40). So the number K of candidates needed for one accepted value is geometric, $\mathrm{P}(K = k) = (1-1/M)^{k-1}/M$: $k-1$ rejections followed by one acceptance. Its mean is $\mathbb{E}\,K = \sum_k k\,(1-1/M)^{k-1}/M = M$ [CasellaBerger (5.6.12)]. So M is the average cost of one draw, counted in candidates.

The best envelope. Since M is the cost, the best envelope is the one with the smallest M. Two choices achieve that. First, take $M = \sup_x f(x)/g(x)$ exactly: this is the smallest value that still keeps Mg above f everywhere. Second, choose g as close in shape to f as possible, so that this supremum is itself close to 1.

Working definition 6.4 (rejection sampling)

Target f (known up to a constant), envelope g that we can sample, and $M = \sup_x f(x)/g(x) < \infty$ [the factor that lifts g above f everywhere]. Draw $V \sim g$, $U \sim \mathrm{Uniform}(0,1)$; keep V if $U\,Mg(V) < f(V)$. *Cost:* on average M candidates per accepted draw; acceptance rate $1/M$. *Requirement:* f/g must be bounded, so g needs tails at least as heavy as f's.

6.10.3 Three worked examples

Example: the muon-pair angle in a box. The simplest envelope is a box. For (6.38) on $[-1,1]$, the largest value of f is $f_{\max} = f(\pm 1) = 3/4$. Take g uniform on $[-1,1]$, so $g = 1/2$ and $M = f_{\max}/g = 3/2$. The recipe becomes three steps: draw x uniform on $[-1,1]$, draw u uniform on $[0, f_{\max}]$, and keep x if $u < f(x)$ [Cowan §3.3]. The predicted acceptance is $1/M = 2/3$: the area under f (one) divided by the area of the box ($2 \times 3/4$). Out of 200000 candidates the script accepts a fraction 0.6666. To check the shape as well, compare the mean of x^2 over accepted values, 0.399, with its exact value, $\int_{-1}^{1} x^2 \cdot \frac{3}{8}(1+x^2)\,\mathrm{d}x = \frac{3}{8}\left(\frac{2}{3} + \frac{2}{5}\right) = \frac{2}{5}$. Figure 6.27 (left and middle) reproduces Cowan's Figure 3.2.

Example: a half-normal under an exponential. Take $f(x) = \sqrt{2/\pi}\, e^{-x^2/2}$ for $x > 0$, the distribution of $|Z|$ for a standard normal Z, and the envelope $g(x) = e^{-x}$, which we sample as

$-\ln(1-u)$. The ratio and its maximum:

$$\frac{f(x)}{g(x)} \overset{(1)}{=} \sqrt{\frac{2}{\pi}}\, e^{\,x-x^2/2} \overset{(2)}{=} \sqrt{\frac{2}{\pi}}\, e^{1/2}\, e^{-(x-1)^2/2} \quad\Longrightarrow\quad M = \sup_x \frac{f}{g} \overset{(3)}{=} \sqrt{\frac{2e}{\pi}} = 1.315. \tag{6.42}$$

(1) Divide the two densities: $e^{-x^2/2}/e^{-x} = e^{x-x^2/2}$.
(2) Complete the square: $x - \frac{1}{2}x^2 = \frac{1}{2} - \frac{1}{2}(x-1)^2$.
(3) The Gaussian factor is at most 1, reached at $x = 1$, where the two curves touch (Figure 6.26).

So a fraction $1/M = 0.7602$ of candidates should survive; the script finds 0.7601. The number of candidates per accepted value averages 1.316, against M, and its histogram is the geometric distribution (Figure 6.28). The accepted values have mean 0.7978, against $\mathbb{E}\,|Z| = \sqrt{2/\pi} = 0.7979$. Attaching a random sign to each accepted value gives a full standard normal. That is a third Gaussian generator, after Box–Muller and the polar method.

Example: a Beta density with non-integer parameters. Take $Y \sim \text{Beta}(a, b)$ with $a = 2.7$ and $b = 6.3$. The recipes of Section 6.9 that build gamma and beta variables from logarithms of uniforms need integer parameters, so they do not apply here, and there is no other simple transformation from uniforms. Rejection needs none: put the density on $[0, 1]$ inside a box of width 1 and height $c = \max_y f_Y(y)$ [CasellaBerger Ex. 5.6.7]. Here g is uniform on $[0, 1]$, so $M = c$. The maximum is at the mode $y = (a-1)/(a+b-2) = 1.7/7$, where $c = 2.67$. So the acceptance should be $1/c = 0.3746$; the script finds 0.3729. The mean of accepted values is 0.3, against $a/(a+b) = 0.3$.

Can a better envelope help? A Beta(2, 6) envelope is closer in shape to the target, and needs only $M = 1.67$ candidates per draw instead of 2.67. But each Beta(2, 6) candidate itself costs several uniforms. So which algorithm is faster depends on the cost of a candidate as well as on M [CasellaBerger Ex. 5.6.9].

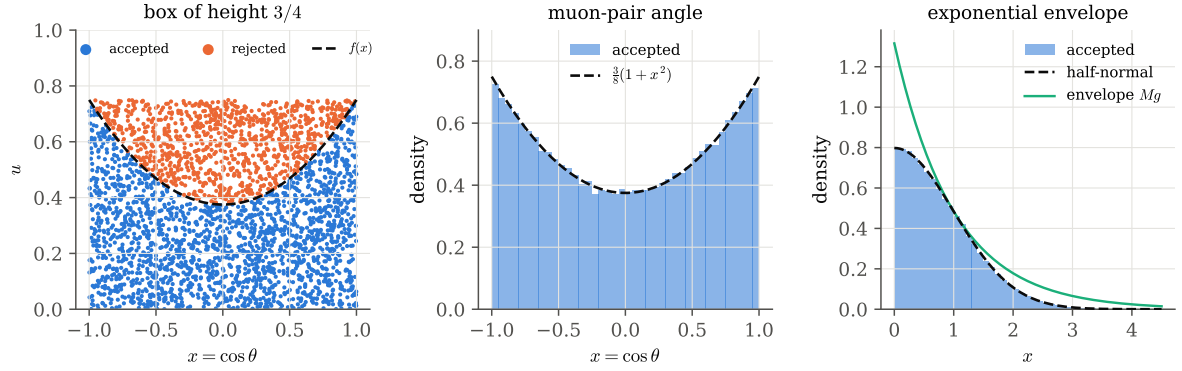


Figure 6.27: Rejection sampling at work. Left: 3000 candidate points uniform in the box of height 3/4; those under $f(x) = \frac{3}{8}(1+x^2)$ (blue) are kept. Middle: the histogram of all kept $x = \cos\theta$ values follows the muon-pair angular distribution. Right: half-normal values kept from exponential candidates, with the envelope Mg (green) touching the target at $x = 1$.

Listing 6.11: Half-normal draws from exponential candidates (`code/ch06/24_rejection.py`)

```
# ------------------------------------------------------------ 2. half-normal under an exponential
M2 = np.sqrt(2.0 * np.e / np.pi)
v = -np.log(1.0 - rng.random(N))                # candidate from g(x) = e^{-x} (inverse CDF)
u = rng.random(N)
ratio = np.sqrt(2.0 / np.pi) * np.exp(-v**2 / 2) / (M2 * np.exp(-v))
acc2 = u < ratio
```

```
45 hn = v[acc2]
```

Line 41 draws the candidates by the inverse CDF of Section 6.8, line 43 computes the acceptance probability $f(V)/Mg(V)$ for all of them at once, and line 44 keeps the candidates whose uniform number falls below it. The loop "repeat until accepted" of the algorithm has become a filter applied to a fixed number of candidates at once. This is how rejection is usually coded in numpy. The price is that N candidates give about N/M accepted values, not exactly N.

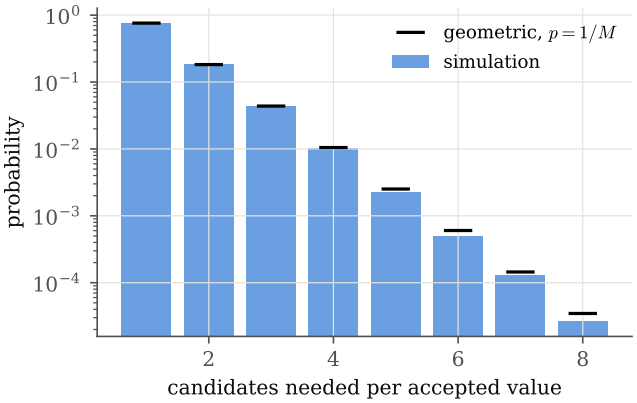


Figure 6.28: How many exponential candidates it takes to get one half-normal value. The frequencies (bars, logarithmic scale) fall by a factor $1 - 1/M$ per extra candidate, the geometric distribution with success probability $1/M$ (black ticks).

6.10.4 When rejection fails: tails and dimension

Rejection fails in two ways. The first is in the tails. The method needs a finite $M = \sup f/g$, so that Mg lies above f everywhere. If the target has heavier tails than the envelope, then far out f is much larger than g, the ratio f/g grows without bound, and no finite M exists. A Cauchy envelope for a Gaussian target works, since the Cauchy tails, falling as x^{-2}, outlast $e^{-x^2/2}$. A Gaussian envelope for a Cauchy target cannot work, however large we make M [CasellaBerger p. 207]. The same condition on the tails comes back, in a sharper form, for importance sampling in Section 6.12.

The second failure is in high dimension, and it is worse because nothing visibly breaks: the acceptance rate just collapses. In d dimensions the envelope has to cover the target in every direction at once, and the wasted volume multiplies. The cleanest example is the uniform distribution in the unit ball, sampled from the cube $[-1, 1]^d$ around it. The acceptance is the volume of the ball over the volume of the cube:

$$\frac{1}{M} \overset{(1)}{=} \frac{\pi^{d/2}/\Gamma(d/2+1)}{2^d}, \qquad \text{so} \qquad \frac{1}{M} = \frac{\pi}{4} = 0.785 \ \text{ for } d = 2,$$
$$\frac{1}{M} \overset{(2)}{=} \frac{\pi^5}{5!\,2^{10}} = 0.0025 \ \text{ for } d = 10, \qquad \frac{1}{M} \overset{(3)}{=} \frac{\pi^{10}}{10!\,2^{20}} = 2.5 \times 10^{-8} \ \text{ for } d = 20. \tag{6.43}$$

(1) The volume of the unit ball in d dimensions is $\pi^{d/2}/\Gamma(d/2+1)$; the cube of side 2 has volume 2^d.
(2) $\Gamma(6) = 5! = 120$, so $306.0/(120 \times 1024)$.
(3) $\Gamma(11) = 10! = 3\,628\,800$, and $\pi^{10} = 93\,648$.

In two dimensions (the polar method of Section 6.9.1) we keep 79% of the candidates. In twenty dimensions we keep one candidate in forty million. A cosmological posterior with twenty

parameters is far more concentrated inside its prior than a ball is inside a cube, so rejection from the prior is hopeless there. This is the practical reason for the Markov chain methods of Chapter 9, which never try to cover the whole space at once.

Rejection also hides inside some recipes that do not use the word. In a dose–response study, ten response probabilities $p_1, \dots, p_{10}$, one per dose, each have an independent Beta posterior, but physics says they must be ordered, $p_1 \le \cdots \le p_{10}$. The recipe is: draw all ten, and throw the draw away if it is not ordered [Wasserman Ex. 24.4, p. 407]. That is rejection sampling. By exactly the argument of (6.41), the kept draws come from the posterior restricted to the ordered region, and the acceptance rate is the posterior probability of that region.

Key idea: Accepting is conditioning

Points uniform under the graph of a density have x-coordinates distributed as that density. Rejection sampling makes points uniform under an easy envelope Mg and keeps those under f; the survivors are exact draws from f, at an average cost of M candidates each. It needs only values of f up to a constant, and it breaks down when f/g is unbounded or when the dimension is large.

6.11 Monte Carlo integration and the $1/\sqrt{N}$ law

Almost every number we want out of a probability model is an integral. The posterior mean of a parameter is $\int \theta\, p(\theta \mid \boldsymbol{d})\, \mathrm{d}\theta$. The marginal posterior of one cosmological parameter is an integral over the other five. The normalising constant of Bayes' theorem, the evidence, is an integral of likelihood times prior over the whole parameter space [Wasserman §24.1, p. 403], [Lambert §12.3]. The acceptance of a detector is an integral of the emission density over the directions that hit it. Most of these integrals have no closed form. Even a one-parameter model with a Poisson likelihood and a log-normal prior gives one, and each added parameter makes it harder [Lambert §12.3]. Grids and quadrature rules work in one or two dimensions. Beyond that the number of grid points, n^D for n points per axis in D dimensions, becomes too large to evaluate [Lambert §12.4–12.5].

The samplers built so far (the inverse CDF, Box–Muller and rejection) give another way, resting on two facts: an average over random draws estimates an expectation, and every integral can be written as an expectation. Lambert tells the first fact as a story about a box with a die inside. We cannot see the die, but we can shake the box and read the face through a small window. The average of many readings estimates the mean of the die, without our knowing how many faces it has or how it is loaded [Lambert §12.6].

The question. If we estimate an integral by averaging a function over N random draws, how large is the error, and how does it depend on N and on the dimension?

6.11.1 An integral is an expectation

Start with a one-dimensional integral $I = \int_a^b h(x)\, \mathrm{d}x$. Multiply and divide by $b - a$ [Wasserman (24.1)]:

$$I \overset{(1)}{=} \int_a^b \underbrace{h(x)\,(b-a)}_{w(x)}\, \underbrace{\frac{1}{b-a}}_{f(x)}\, \mathrm{d}x \overset{(2)}{=} \mathbb{E}_f\big[w(X)\big], \qquad X \sim \mathrm{Uniform}(a,b). \tag{6.44}$$

(1) Nothing has changed: we multiplied and divided by the same constant.

(2) $f(x) = 1/(b-a)$ is the uniform density on $[a, b]$, and an integral of a function against a density is the expectation of that function.

More generally, any integral of the form $I = \int h(x)f(x)\,dx$ with f a density is $\mathbb{E}_f[h(X)]$ [Wasserman (24.3)], and the variable x can have any number of dimensions. Draw $X_1, \dots, X_N$ independently from f and average:

$$\hat{I} = \frac{1}{N}\sum_{i=1}^{N} h(X_i). \tag{6.45}$$

By the law of large numbers (Chapter 3), $\hat{I} \to I$ as $N \to \infty$ [Wasserman (24.2)], [CasellaBerger (5.6.4)].

Working definition 6.5 (Monte Carlo estimate of an integral)

Write the integral as an expectation, $I = \mathbb{E}_f[h(X)]$ [choose a density f we can sample, and put everything else into h]. Draw N independent $X_i \sim f$ and report

$$\hat{I} = \frac{1}{N}\sum_i h(X_i) \;\pm\; \frac{s}{\sqrt{N}}, \qquad s^2 = \frac{1}{N-1}\sum_i \big(h(X_i) - \hat{I}\big)^2.$$

The error bar is itself estimated from the same draws.

6.11.2 Why the error falls as $1/\sqrt{N}$

The estimator $\hat{I}$ of (6.45) is an average of N independent terms $h(X_i)$. How far is it from I? Call $\mu = \mathbb{E}_f[h(X)] = I$ and $\sigma^2 = \mathrm{Var}_f[h(X)]$ the mean and variance of a single term. The mean and the variance of the average then follow in two lines each:

$$\mathbb{E}[\hat{I}] \overset{(1)}{=} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[h(X_i)] \overset{(2)}{=} I,$$

$$\mathrm{Var}[\hat{I}] \overset{(3)}{=} \frac{1}{N^2}\sum_{i=1}^{N} \mathrm{Var}[h(X_i)] \overset{(4)}{=} \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N}. \tag{6.46}$$

(1) Expectation is linear.
(2) Each X_i has density f, so each term has mean I.
(3) The X_i are independent, so the variance of the sum is the sum of the variances, and a constant factor $1/N$ comes out squared.
(4) Each term has variance σ^2.

So $\hat{I}$ is unbiased, and its standard error, the square root of its variance, is $\sigma/\sqrt{N}$. This is the same law that made the error of a sample mean shrink in Chapter 3. The central limit theorem (Chapter 2) adds the shape: for large N, $\hat{I}$ is approximately Gaussian, $\hat{I} \approx \mathcal{N}(I, \sigma^2/N)$. We do not know σ, so we replace it by s, the standard deviation of the N values $h(X_i)$, $s^2 = \frac{1}{N-1}\sum_i \big(h(X_i) - \hat{I}\big)^2$. Then $\hat{I} \pm 1.96\, s/\sqrt{N}$ is an approximate 95% confidence interval: in 95% of repeated Monte Carlo runs it covers the true I [Wasserman p. 405]. The CLT needs $\sigma^2 < \infty$; Section 6.12 shows what happens when it is not.

Two facts about $\sigma/\sqrt{N}$ matter most. First, the factor $1/\sqrt{N}$ is slow: to gain one decimal digit we need a hundred times more draws. Second, nothing in (6.46) refers to the dimension of x. A thousand-dimensional integral has error $\sigma/\sqrt{N}$ exactly like a one-dimensional one; only σ can depend on the problem.

Key idea: The Monte Carlo error law

The average of h over N independent draws from f estimates $\int hf$ with standard error $\sigma/\sqrt{N}$, where $\sigma^2 = \mathrm{Var}_f[h]$. The rate $N^{-1/2}$ does not depend on the dimension; the constant σ does depend on how well f matches where h is large, and every variance-reduction method works on σ.

Example: $\int_0^1 x^3\,\mathrm{d}x$. Here f is uniform on $[0,1]$ and $h(x) = x^3$, so $I = 1/4$ and $\sigma^2 = \mathbb{E}[U^6] - (\mathbb{E}[U^3])^2 = \frac{1}{7} - \frac{1}{16} = \frac{9}{112}$, that is $\sigma = 0.2835$. With $N = 10000$ uniforms the script finds $\hat{I} = 0.2509 \pm 0.002818$, where the error bar is $s/\sqrt{N}$; the prediction is $\sigma/\sqrt{N} = 0.0028$, and Wasserman reports 0.248 ± 0.0028 [Wasserman Ex. 24.1]. To see the law itself, the script repeats the estimate 2000 times at each of thirteen values of N from 10 to 10^5 and measures the root-mean-square error. On a log–log plot the points fall on a line of slope -0.5014 (Figure 6.29, left), and at $N = 10^5$ the measured RMS error is 0.9852 times $\sigma/\sqrt{N}$.

Example: a probability is the mean of an indicator. To compute $\Phi(2) = \mathrm{P}(Z \le 2)$ for $Z \sim \mathcal{N}(0,1)$, write it as $\int h(s)\,\varphi(s)\,\mathrm{d}s$ with φ the standard normal density and $h(s) = 1$ if $s \le 2$, 0 otherwise [Wasserman Ex. 24.2]. The Monte Carlo estimate is then the fraction of normal draws below 2. Each term is a Bernoulli variable with $\sigma^2 = p(1-p)$, so the standard error is the binomial $\sqrt{p(1-p)/N}$. The script gives 0.9761 ± 0.001527 for $N = 10^4$ and $0.9775 \pm 4.691 \times 10^{-4}$ for $N = 10^5$, against 0.9772. Every probability of an event can be computed this way: simulate the experiment and count. Casella and Berger's example asks for the probability that at least t of c electronic components with exponential lifetimes survive h hours; the Monte Carlo answer is the fraction of simulated batches in which they do [CasellaBerger Ex. 5.6.2].

6.11.3 Grids lose in high dimension

A grid rule beats $1/\sqrt{N}$ easily in one dimension, and loses that advantage as the dimension grows. Take the midpoint rule with spacing Δ: it replaces h on each cell by its value at the cell's midpoint. For a smooth integrand the midpoint approximation has an error of order Δ^2. To see where this power comes from, expand h about the midpoint. The linear term is odd and integrates to zero over the cell, leaving the quadratic term as the first contribution to the error. Now spend N points in d dimensions. A grid gives only $n = N^{1/d}$ points per axis, so

$$\text{grid error} \;\propto\; \Delta^2 \overset{(1)}{=} \left(\frac{1}{n}\right)^2 \overset{(2)}{=} N^{-2/d}, \qquad \text{Monte Carlo error} \;=\; \sigma\, N^{-1/2}. \tag{6.47}$$

(1) On the unit cube the spacing is $\Delta = 1/n$.
(2) $N = n^d$ points on a grid with n points per axis.

The two rates are equal when $2/d = 1/2$, at $d = 4$. Below four dimensions the grid wins; above it Monte Carlo wins, and by a margin that grows with d [Cowan §3.4]. (Gaussian quadrature or Simpson's rule raise the power 2, which moves the crossover but not the conclusion; the grid still needs n^d points.)

Example: the same integral in 1, 2, 4 and 8 dimensions. On the unit cube take the product

$$h(\boldsymbol{x}) = \prod_{i=1}^{d} \tfrac{\pi}{2}\sin(\pi x_i).$$

Each factor integrates to one, so $I_d = 1$ in every dimension. For Monte Carlo with uniform draws,

$$\sigma_d^2 \overset{(1)}{=} \mathbb{E}[h^2] - I_d^2 \overset{(2)}{=} \prod_{i=1}^{d} \mathbb{E}\left[\tfrac{\pi^2}{4}\sin^2(\pi U_i)\right] - 1 \overset{(3)}{=} \left(\frac{\pi^2}{8}\right)^d - 1. \tag{6.48}$$

(1) The variance is the mean square minus the square of the mean.
(2) The coordinates are independent, so the expectation of the product is the product of the expectations.
(3) $\mathbb{E}[\sin^2(\pi U)] = \frac{1}{2}$ for U uniform on $[0,1]$.

So $\sigma_1 = 0.4834$ and $\sigma_8 = 2.09$. Figure 6.29 (right) compares the two methods for the same number of function evaluations. With about 10^6 points, the grid error is 1.028×10^{-13} in one dimension, 4.114×10^{-7} in two, 0.002437 in four (456976 points) and 0.09592 in eight ($6^8 = 1679616$ points), while the Monte Carlo errors are 3.418×10^{-4}, 5.110×10^{-4}, 0.001697 and 0.001612. In eight dimensions a grid of 1.7 million points has only six points per axis, and it is sixty times worse than the same number of random points.

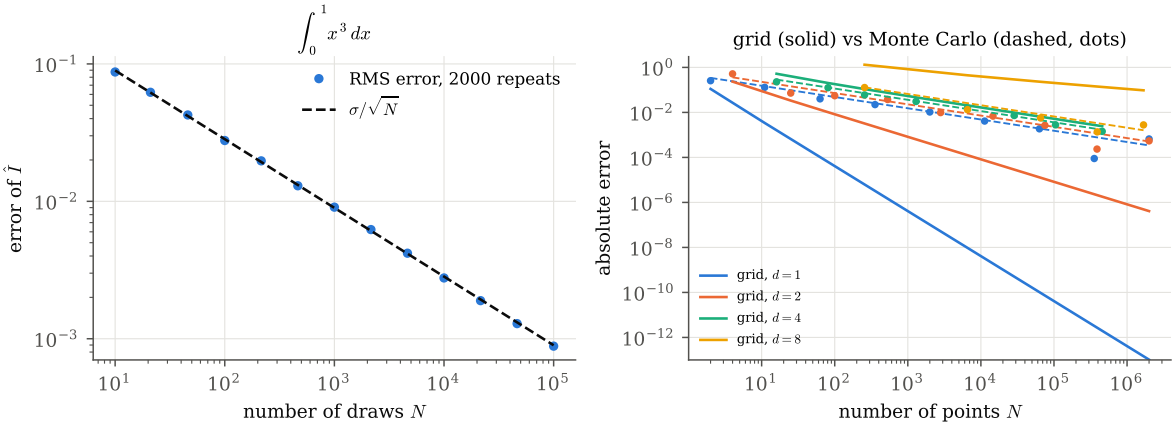


Figure 6.29: The $1/\sqrt{N}$ law. Left: RMS error of the Monte Carlo estimate of $\int_0^1 x^3 \mathrm{d}x$ over 2000 repeats, against $\sigma/\sqrt{N}$ (black dashed). Right: the error for $\int \prod \frac{\pi}{2}\sin(\pi x_i)\,\mathrm{d}^d x = 1$ with N points. Solid lines: midpoint grid, error $\propto N^{-2/d}$. Dashed lines and dots: Monte Carlo, error $\sigma_d/\sqrt{N}$ with slope $-1/2$ in every dimension. The grid wins for $d = 1, 2$, roughly ties at $d = 4$ and loses badly at $d = 8$.

Watch out: Dimension-independent is not dimension-free

The *rate* $N^{-1/2}$ does not depend on d, but σ_d can grow exponentially with d, as (6.48) shows: $(\pi^2/8)^{d/2}$. It grows much faster when the integrand is concentrated in a small part of the region we sample. The evidence estimated by averaging the likelihood over prior draws was exactly that failure (Chapter 5). Monte Carlo beats the curse of dimensionality only when the sampling density f follows the integrand, which is what importance sampling and MCMC are for.

Example: a posterior by simulation. Integrals over posteriors are where Monte Carlo pays most. Suppose $X \sim \text{Binomial}(n, p_1)$ and $Y \sim \text{Binomial}(m, p_2)$ are the numbers of events passing a selection cut in two runs, out of n and m events. We want the difference of the two efficiencies, $\delta = p_2 - p_1$ [Wasserman Ex. 24.3]. With flat priors the posterior factorises into $p_1 \mid X \sim \text{Beta}(X+1, n-X+1)$ and $p_2 \mid Y \sim \text{Beta}(Y+1, m-Y+1)$ (Chapter 5). So we can draw

(p_1, p_2) pairs exactly, form $\delta = p_2 - p_1$ for each pair, and make a histogram. The posterior of δ is a double integral, and we never have to write it down.

For $n = m = 10$, $X = 8$, $Y = 6$, 100000 draws give:

- a posterior mean $-0.1667 \pm 5.748 \times 10^{-4}$ (exact: $\frac{7}{12} - \frac{9}{12} = -\frac{1}{6}$);
- a 95% posterior interval $[-0.5134,\ 0.1967]$, from the 2.5% and 97.5% quantiles of the draws (Wasserman: $(-0.52, 0.20)$ from 1000 draws);
- a posterior probability $\mathrm{P}(\delta < 0 \mid \boldsymbol{d}) = 0.8196$ that the second run has the lower efficiency (Figure 6.30).

Each of these is a summary of the draws standing in for the same summary of the posterior. A quantile of the draws estimates a quantile of the posterior. The fraction of draws in a region estimates the probability of the region. The histogram estimates the density, which is the subject of Section 6.13.

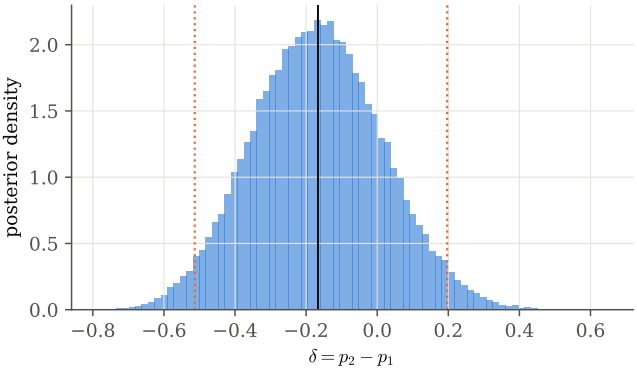


Figure 6.30: Posterior of the difference $\delta = p_2 - p_1$ of two binomial probabilities, from 10^5 pairs of Beta draws. The solid line is the posterior mean, the dotted lines the 2.5% and 97.5% quantiles of the draws.

6.11.4 Example: the acceptance of a detector

A radioactive source sits at distance d in front of a flat square scintillator of side a, centred on the axis. It emits isotropically, that is, equally in all directions. What fraction of the decays hits the scintillator? The answer is the **geometric acceptance** $\epsilon = \Omega/4\pi$ [the solid angle subtended by the detector as a fraction of the full sphere]. In the Monte Carlo form it is the mean of an indicator: emit N particles in isotropic directions (Section 6.8.2: $\cos\theta = 2u_1 - 1$, $\phi = 2\pi u_2$), follow each straight line to the plane $z = d$, and count the fraction that lands inside the square.

Verification. For the plain square the solid angle can be computed. With the detector in the plane $z = d$, a surface element $\mathrm{d}x\,\mathrm{d}y$ at distance $\rho = \sqrt{x^2 + y^2 + d^2}$ subtends $\mathrm{d}\Omega = d\,\mathrm{d}x\,\mathrm{d}y/\rho^3$ (area times the cosine of the angle of incidence d/ρ, divided by ρ^2). Doing the y integral and then the x integral with the standard antiderivatives gives

$$\Omega \overset{(1)}{=} \int_{-a/2}^{a/2}\int_{-a/2}^{a/2} \frac{d\,\mathrm{d}y\,\mathrm{d}x}{(x^2+y^2+d^2)^{3/2}} \overset{(2)}{=} \int_{-a/2}^{a/2} \frac{a\,d\,\mathrm{d}x}{(x^2+d^2)\sqrt{x^2+d^2+a^2/4}}$$
$$\overset{(3)}{=} 4\arctan\frac{a^2}{2d\sqrt{2a^2+4d^2}} \overset{(4)}{=} 4\arcsin\frac{a^2}{a^2+4d^2}. \tag{6.49}$$

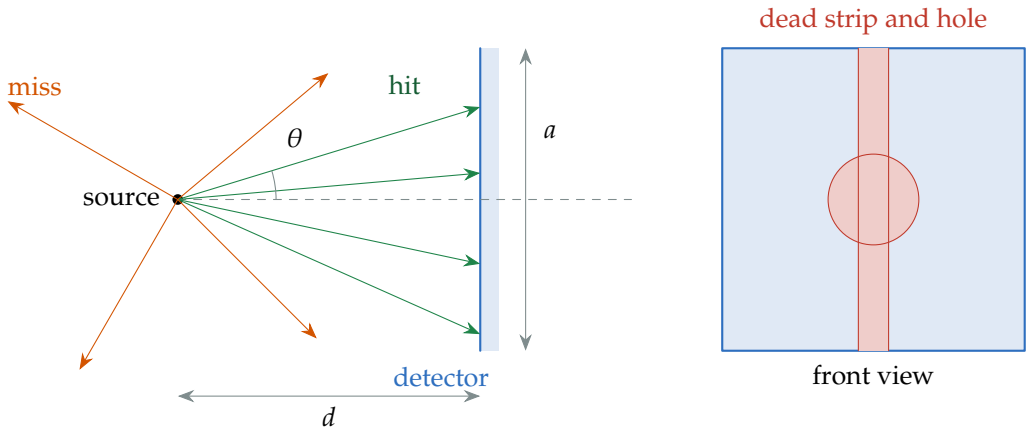


Figure 6.31: Left: side view of a point source at distance d from a square detector of side a. Isotropic rays either hit the detector (green) or miss it (orange); the acceptance is the fraction that hits. Right: the same detector seen from the source, with an insensitive strip and a central hole, the kind of geometry for which no closed form exists.

(1) The solid angle is the integral of $\mathrm{d}\Omega$ over the square.

(2) With $q^2 = x^2 + d^2$: $\int_{-a/2}^{a/2} \mathrm{d}y/(q^2+y^2)^{3/2} = \left[y/(q^2\sqrt{q^2+y^2})\right]_{-a/2}^{a/2} = a/\left(q^2\sqrt{q^2+a^2/4}\right)$.

(3) With $p = d$ and $r^2 = d^2 + a^2/4$, use the antiderivative (check it by differentiating)

$$\int \frac{\mathrm{d}x}{(x^2+p^2)\sqrt{x^2+r^2}} = \frac{1}{p\sqrt{r^2-p^2}} \arctan \frac{x\sqrt{r^2-p^2}}{p\sqrt{x^2+r^2}},$$

with $\sqrt{r^2-p^2} = a/2$. The two limits contribute equally.

(4) If $\tan\alpha = T$ then $\sin\alpha = T/\sqrt{1+T^2}$; here $1+T^2 = (a^2+4d^2)^2/\left(4d^2(2a^2+4d^2)\right)$.

For $d = a/2$ the square is one face of a cube centred on the source, so by symmetry it must catch exactly $1/6$ of the particles; indeed (6.49) gives $4\arcsin\frac{1}{2} = 2\pi/3 = 4\pi/6$. With $a = 10\,\mathrm{cm}$ and $N = 1000000$ particles the Monte Carlo acceptance is $0.1668 \pm 3.728\times10^{-4}$ at $d = 5\,\mathrm{cm}$ (exact $1/6$, pull 0.473), $0.06375 \pm 2.443\times10^{-4}$ at $d = 10\,\mathrm{cm}$ (exact 0.06409, pull -1.425) and $0.01875 \pm 1.356\times10^{-4}$ at $d = 20\,\mathrm{cm}$ (exact 0.01873). The error bars are binomial, $\sqrt{\epsilon(1-\epsilon)/N}$, and the pulls [the difference from the exact value in units of the error bar] are of order one, as they should be.

Listing 6.12: Isotropic emission and the crossing point with the detector plane (`code/ch06/26_acceptance.py`)

```
33 def isotropic(n):
34     cth = 2.0 * rng.random(n) - 1.0            # cos(theta) uniform on [-1, 1]
35     phi = 2.0 * np.pi * rng.random(n)          # phi uniform on [0, 2 pi)
36     return cth, phi
37
38
39 def hits(cth, phi, a, d):
40     """Crossing point with the plane z = d for upward particles (NaN otherwise)."""
41     up = cth > 0
42     sth = np.sqrt(1.0 - cth**2)
43     with np.errstate(divide="ignore", invalid="ignore"):
44         rho = np.where(up, d * sth / cth, np.nan)
45     return rho * np.cos(phi), rho * np.sin(phi)
```

Lines 34–35 are the isotropic directions of Section 6.8.2. Lines 41–45 follow each upward-going particle to the plane $z = d$: it travels a transverse distance $d\tan\theta$ before reaching it, in

the azimuthal direction ϕ.

Research. Now put a 1 cm insensitive strip down the middle of the scintillator (a gap between two tiles) and a hole of radius 1.5 cm for the beam pipe (Figure 6.31, right). No formula is at hand, but the Monte Carlo number is as easy as before: at $d = 5$ cm the live acceptance is $0.1321 \pm 3.386 \times 10^{-4}$, a fraction 0.7925 of the full square's. The same code with realistic geometry, a position-dependent efficiency, multiple scattering and energy thresholds is how every experiment computes its acceptance (Figure 6.32).

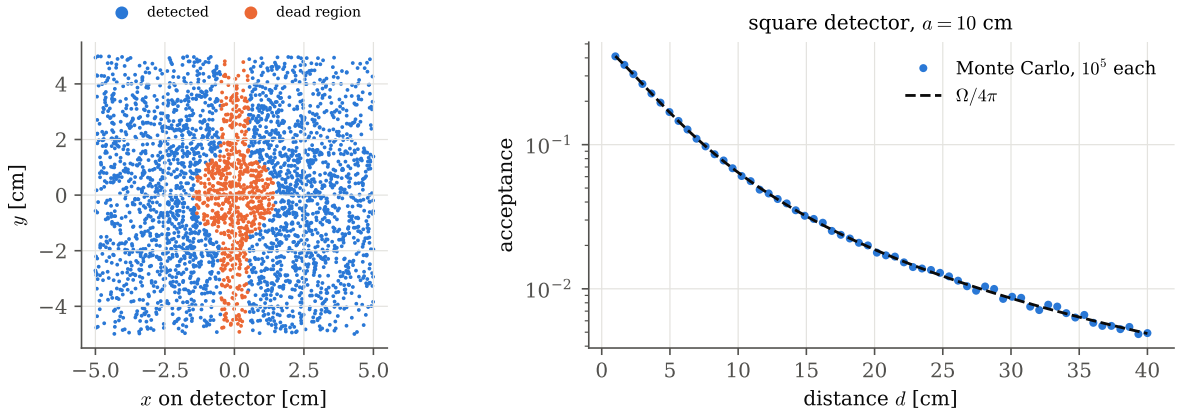


Figure 6.32: Detector acceptance by Monte Carlo. Left: where the particles that reach the square cross it, for $d = 5$ cm; those in the dead strip and the central hole (orange) are lost. Right: acceptance of the plain square against distance, 10^5 particles per point, on the exact $\Omega/4\pi$ of (6.49).

How this is used in real research: Why physics runs on sampling

The two uses of the acceptance calculation are the two uses of Monte Carlo in research. Against (6.49) it *verifies*: the code, the random numbers and the error bars are checked against a known answer. With the dead strip it *is the answer*: no formula exists and the simulation is the instrument. A collider analysis chains the two stages Cowan describes [Cowan §3.4]. An **event generator** such as Pythia [161] samples the four-momenta of the final-state particles from the differential cross section: hard scattering, parton showers and hadronisation, each by inverse-CDF, rejection and veto steps of exactly the kinds in this chapter. A **detector simulation** such as Geant4 [162] then samples each particle's interactions with the material: ionisation, multiple scattering, showers. The output is treated like real data: it gives the acceptance and efficiency, the response matrix for unfolding, and the distribution of a test statistic under the background-only hypothesis (Chapter 4). In cosmology the same role is played by simulated skies pushed through the mask, beam and noise of the instrument (Chapter 8). In the pipeline from physical model to inference, sampling is how we *derive the sampling distribution of a summary* whenever the algebra gives out.

6.12 Importance sampling, and why g needs thicker tails than f

A search for a rare background fluctuation needs the probability that a standard normal test statistic exceeds 3, $\mathrm{P}(Z > 3) = 0.00135$. Estimated by plain Monte Carlo with $N = 100$ draws, it is the fraction of draws above 3 (Section 6.11.2). Most runs of 100 draws contain no such draw at all: in 20000 repeated runs, a fraction 0.8739 returned exactly zero. The integrand is the

indicator $h(x) = \mathbf{1}[x > 3]$, so only draws above 3 contribute. A run in which every draw falls below the threshold therefore estimates the tail probability as zero. So the trouble is not the factor $1/\sqrt{N}$ but the constant σ in front of it: we sampled from a density that rarely visits the region that matters.

A second problem has the same shape. In Bayesian inference the density we need to average over, the posterior, is usually one we cannot sample at all, while a simpler density that resembles it is easy to sample [Wasserman §24.3]. Both problems have the same fix: draw from a density of our choosing, and correct each draw with a weight.

The question. If we draw from a density g instead of f, how do we still estimate $\int hf$, and which g make the estimate better or worse?

6.12.1 Reweighting

Let g be any density that is positive wherever $h(x)f(x) \neq 0$. Multiply and divide by it [Wasserman (24.4)]:

$$I = \int h(x)f(x)\,\mathrm{d}x \overset{(1)}{=} \int \frac{h(x)f(x)}{g(x)}\,g(x)\,\mathrm{d}x \overset{(2)}{=} \mathbb{E}_g\Big[h(X)\,w(X)\Big], \qquad w(x) \equiv \frac{f(x)}{g(x)}. \tag{6.50}$$

(1) Multiplying and dividing by $g(x)$ changes nothing where $g > 0$; where $g = 0$ the integrand hf vanishes by assumption, so no part of the integral is lost.
(2) An integral against the density g is an expectation under g.

So draw $X_1, \ldots, X_N \sim g$ and average the weighted values [Wasserman (24.5)]:

$$\hat{I}_g = \frac{1}{N}\sum_{i=1}^{N} h(X_i)\,w(X_i). \tag{6.51}$$

The **importance weight** $w = f/g$ [how much more often f would have produced this value than g did] corrects for sampling from the wrong density: values that g over-produces get weights below one, values it under-produces get weights above one. By (6.46) applied to the terms hw under g, $\hat{I}_g$ is unbiased and its variance is $\mathrm{Var}_g[hw]/N$, with [Wasserman (24.6)]

$$\mathrm{Var}_g[hw] \overset{(1)}{=} \mathbb{E}_g\big[h^2w^2\big] - I^2 \overset{(2)}{=} \int \frac{h^2(x)f^2(x)}{g^2(x)}\,g(x)\,\mathrm{d}x - I^2 \overset{(3)}{=} \int \frac{h^2(x)f^2(x)}{g(x)}\,\mathrm{d}x - I^2. \tag{6.52}$$

(1) Variance is the mean square minus the square of the mean, and the mean of hw under g is I by (6.50).
(2) Write the expectation under g as an integral, with $w = f/g$.
(3) One factor of g cancels.

Everything about the quality of importance sampling is in the last integral, and the key feature is that g sits in its *denominator*. Where g is much smaller than f (and $h \neq 0$), the integrand h^2f^2/g is large. If g falls off faster than f in the tails, the integrand grows there instead of dying away, and the integral diverges.

Working definition 6.6 (importance sampling)

To estimate $I = \int hf$, draw $X_i \sim g$ and average $h(X_i)\,f(X_i)/g(X_i)$ [each draw weighted by how much likelier it is under f than under g]. *Unbiased* whenever $g > 0$ where $hf \neq 0$.

Standard error $\sqrt{(\int h^2 f^2/g - I^2)/N}$, estimated by the sample standard deviation of the $h_i w_i$ over $\sqrt{N}$ when it is finite.

6.12.2 The tail rule, shown

We can see the divergence in the simplest possible case. Let the target be the standard normal, $f = \mathcal{N}(0,1)$, and estimate its normalisation: take $h = 1$, so $I = \int f = 1$. Put another way, if we knew only the unnormalised curve $e^{-x^2/2}$, we would be estimating its area $Z = \sqrt{2\pi}$, and the ratio $\hat{Z}/Z$ is the same number as $\hat{I}$. This is the evidence problem of Chapter 5 in miniature. Sample from $g = \mathcal{N}(0, s^2)$, a Gaussian of adjustable width s. With $h = 1$ the second moment of the terms in (6.52) is the mean square weight,

$$\mathbb{E}_g[w^2] = \int \frac{f^2}{g}\,\mathrm{d}x \overset{(1)}{=} \int \frac{e^{-x^2}/2\pi}{e^{-x^2/2s^2}/\sqrt{2\pi}\,s}\,\mathrm{d}x \overset{(2)}{=} \frac{s}{\sqrt{2\pi}} \int \exp\!\left[-x^2\left(1 - \frac{1}{2s^2}\right)\right]\mathrm{d}x$$
$$\overset{(3)}{=} \frac{s}{\sqrt{2\pi}}\sqrt{\frac{\pi}{1 - 1/2s^2}} \overset{(4)}{=} \frac{s^2}{\sqrt{2s^2 - 1}} \qquad \text{if } s^2 > \tfrac{1}{2}, \qquad \text{and } \infty \text{ if } s^2 \le \tfrac{1}{2}. \quad (6.53)$$

(1) $f^2 = e^{-x^2}/2\pi$ and $g = e^{-x^2/2s^2}/(\sqrt{2\pi}\,s)$.
(2) Collect the constants and the exponents: $-x^2 + x^2/2s^2$.
(3) $\int e^{-cx^2}\mathrm{d}x = \sqrt{\pi/c}$ for $c > 0$. If $c = 1 - 1/2s^2 \le 0$ the integrand does not decay and the integral is infinite.
(4) $s\sqrt{1/(2 - 1/s^2)} = s \cdot s/\sqrt{2s^2 - 1}$.

Since $I = 1$, the variance of one weight is $\mathbb{E}_g[w^2] - 1 = s^2/\sqrt{2s^2 - 1} - 1$, and this is also its variance relative to the answer. It is zero at $s = 1$, when $g = f$ and every weight is exactly 1. For $s > 1$ (g wider than f) it grows only slowly: at $s = 1.5$ it is 0.2027. For $s < 1$ (g narrower) it grows fast and becomes infinite at $s = 1/\sqrt{2}$, although g is still a perfectly good Gaussian that covers the whole real line. Figure 6.33 shows the measured scatter of $\hat{Z}/Z$ over 400 runs of $N = 1000$ draws against s: on the theory curve for $s > 1/\sqrt{2}$, erratic and large below.

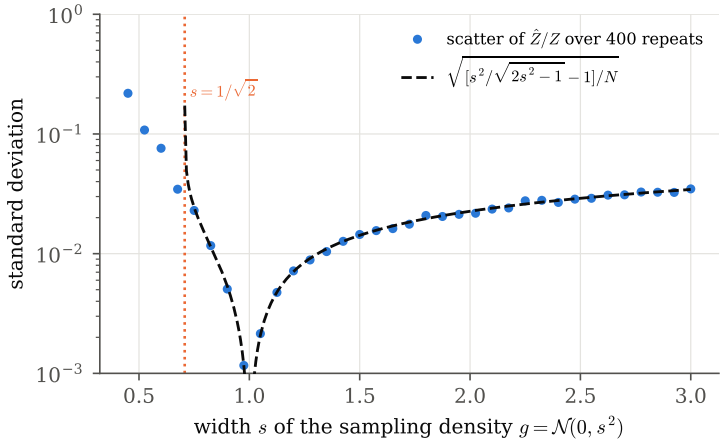


Figure 6.33: How the width of the sampling density controls importance sampling. Each point is the standard deviation of the estimate of the normalisation of a standard normal, from 400 runs of 1000 draws from $g = \mathcal{N}(0, s^2)$. It vanishes at $s = 1$ ($g = f$), rises gently for wider g, and diverges at $s = 1/\sqrt{2}$; below that the variance is infinite and the scatter over a finite number of runs is erratic.

What does an infinite variance look like in a single run? The weights are $w = f/g =$

$s\,e^{x^2(1/2s^2-1/2)}$. For $s = 0.5$ that is $\frac{1}{2}e^{3x^2/2}$: modest near the centre, and astronomically large for the rare draw far out in the tail of g. The mean of w is still finite, so the estimate still converges, by the law of large numbers. But the central limit theorem needs a finite variance, so it no longer applies, and the error does not shrink as $1/\sqrt{N}$. In practice, most of the time the running estimate sits a little *low*, because the large weights that should balance it have not been drawn yet. Then one draw lands in the tail and the estimate jumps (Figure 6.34, left). With 200000 draws, three independent runs at $s = 0.5$ end with a spread of 0.03349 in $\hat{Z}/Z$, against 7.316×10^{-4} at $s = 1.5$. The **effective sample size** $(\sum w_i)^2/\sum w_i^2$ [the number of equally weighted draws that would carry the same information] is a fraction 0.03809 of N at $s = 0.5$ and 0.8318 at $s = 1.5$.

Key idea: Thicker tails than the target

A basic rule in importance sampling is to sample from a density g with thicker tails than f [Wasserman p. 409]. The variance of the estimate contains $\int h^2 f^2/g$, with g in the denominator. If g falls off faster than f, the weights f/g grow without bound in the tails and the variance is infinite. A good g is similar in shape to f (or to $|h|f$) and a little wider.

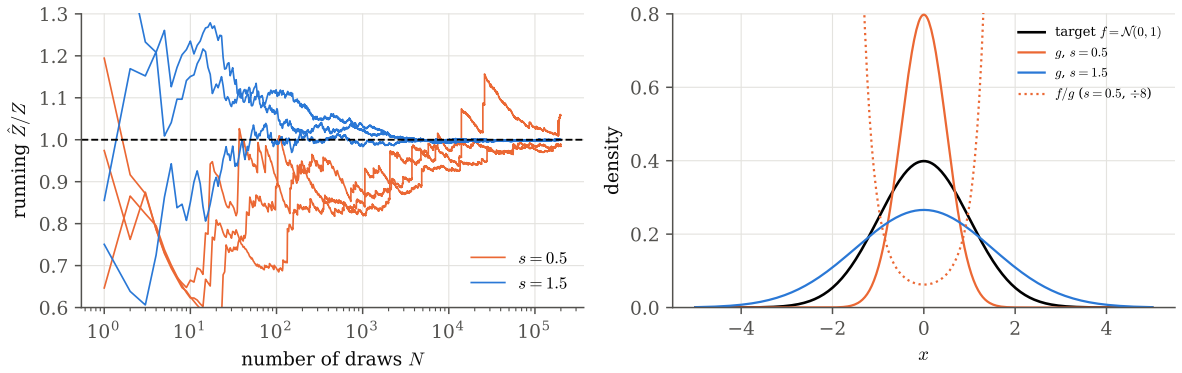


Figure 6.34: Left: running estimates of the normalisation of a standard normal, three runs each, with a narrow sampling density ($s = 0.5$, orange) and a wide one ($s = 1.5$, blue). The narrow one drifts and jumps whenever a rare tail draw arrives with a huge weight. Right: the target and the two sampling densities; the dotted curve is the weight f/g for $s = 0.5$ (scaled down by 8), which grows without bound in the tails.

Watch out: The error bar lies before the estimate does

With infinite variance the sample standard deviation of the weights is finite in every run, so $s/\sqrt{N}$ always prints a number. It is too small most of the time, because the draws that would reveal the true spread are the rare ones that have not happened yet. Check the effective sample size and the largest single weight's share of the total, and plot the running estimate.

6.12.3 The best g, and a tail probability

Which g minimises the variance? In (6.52) the term I^2 is fixed, so only the second moment $\mathbb{E}_g[h^2w^2]$ depends on g. We first find a floor below which it cannot go, whatever g we pick. The tool is the inequality $\mathbb{E}[Y^2] \geq (\mathbb{E}|Y|)^2$, a case of Jensen's inequality, which holds because the variance of $|Y|$ cannot be negative. With $Y = hw$ it gives

$$\mathbb{E}_g\left[h^2w^2\right] \overset{(1)}{\geq} \left(\mathbb{E}_g\left[|h|w\right]\right)^2 \overset{(2)}{=} \left(\int |h(x)|\,\frac{f(x)}{g(x)}\,g(x)\,\mathrm{d}x\right)^2 \overset{(3)}{=} \left(\int |h|f\,\mathrm{d}x\right)^2. \tag{6.54}$$

(1) $\mathbb{E}[Y^2] - (\mathbb{E}|Y|)^2 = \mathrm{Var}|Y| \geq 0$, with $Y = hw$.
(2) Write the expectation as an integral against g.
(3) g cancels.

Can any g reach this floor? Equality needs $|h|w$ to have zero variance, that is, to take the same value at every draw. The choice $g^*(x) = |h(x)|f(x)/\int |h|f$ does exactly that: then $|h|w = |h|f/g^* = \int |h|f$ for every draw [Wasserman Thm 24.5]. For $h \geq 0$ the estimator itself then has zero variance: every single draw returns the exact answer. This is only of theoretical interest. To normalise g^* we need $\int |h|f$, which is the integral we wanted in the first place [Wasserman p. 410]. What it teaches in practice is a guideline: make g resemble $|h|f$, the place where the integrand actually lives, and give it thicker tails.

Example: $\mathrm{P}(Z > 3)$. For $h = \mathbf{1}[x > 3]$ and $f = \mathcal{N}(0,1)$, the ideal g^* is the normal density truncated to $x > 3$. A simple imitation is $g = \mathcal{N}(4,1)$, centred beyond the threshold [Wasserman Ex. 24.6]. The weight is $w = f/g = \exp[-x^2/2 + (x-4)^2/2] = e^{8-4x}$, and the second moment of the estimator's terms can be done exactly:

$$\mathbb{E}_g[h^2w^2] \overset{(1)}{=} \int_3^\infty \frac{f^2}{g}\,\mathrm{d}x \overset{(2)}{=} \int_3^\infty \frac{1}{\sqrt{2\pi}}\, e^{-x^2+(x-4)^2/2}\,\mathrm{d}x \overset{(3)}{=} e^{16}\int_3^\infty \frac{e^{-(x+4)^2/2}}{\sqrt{2\pi}}\,\mathrm{d}x \overset{(4)}{=} e^{16}\big[1-\Phi(7)\big]. \tag{6.55}$$

(1) (6.52) with $h^2 = h = \mathbf{1}[x > 3]$.
(2) $f^2/g = \big(e^{-x^2}/2\pi\big)\,\big/\,\big(e^{-(x-4)^2/2}/\sqrt{2\pi}\big)$.
(3) Complete the square: $-x^2 + \frac{1}{2}(x-4)^2 = -\frac{1}{2}x^2 - 4x + 8 = -\frac{1}{2}(x+4)^2 + 16$.
(4) Substitute $t = x + 4$: the integral is the normal tail beyond 7.

This is 1.137×10^{-5}, so the standard deviation of $\hat{I}_g$ with $N = 100$ is $\sqrt{(1.137 \times 10^{-5} - p^2)/100} = 3.090 \times 10^{-4}$, against the plain binomial $\sqrt{p(1-p)/100} = 0.003672$. Over 20000 repeated runs the script measures 0.003667 for plain Monte Carlo and 3.086×10^{-4} for importance sampling, a gain of 11.88 in standard deviation, or about 140 in the number of draws needed for a given precision.[1] The means of the two estimators are 0.001346 and 0.001352, both unbiased for $p = 0.00135$.

The same device is how particle physicists reweight simulated events. Events generated under one set of parameters (a parton distribution set, a value of a coupling, a Wilson coefficient) are reused for another by the weight $w = f_{\text{new}}/f_{\text{old}}$ of each event's kinematics, and the effective sample size tells us when the old sample is too far from the new model to be trusted. The tail rule applies exactly: reweighting towards a harder spectrum than the one generated puts huge weights on the few high-energy events.

6.12.4 Self-normalised importance sampling for posteriors

For a posterior we know $f(\theta) \propto \mathcal{L}(\theta)\pi(\theta)$, the likelihood times the prior, only up to the evidence, which is the unknown normalising constant. So importance sampling as written cannot be used directly: the weight f/g needs f itself. The way round it is to work with the unnormalised posterior $\tilde{f} = \mathcal{L}\pi$ and the weight $w = \tilde{f}/g$. The posterior mean is a ratio of two

[1] Wasserman reports the reduction as "a factor of 20" [Wasserman p. 410], from simulated values 0.0039 and 0.0002 that he labels as variances but that are standard deviations. (A variance of 0.0039 is impossible here: the plain estimator's variance is $p(1-p)/100 \approx 1.3 \times 10^{-5}$.) The exact standard deviations from (6.55) are 0.0037 and 0.00031, a factor of 11.9, which our simulation reproduces; his second value, 0.0002, is low, and the ratio of his rounded numbers, 19.5, is quoted as 20.

integrals, and each can be estimated from the same draws $\theta_j \sim g$ [Wasserman Ex. 24.7]:

$$\mathbb{E}[\theta \mid \boldsymbol{d}] \overset{(1)}{=} \frac{\int \theta\, \tilde{f}(\theta)\, \mathrm{d}\theta}{\int \tilde{f}(\theta)\, \mathrm{d}\theta} \overset{(2)}{=} \frac{\mathbb{E}_g[\theta\, w(\theta)]}{\mathbb{E}_g[w(\theta)]} \approx \frac{\sum_j \theta_j\, w_j}{\sum_j w_j}. \tag{6.56}$$

(1) The evidence $\int \tilde{f}$ normalises the posterior, so it divides the unnormalised integral.
(2) Importance sampling (6.50) applied to numerator and denominator separately.

The unknown constant cancels between numerator and denominator. The price is a small bias: a ratio of two unbiased averages is not itself unbiased for finite N. The bias vanishes as N grows (the estimator is **consistent**), and the tail rule applies to it just as before.

Example: a measurement with an outlier. Model three measurements as $x_i = \theta + \epsilon_i$ with Student-t noise of $\nu = 3$ degrees of freedom, a heavy-tailed noise model that tolerates occasional wild values [Wasserman Ex. 24.7], and a flat prior on θ. Take the data $x = (-4, 0.5, 1)$, where the first point is an outlier. The posterior mean, by numerical quadrature, is 0.1204.

The tails decide the sampler. The t_3 density is proportional to $(1 + t^2/3)^{-2}$, so each likelihood factor falls off as $|\theta|^{-4}$, and the posterior, a product of three such factors, has polynomial tails $\propto |\theta|^{-12}$. Try $g = \mathcal{N}(0, 1)$, whose tails are Gaussian: the weight $\tilde{f}/g$ grows like $e^{\theta^2/2}|\theta|^{-12}$, and the variance (6.52) is infinite. Try $g = \mathrm{Cauchy}(0, 1)$, with tails $\propto \theta^{-2}$, heavier than the posterior's: the weight is bounded.

The numbers. Over 2000 runs of 1000 draws each, the Normal sampler gives estimates with median 0.1471, standard deviation 0.1335 and a 1–99% range $[-0.2624, 0.2387]$; the Cauchy sampler gives median 0.1202, standard deviation 0.03204 and range $[0.04202, 0.1915]$. The typical Normal-sampler answer is pulled towards the bulk of the data. The reason is that the region around the outlier, which pulls the posterior mean down, is almost never sampled. Wasserman's two-point version shows the same: -0.74 from a Normal sampler and -0.53 from a Cauchy sampler, against the true -0.54 [Wasserman p. 411].

Listing 6.13: Self-normalised importance sampling of a posterior mean (`code/ch06/27_importance.py`)

```
120 for name, g in [("Norm", stats.norm(loc=0.0, scale=1.0)), ("Cauchy", stats.cauchy(loc=0.0,
        scale=1.0))]:
121     ests = []
122     for _ in range(2000):
123         th = g.rvs(size=1000, random_state=rng)
124         w = like(th) / g.pdf(th)
125         ests.append(np.sum(w * th) / np.sum(w))     # self-normalised: the constant of f
        cancels
```

Line 124 is the weight, the unnormalised posterior over the sampling density, and line 125 is (6.56). Nothing in the loop knows the evidence.

Importance sampling returns in the rest of the book. Nested sampling and the evidence estimators of Chapter 5 are importance sampling with cleverly chosen g. An MCMC chain (Chapter 9) can be reweighted to a slightly different prior or likelihood with the weights $p_{\mathrm{new}}/p_{\mathrm{old}}$, subject to the same tail rule. And Casella and Berger's condition $M < \infty$ for rejection sampling (Section 6.10.4) is the same requirement in its strongest form: rejection needs f/g bounded, importance sampling needs only $\int f^2/g < \infty$.

6.13 A histogram of samples approximates the density

To see the distribution that a pile of samples came from, we histogram them. Every method so far ends with such a pile: decay times, directions, posterior draws of the difference $\delta = p_2 - p_1$ of two efficiencies. From Chapter 9 on, a Markov chain will hand us a pile of parameter values and nothing else, with no formula for the posterior at all. Wasserman states the idea for posterior draws: if $\theta_1, \dots, \theta_B$ are drawn from $p(\theta \mid \boldsymbol{d})$, then a histogram of $\theta_1, \dots, \theta_B$ approximates the posterior density [Wasserman §11.4, p. 180]. Kruschke builds his whole approach to Bayesian computation on it. A distribution is represented by a large sample of representative values, as an opinion poll represents a population, and the larger the sample, the better the representation [Kruschke §7.1].

The idea is not special to posteriors. It also gives the probability distribution of the position of a particle: we plot the histogram of x for a large set of random values of the time t. Take a classical oscillator $x(t) = A\cos\omega t$, observed at a random moment. It is found at position x with some probability density, and we can obtain that density without any algebra: sample many random times, compute x at each, and histogram. Here the sampling prescription is "the time is uniform". As Section 1.11.3 showed, the equation of motion alone fixes no distribution; choosing the position uniformly instead would give a different one.

The question. How close is a histogram of N samples to the density that produced them, and what error bar belongs on each bin?

6.13.1 What a bin count is

Divide the range into bins of width Δ, and let bin k be the interval $[b_k, b_{k+1})$. Each sample falls into bin k with probability

$$p_k = \int_{b_k}^{b_{k+1}} f(x)\,\mathrm{d}x = F(b_{k+1}) - F(b_k),$$

independently of the other samples. So the count n_k in bin k is the number of successes in N independent trials with success probability p_k: it is Binomial(N, p_k) (Chapter 2). The natural estimate of the density is the fraction of samples in the bin per unit length, $\hat{f}_k = n_k/(N\Delta)$.

Working definition 6.7 (histogram density estimate)

$\hat{f}_k = n_k/(N\Delta)$ [the fraction of the N samples that fell in bin k, divided by the bin width]. It estimates the *bin-averaged* density p_k/Δ, with standard error

$$\delta\hat{f}_k = \frac{\sqrt{Np_k(1-p_k)}}{N\Delta} \approx \frac{\sqrt{n_k(1-n_k/N)}}{N\Delta} \approx \frac{\sqrt{n_k}}{N\Delta} \quad (p_k \ll 1).$$

The mean and variance follow from those of the binomial:

$$\mathbb{E}[\hat{f}_k] \overset{(1)}{=} \frac{\mathbb{E}[n_k]}{N\Delta} \overset{(2)}{=} \frac{Np_k}{N\Delta} = \frac{p_k}{\Delta}, \qquad \mathrm{Var}[\hat{f}_k] \overset{(3)}{=} \frac{\mathrm{Var}[n_k]}{N^2\Delta^2} \overset{(4)}{=} \frac{p_k(1-p_k)}{N\Delta^2}, \qquad \frac{\delta\hat{f}_k}{\mathbb{E}[\hat{f}_k]} \overset{(5)}{=} \sqrt{\frac{1-p_k}{Np_k}}. \tag{6.57}$$

(1) $N\Delta$ is a constant.
(2) A binomial count has mean Np_k.
(3) A constant factor comes out of a variance squared.

(4) A binomial count has variance $Np_k(1-p_k)$.
(5) The standard deviation divided by the mean.

So the relative error of a bin is about $1/\sqrt{Np_k}$, one over the square root of the expected count in that bin. A bin expected to hold 100 samples is good to 10%, whatever N is. When there are many bins, each p_k is small, $1-p_k \approx 1$, and the binomial is close to a Poisson distribution with mean Np_k. This gives the familiar $\sqrt{n}$ error bar.

Are the bins independent of each other? With N fixed, not quite: the counts must add up to N, so a sample in one bin is a sample missing from the others. They are multinomial and slightly anticorrelated, $\mathrm{Cov}(n_j, n_k) = -Np_jp_k$, which is negligible when all the p_k are small. If N is itself a Poisson count, as in a counting experiment with a fixed exposure, the bin counts are exactly independent Poisson variables.

What a bin estimates is the density *averaged over the bin*, not its value at the centre. Expand f about the bin centre c to see the difference:

$$\frac{p_k}{\Delta} \overset{(1)}{=} \frac{1}{\Delta}\int_{-\Delta/2}^{\Delta/2} f(c+s)\,\mathrm{d}s \overset{(2)}{=} \frac{1}{\Delta}\int_{-\Delta/2}^{\Delta/2}\Big[f(c) + f'(c)\,s + \tfrac{1}{2}f''(c)\,s^2 + \dots\Big]\mathrm{d}s \overset{(3)}{=} f(c) + \frac{\Delta^2}{24}f''(c) + \dots \tag{6.58}$$

(1) Substitute $x = c + s$ in the integral over the bin.
(2) Taylor expansion of f about the centre.
(3) The linear term integrates to zero over a symmetric interval, and $\int_{-\Delta/2}^{\Delta/2} s^2\,\mathrm{d}s = \Delta^3/12$.

So a histogram is biased by $\Delta^2 f''/24$ as an estimate of f at the bin centres, and the bias is largest where the density curves sharply.

How wide should the bins be? Narrow bins reduce the bias but, by (6.57), raise the relative error, since each bin holds fewer samples. The squared bias falls like Δ^4 and the variance rises like $1/(N\Delta)$. Their sum is smallest where the two are balanced, $\Delta^4 \propto 1/(N\Delta)$, that is $\Delta \propto N^{-1/5}$. Both terms are then of order $N^{-4/5}$, so the error falls like $N^{-2/5}$: the histogram converges to the density more slowly than $1/\sqrt{N}$. For comparing a histogram with a model, the bias can be avoided altogether: compare n_k with the model's expected count Np_k computed from the CDF, which is exact for any bin width.

6.13.2 The oscillator seen at random times

We now put the histogram to work on the oscillator of the opening of this section, $x(t) = A\cos\omega t$ with amplitude A and angular frequency ω. We pick the observation time t uniformly over one period $T = 2\pi/\omega$. To have something to compare the histogram with, we first need the exact density of x. It follows from the change of variables of Chapter 1. Each value of x in $(-A, A)$ is reached at two times per period, once moving left and once moving right, and at each of them the speed is $|\mathrm{d}x/\mathrm{d}t| = A\omega|\sin\omega t| = \omega\sqrt{A^2 - x^2}$:

$$p(x) \overset{(1)}{=} \sum_{2\text{ times}} p_t(t)\left|\frac{\mathrm{d}t}{\mathrm{d}x}\right| \overset{(2)}{=} 2\cdot\frac{1}{T}\cdot\frac{1}{\omega\sqrt{A^2-x^2}} \overset{(3)}{=} \frac{1}{\pi\sqrt{A^2-x^2}}, \qquad F(x) \overset{(4)}{=} \frac{1}{2} + \frac{1}{\pi}\arcsin\frac{x}{A}. \tag{6.59}$$

(1) The probability of finding x in $\mathrm{d}x$ is the probability that t falls in the time intervals that map onto $\mathrm{d}x$, summed over both branches.
(2) t is uniform over the period, $p_t = 1/T$, and the speed is the same on both branches.
(3) $\omega T = 2\pi$.
(4) Integrate: $\int \mathrm{d}x/\sqrt{A^2-x^2} = \arcsin(x/A)$, with $F(-A) = 0$.

This is the arcsine density of Section 1.11.3. It diverges at the turning points $x = \pm A$, where the oscillator is slow and spends most of its time. It is lowest at the centre, where the oscillator moves fastest. The divergence is integrable, so the total probability is still one. Classical statistical mechanics makes the same statement: picking a random time is the same as picking a random point on the phase-space orbit, so the fraction of time the system spends in a region is the probability of finding it there. A Markov chain (Chapter 9) uses exactly this: the fraction of time the chain spends in a region is its estimate of the region's probability.

Listing 6.14: The oscillator position at random times, histogrammed with binomial error bars (`code/ch06/28_histogram_density.py`)

```
A, omega, N, K = 1.0, 2 * np.pi, 10_000, 40
t = rng.random(N) * (2 * np.pi / omega)          # random times over one period
x = A * np.cos(omega * t)
edges = np.linspace(-A, A, K + 1)
Delta = edges[1] - edges[0]
n_k, _ = np.histogram(x, bins=edges)
F = lambda x: 0.5 + np.arcsin(np.clip(x / A, -1, 1)) / np.pi
p_k = np.diff(F(edges))
dens = n_k / (N * Delta)
err = np.sqrt(n_k * (1 - n_k / N)) / (N * Delta)
pull = (n_k - N * p_k) / np.sqrt(N * p_k * (1 - p_k))
chi2 = np.sum((n_k - N * p_k) ** 2 / (N * p_k))  # Pearson chi^2, K - 1 degrees of freedom
```

Line 31 draws the random times and line 32 evaluates the trajectory at each; nothing else is needed to generate the distribution. Lines 33–38 build the histogram and convert counts into a density. Line 37 computes the exact bin probabilities p_k from the CDF (6.59), line 39 the error bars of (6.57), and lines 40–41 the pull of each bin and the χ^2 of the whole histogram.

Example: how good is it? With $N = 10000$ random times and $K = 40$ bins on $[-A, A]$ (Figure 6.35), the Pearson statistic $\sum_k (n_k - Np_k)^2/Np_k$ is 35.74 for 39 degrees of freedom (K bins minus one, because the counts must add up to N), a p-value of 0.6193 (Chapter 4). The pulls $(n_k - Np_k)/\sqrt{Np_k(1-p_k)}$ have standard deviation 0.9676, close to one, which says that the binomial error bars are the right size. The relative error in a central bin is 0.07862 and in the end bin 0.02982, because the end bins hold many more samples.

Example: the bin average is not the value at the centre. In the central bin the bin-averaged density $p_k/\Delta = 0.3184$ equals the density at the centre, 0.3184, to four digits, because the density is flat there. In the last bin, $[0.95A, A)$, the average is $2.022/A$ while the density at the bin centre is only $1.433/A$: the divergence at $x = A$ lies inside the bin, and the Taylor expansion (6.58) fails completely there. A histogram drawn over the curve $1/\pi\sqrt{A^2 - x^2}$ therefore seems to overshoot at the ends, and a χ^2 against the density at the bin centres would reject a perfect sample. The green staircase in Figure 6.35, the exact bin averages, is what the histogram estimates.

Example: summaries from a representative sample. Kruschke's illustration represents a Beta(15, 7) density, which could be the posterior for the probability of heads after 14 heads in 20 tosses with a flat prior, by random samples of increasing size [Kruschke Fig. 7.1]. The exact mode is 0.7 and the exact 95% highest-density interval (the shortest interval holding 95% of the probability, Chapter 5) is $[0.4907, 0.8639]$. From the samples we estimate the mode as the centre of the tallest histogram bin and the HDI as the shortest interval containing 95% of the sorted draws. With $N = 500$ the script finds a mode of 0.67 and an HDI of $[0.4766, 0.8445]$; with $N = 5000$, 0.71 and $[0.4882, 0.8656]$; with $N = 50000$, 0.71 and $[0.4907, 0.8637]$ (Figure 6.36). The interval limits settle to within 0.01 of the exact ones at $N = 5000$, as the $1/\sqrt{N}$ law predicts.

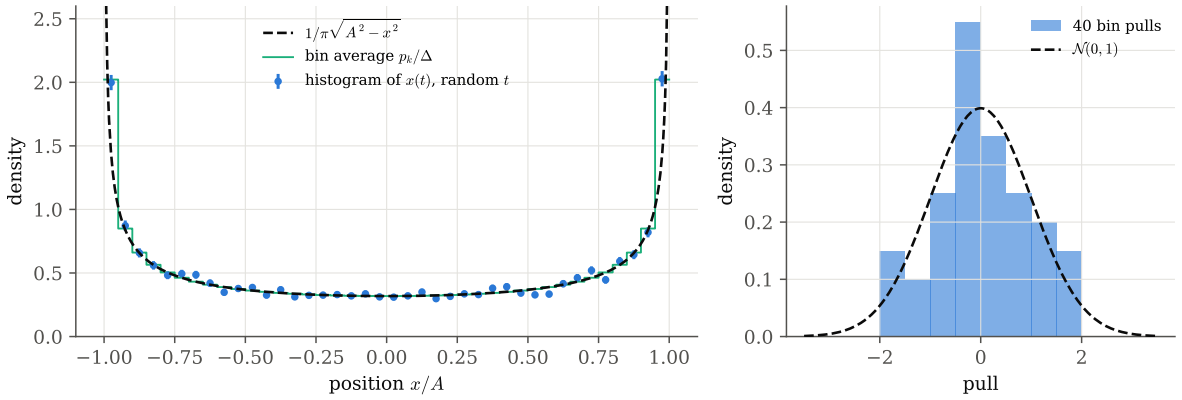


Figure 6.35: The position of a classical oscillator at 10^4 random times. Left: the histogram as a density with binomial error bars, the arcsine density (black dashed) and its exact bin averages (green staircase); the histogram follows the staircase, which lies above the curve in the end bins because the density diverges inside them. Right: the pulls of the 40 bins are distributed as a standard normal.

The mode is the slowest summary to settle, because it depends on the height of a single bin, which is the noisiest thing a histogram offers. Kruschke's numbers (0.723 and $[0.477, 0.875]$ at $N = 500$) scatter about the truth in the same way.

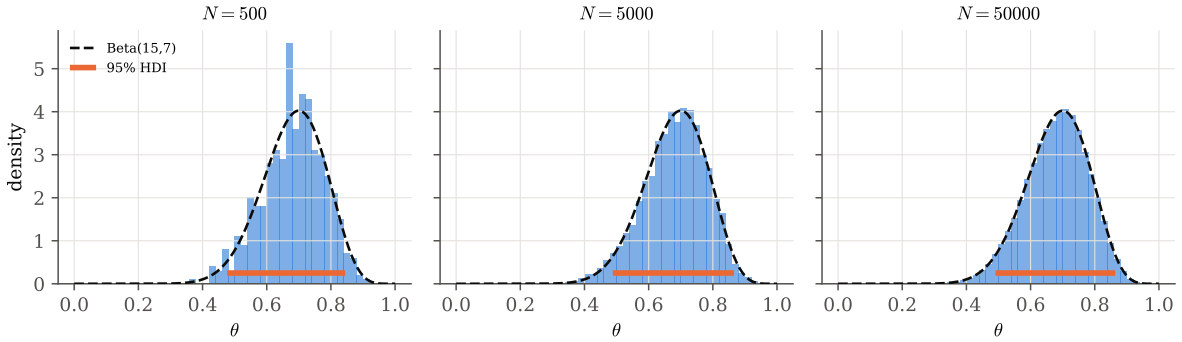


Figure 6.36: A Beta(15, 7) density represented by 500, 5000 and 50 000 random samples, with the 95% highest-density interval found from each sample (orange bar). The histogram becomes smoother and the interval converges on the exact one as N grows.

Watch out: Empty bins and the density at the centre

Two mistakes are common when a simulated histogram is compared with a model. The first is to compare $n_k/(N\Delta)$ with f at the bin centre instead of with p_k/Δ; it fails wherever f curves sharply or diverges. The second is to put bins with zero or very few expected counts into a Pearson χ^2, where the Gaussian approximation behind χ^2 fails; use the Poisson likelihood of Chapter 3 instead, or merge the sparse bins.

6.14 Unfolding: undoing the detector with a response matrix from simulation

A calorimeter never tells us the energy a photon had. It tells us the energy it measured, which is the true energy plus a random error, and sometimes it tells us nothing because the photon

went through a crack. So the histogram an experiment records is not the spectrum nature produced. It is that spectrum smeared out by the resolution and thinned out by the efficiency. A sharp peak comes out broad, two nearby peaks merge, a steep edge is softened. To compare our measurement with another experiment's, whose detector smears differently, or with a theory written down twenty years later, we would like to report the spectrum *before* the detector touched it. Undoing the detector is called **unfolding** [Cowan §11].

We have met both halves of this before. In Example 1.43 a falling exponential spectrum of true photon energies (mean $\beta = 20\,\text{GeV}$) was pushed through a Gaussian resolution of $\sigma = 5\,\text{GeV}$, and the measured spectrum came out as the marginal $f(x) = \int f(x \mid E) f(E)\,\mathrm{d}E$: the tail was raised by a constant factor and some events were measured at negative energy. That was the forward direction, from truth to measurement, and it had a closed form because the detector was a perfect Gaussian. In Section 6.11.4 the detector had a dead strip and a hole, no closed form existed, and a Monte Carlo of particles thrown at it was the only way to learn what fraction it caught; there we noted that the same kind of simulation supplies the "response matrix for unfolding". And in Section 6.13.1 we learned what a bin count is: a binomial, nearly Poisson, number whose error is about the square root of its expectation. Unfolding puts the three together and turns them around. The forward map is now a matrix that we learn by simulation, the data are Poisson bin counts, and the direction is backward, from the measured histogram to the true one. The backward direction is where all the difficulty lives.

The question. Given a measured histogram $n_1, \dots, n_N$ and a simulation of the detector, what is our best estimate of the true histogram, and how far can we trust the error bars we put on it?

6.14.1 From the true histogram to the measured one

Every event has two numbers attached to it: its true value y (the energy the photon really had), which we never see, and its measured value x. Two things can happen between them [Cowan §11.1]. The event may not be recorded at all; the probability that it is recorded is the **efficiency** $\varepsilon(y)$, which can depend on the true value (low energies may fall below threshold). And if it is recorded, the measured value scatters around the true one with a conditional density $s(x \mid y)$, the **resolution function** [the density of the measured x for an event whose true value is y, given that it was recorded], normalised so that $\int s(x \mid y)\,\mathrm{d}x = 1$.

We describe the truth by a histogram, because we do not know its functional form. Split the range of y into M bins. Let μ_j be the *expected* number of events whose true value falls in bin j; this is $\mu_{\text{tot}}\, p_j$, with $p_j = \int_{\text{bin } j} f_{\text{true}}(y)\,\mathrm{d}y$ the probability of bin j and μ_{tot} the expected total. The vector $\boldsymbol{\mu} = (\mu_1, \dots, \mu_M)$ is the **true histogram** we want. The measured values are histogrammed into N bins with counts $\boldsymbol{n} = (n_1, \dots, n_N)$ and expectations $\nu_i = \mathbb{E}[n_i]$. How is ν_i built from $\boldsymbol{\mu}$? An event lands in observed bin i by first having some true value y, then being recorded, then being measured inside bin i. Summing over the ways this can happen is the law of total probability (Chapter 1):

$$\begin{aligned} \nu_i &\overset{(1)}{=} \mu_{\text{tot}} \int \mathrm{d}y\; f_{\text{true}}(y)\, \varepsilon(y) \int_{\text{bin } i} s(x \mid y)\,\mathrm{d}x \\ &\overset{(2)}{=} \sum_{j=1}^{M} \mu_{\text{tot}} \int_{\text{bin } j} \mathrm{d}y\; f_{\text{true}}(y)\, \varepsilon(y) \int_{\text{bin } i} s(x \mid y)\,\mathrm{d}x \overset{(3)}{=} \sum_{j=1}^{M} R_{ij}\, \mu_j, \end{aligned} \tag{6.60}$$

with

$$R_{ij} = \frac{\int_{\text{bin } j} \mathrm{d}y\, f_{\text{true}}(y)\, \varepsilon(y) \int_{\text{bin } i} s(x \mid y)\,\mathrm{d}x}{\int_{\text{bin } j} f_{\text{true}}(y)\,\mathrm{d}y} = \frac{\text{P(true in bin } j \text{ and observed in bin } i)}{\text{P(true in bin } j)}. \tag{6.61}$$

(1) The expected number in bin i is μ_{tot} times the probability that one event ends there: average over the true value y (density f_{true}) the probability of being recorded (ε) and then measured inside bin i (the integral of s).
(2) Cut the y integral into the M true bins.
(3) Multiply and divide each term by $\mu_j = \mu_{\text{tot}} \int_{\text{bin } j} f_{\text{true}}\, \mathrm{d}y$; the ratio left over is R_{ij}.

The second form of (6.61) is a joint probability divided by a marginal, so R_{ij} is a conditional probability.

Working definition 6.8 (response matrix)

$R_{ij} = \mathrm{P}(\text{observed in bin } i \,|\, \text{true value in bin } j)$ [of the events created in true bin j, the fraction that are recorded and land in measured bin i]. It maps the true histogram to the expected measured one, $\boldsymbol{\nu} = \mathbf{R}\boldsymbol{\mu} + \boldsymbol{\beta}$, where β_i is the expected background in bin i [Cowan eqs. 11.9, 11.12].

Each column of $\mathbf{R}$ says where the events of one true bin go. If we add up a column, we add up the probabilities of all the places a recorded event can land, and what is missing is the probability that it was not recorded:

$$\sum_{i=1}^{N} R_{ij} \overset{(1)}{=} \frac{\int_{\text{bin } j} \mathrm{d}y\, f_{\text{true}}(y)\, \varepsilon(y) \int s(x \,|\, y)\, \mathrm{d}x}{\int_{\text{bin } j} f_{\text{true}}(y)\, \mathrm{d}y} \overset{(2)}{=} \frac{\int_{\text{bin } j} f_{\text{true}}(y)\, \varepsilon(y)\, \mathrm{d}y}{\int_{\text{bin } j} f_{\text{true}}(y)\, \mathrm{d}y} \equiv \varepsilon_j. \tag{6.62}$$

(1) The N observed bins together cover the whole measured range, so the sum of the integrals over bin i is one integral over all x.
(2) $\int s(x \,|\, y)\, \mathrm{d}x = 1$; what remains is the efficiency averaged over bin j.

So the column sums are the efficiencies of the true bins, and they are less than one whenever events can be lost. The off-diagonal elements are **migrations**: events created in one bin and measured in another. They are what smears out fine structure.

The response matrix comes from simulation. Look again at (6.61): it contains f_{true}, the very thing we want to find. Inside a narrow bin, however, the efficiency and resolution hardly change with y, and then f_{true} cancels between numerator and denominator. In practice we therefore build $\mathbf{R}$ from a simulation that uses a plausible guess for the truth, exactly as in Section 6.11.4: generate N_j events with true value in bin j, push each through the detector simulation, and count how many, N_{ij}, land in measured bin i. Then $\hat{R}_{ij} = N_{ij}/N_j$, a binomial fraction with error $\sqrt{R_{ij}(1 - R_{ij})/N_j}$ (Section 6.13.1). The residual dependence on the guessed truth is a systematic error of the method, and so is any imperfection of the detector simulation [Cowan p. 157].

Example: a two-peak spectrum and a Gaussian detector. We use a test case modelled on Cowan's [Cowan Fig. 11.1]. The true variable y lives on $[0, 1]$, split into $M = N = 20$ bins of width $\Delta = 0.05$. The truth is a flat floor holding 20% of the events plus two Gaussian peaks of width 0.08 at $y = 0.3$ and 0.7, with $\mu_{\text{tot}} = 20\,000$ expected events. The detector records $x = y + \mathcal{N}(0, \sigma^2)$ with $\sigma = 1.5$ bin widths; an event smeared outside $[0, 1]$ is lost, which is our efficiency. We build $\mathbf{R}$ from 2×10^6 simulated events generated *flat* in y, since the analyser does not know the truth.

Listing 6.15: The response matrix from simulated events (`code/ch06/30_unfolding.py`)

```
N_mc = 2_000_000
y_mc = rng.random(N_mc)                    # simulated truth, flat
x_mc = y_mc + sigma * rng.standard_normal(N_mc)
```

```
59 H, _, _ = np.histogram2d(x_mc, y_mc, bins=[edges, edges])   # H[i, j]: observed i, true j
60 gen_j = np.histogram(y_mc, bins=edges)[0]
61 R = H / gen_j                              # R_ij = N(obs i, true j) / N(true j)
62 eff = R.sum(axis=0)                        # column sums = efficiency of each true bin
63 nu = R @ mu                                # expected observed histogram
```

Lines 57–58 are the whole detector: a flat truth and a Gaussian error. Line 59 fills the two-dimensional histogram N_{ij} of (measured, true) bins, line 60 counts the generated events per true bin, and line 61 divides, which is $\hat{R}_{ij} = N_{ij}/N_j$. Line 62 takes the column sums (6.62) and line 63 folds the truth, $\boldsymbol{\nu} = \mathbf{R}\boldsymbol{\mu}$.

An event created in a central bin stays there with probability 25.7%, moves one bin left or right with probability 41.6% in total, and moves two or more bins with probability 32.7%. The central columns sum to 1, while the two edge columns sum to 0.628: more than a third of the events created in an edge bin are smeared out of the range. The binomial error of a central element is 0.62% of its value. Had we simulated with the true two-peak shape instead of a flat one, the folded histogram $\mathbf{R}\boldsymbol{\mu}$ would change by up to 3.94% in some bin. That is the size of the model dependence that our bins, which are not narrow compared with the peaks, leave behind. Figure 6.37 shows the chain from truth to data.

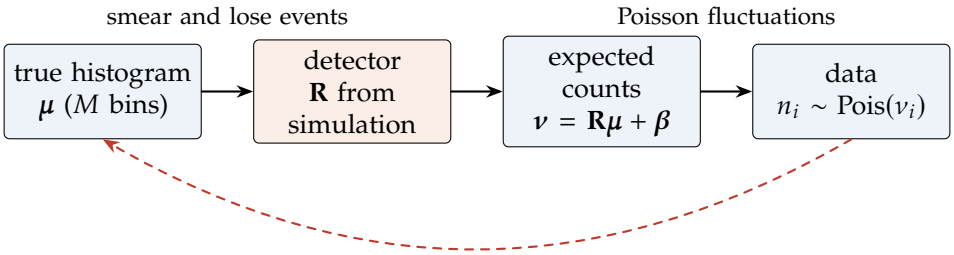


Figure 6.37: The forward chain. The true histogram is pushed through the response matrix, which smears it and loses some events, and adds a background; the counts we record fluctuate about the result. Unfolding is the dashed return path, from the data back to the true histogram.

6.14.2 The obvious answer: invert the matrix

If $M = N$ and $\mathbf{R}$ is invertible, the forward relation can be solved for the truth, $\boldsymbol{\mu} = \mathbf{R}^{-1}(\boldsymbol{\nu} - \boldsymbol{\beta})$. We do not know $\boldsymbol{\nu}$, but the data $\boldsymbol{n}$ estimate it without bias, so the obvious estimator is [Cowan §11.2]

$$\hat{\boldsymbol{\mu}} = \mathbf{R}^{-1}(\boldsymbol{n} - \boldsymbol{\beta}). \tag{6.63}$$

Is it a good estimator? Its mean and covariance follow from linearity, with $\mathbf{V}$ the covariance matrix of the data, $V_{ij} = \text{Cov}(n_i, n_j)$:

$$\begin{aligned} \mathbb{E}[\hat{\boldsymbol{\mu}}] &\overset{(1)}{=} \mathbf{R}^{-1}\big(\mathbb{E}[\boldsymbol{n}] - \boldsymbol{\beta}\big) \overset{(2)}{=} \mathbf{R}^{-1}(\boldsymbol{\nu} - \boldsymbol{\beta}) \overset{(3)}{=} \boldsymbol{\mu}, \\ \mathbf{U} \equiv \text{Cov}(\hat{\boldsymbol{\mu}}) &\overset{(4)}{=} \mathbf{R}^{-1}\,\mathbf{V}\,(\mathbf{R}^{-1})^{\mathsf{T}}, \qquad U_{ij} \overset{(5)}{=} \sum_k (R^{-1})_{ik}(R^{-1})_{jk}\,\nu_k. \end{aligned} \tag{6.64}$$

(1) The expectation of a fixed matrix times a random vector is the matrix times the expectation; $\boldsymbol{\beta}$ is a constant.
(2) $\mathbb{E}[n_i] = \nu_i$ by definition.
(3) The forward relation $\boldsymbol{\nu} = \mathbf{R}\boldsymbol{\mu} + \boldsymbol{\beta}$.
(4) The covariance of a linear map, $\text{Cov}(\mathbf{A}\boldsymbol{n}) = \mathbf{A}\,\text{Cov}(\boldsymbol{n})\mathbf{A}^{\mathsf{T}}$, as in (6.36); the constant $\boldsymbol{\beta}$ does not fluctuate.

(5) Independent Poisson counts have $V_{kl} = \delta_{kl} \nu_k$.

So the naive inverse is *unbiased*. It is also the maximum-likelihood estimator. With independent Poisson counts the log-likelihood (Section 3.11.1) is $\ln L(\boldsymbol{\mu}) = \sum_i (n_i \ln \nu_i - \nu_i)$ up to a constant, with $\nu_i = \sum_j R_{ij} \mu_j + \beta_i$, and setting its derivatives to zero gives

$$\frac{\partial \ln L}{\partial \mu_k} \overset{(1)}{=} \sum_i \frac{\partial \ln L}{\partial \nu_i} \frac{\partial \nu_i}{\partial \mu_k} \overset{(2)}{=} \sum_i \left(\frac{n_i}{\nu_i} - 1 \right) R_{ik} = 0. \tag{6.65}$$

(1) Chain rule through the ν_i.
(2) $\partial (n_i \ln \nu_i - \nu_i)/\partial \nu_i = n_i/\nu_i - 1$ and $\partial \nu_i / \partial \mu_k = R_{ik}$.

For an invertible $\mathbf{R}$ the only way all M sums vanish is $n_i/\nu_i - 1 = 0$ for every i, so $\hat{\boldsymbol{\nu}} = \boldsymbol{n}$ and $\hat{\boldsymbol{\mu}}$ is (6.63). Its covariance is even the smallest any unbiased estimator can have: the Fisher information (Section 3.12.3) is $-\mathbb{E}[\partial^2 \ln L / \partial \mu_k \partial \mu_l] = \sum_i R_{ik} R_{il} / \nu_i$, that is $\mathbf{R}^\mathsf{T} \mathbf{V}^{-1} \mathbf{R}$, and its inverse is $\mathbf{R}^{-1} \mathbf{V} (\mathbf{R}^\mathsf{T})^{-1}$, the $\mathbf{U}$ of (6.64) [Cowan eqs. 11.21–11.26]. Unbiased, maximum likelihood, minimum variance: on paper it is the best estimator there is.

Example: what it gives. Figure 6.38 shows our two-peak truth (a), its folded version $\boldsymbol{\nu} = \mathbf{R}\boldsymbol{\mu}$ (b), one Poisson data set (c), and the naive inverse of that data set (d). The data have relative errors between 2.5% and 7.1%. The inverse oscillates wildly from bin to bin, with values of both signs in the tens of thousands, and its error bars are as large as the values. In the central bin the true content is 436 and the standard deviation of the estimate is 32000, which is 1120 times the Poisson error of a single measured bin. Neighbouring bins have correlation coefficient −0.726: when one bin goes up, its neighbour goes down. A toy study (Section 6.14.6) confirms that the estimator is nevertheless unbiased: the mean of 2000 toy unfoldings agrees with $\boldsymbol{\mu}$ in every bin, the largest deviation being 1.98 standard errors of that mean.

Why it fails: the simplest smearing first. The failure is not a bug, and to see where it comes from we shrink the problem to two bins [163]. Let an event stay in its bin with probability $(1+\epsilon)/2$ and cross to the other with probability $(1-\epsilon)/2$, with $0 < \epsilon \leq 1$. A perfect detector has $\epsilon = 1$; a detector that can hardly tell the bins apart has $\epsilon \ll 1$. No event is lost, so

$$\mathbf{R} = \frac{1}{2} \begin{pmatrix} 1+\epsilon & 1-\epsilon \\ 1-\epsilon & 1+\epsilon \end{pmatrix}.$$

Which patterns does this matrix leave alone, apart from a factor? Try the sum pattern $\boldsymbol{u}_+ = (1,1)/\sqrt{2}$ and the difference pattern $\boldsymbol{u}_- = (1,-1)/\sqrt{2}$:

$$\mathbf{R}\boldsymbol{u}_+ \overset{(1)}{=} \frac{1}{2\sqrt{2}} \begin{pmatrix} (1+\epsilon)+(1-\epsilon) \\ (1-\epsilon)+(1+\epsilon) \end{pmatrix} = 1 \cdot \boldsymbol{u}_+, \qquad \mathbf{R}\boldsymbol{u}_- \overset{(2)}{=} \frac{1}{2\sqrt{2}} \begin{pmatrix} (1+\epsilon)-(1-\epsilon) \\ (1-\epsilon)-(1+\epsilon) \end{pmatrix} = \epsilon\, \boldsymbol{u}_-. \tag{6.66}$$

(1) Each component is a row of $\mathbf{R}$ times $(1,1)/\sqrt{2}$.
(2) The same with $(1,-1)/\sqrt{2}$.

So the detector passes the total number of events unchanged (eigenvalue 1) and multiplies the difference between the bins by ϵ (eigenvalue ϵ). Write the truth in these patterns, $\boldsymbol{\mu} = s\,\boldsymbol{u}_+ + d\,\boldsymbol{u}_-$, with $s = (\mu_1+\mu_2)/\sqrt{2}$ and $d = (\mu_1-\mu_2)/\sqrt{2}$. Then $\boldsymbol{\nu} = s\,\boldsymbol{u}_+ + \epsilon d\,\boldsymbol{u}_-$: the measured difference is the true difference shrunk by ϵ. Inverting means dividing the measured difference by ϵ.

Now add the noise. Suppose both bins expect about $\bar{\nu}$ events, so that $\mathbf{V} \approx \bar{\nu}\,\mathbb{1}$. A rotation to the patterns $\boldsymbol{u}_\pm$ keeps this covariance as it is, so the measured sum and measured difference,

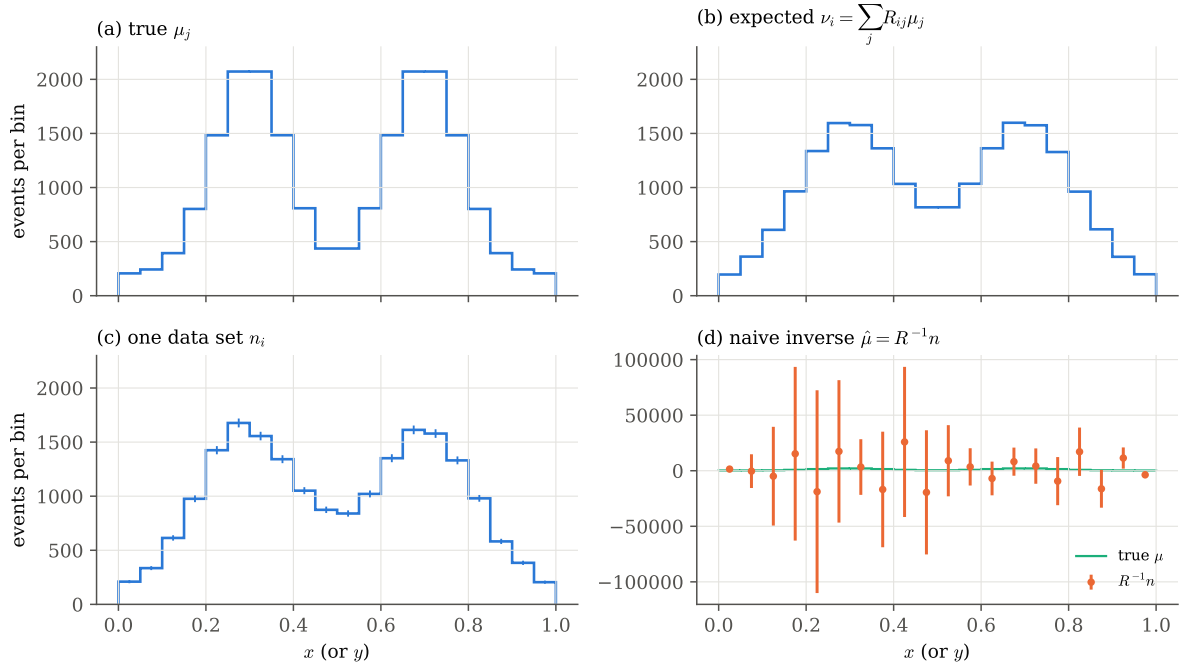


Figure 6.38: (a) The true histogram: two peaks on a flat floor. (b) What the detector makes of it on average: the peaks are lower and broader and the dip is filled in. (c) One simulated data set, with Poisson error bars of a few per cent. (d) The data set unfolded by inverting the response matrix. Note the vertical scale: the estimates swing between about $\pm 10^5$ and the true histogram (green) is a thin line near zero.

$n_s = \boldsymbol{u}_+ \cdot \boldsymbol{n}$ and $n_d = \boldsymbol{u}_- \cdot \boldsymbol{n}$, each carry noise of variance $\bar{\nu}$, independently. The inverse estimates $\hat{s} = n_s$ and $\hat{d} = n_d/\epsilon$, so

$$\mathrm{Var}(\hat{d}) \overset{(1)}{=} \frac{\bar{\nu}}{\epsilon^2}, \qquad \mathrm{Var}(\hat{\mu}_1) \overset{(2)}{=} \frac{\mathrm{Var}(\hat{s}) + \mathrm{Var}(\hat{d})}{2} = \frac{\bar{\nu}}{2}\left(1 + \frac{1}{\epsilon^2}\right),$$
$$\rho(\hat{\mu}_1, \hat{\mu}_2) \overset{(3)}{=} \frac{\mathrm{Var}(\hat{s}) - \mathrm{Var}(\hat{d})}{\mathrm{Var}(\hat{s}) + \mathrm{Var}(\hat{d})} = \frac{\epsilon^2 - 1}{\epsilon^2 + 1}. \tag{6.67}$$

(1) Dividing by ϵ multiplies a variance by $1/\epsilon^2$.
(2) $\hat{\mu}_1 = (\hat{s} + \hat{d})/\sqrt{2}$ and $\hat{s}, \hat{d}$ are independent.
(3) $\hat{\mu}_2 = (\hat{s} - \hat{d})/\sqrt{2}$, so $\mathrm{Cov}(\hat{\mu}_1, \hat{\mu}_2) = (\mathrm{Var}\,\hat{s} - \mathrm{Var}\,\hat{d})/2$; divide by the common variance of (2).

Put in numbers: $\bar{\nu} = 1000$ and $\epsilon = 0.1$. Each measured bin is good to $\sqrt{1000} \approx 32$ events, but each unfolded bin has standard deviation $\sqrt{500 \times 101} \approx 225$, seven times worse, and the two unfolded bins are correlated with $\rho = -0.98$. The reason is plain. With $\epsilon = 0.1$ the detector shrinks a true difference of 100 events to a measured difference of 10, which is smaller than the Poisson noise of that difference, $\sqrt{2 \times 1000} \approx 45$. The measured difference is then mostly noise, and dividing by ϵ blows that noise up ten times and paints it back onto the two bins with opposite signs.

Gaussian smearing: many patterns, each shrunk. A real detector has many bins, and the same reasoning applies pattern by pattern. Take Gaussian smearing on a ring of M bins of width Δ (a periodic range of length $L = M\Delta$, so that no bin is at an edge), and let $R_{ij} = r(i - j)$ depend only on how many bins the event moved, with $r(m) \approx \Delta\, g(m\Delta)$ and g the Gaussian density of standard deviation σ. Which patterns does it leave alone? On a ring every bin looks

the same, so try a wave, $v_j = e^{\mathrm{i}qj\Delta}/\sqrt{M}$ with wavenumber $q = 2\pi k/L$, $k = 0, 1, \ldots, M-1$:

$$(\mathbf{R}\boldsymbol{v})_i \overset{(1)}{=} \sum_j r(i-j)\,\frac{e^{\mathrm{i}qj\Delta}}{\sqrt{M}} \overset{(2)}{=} \frac{e^{\mathrm{i}qi\Delta}}{\sqrt{M}} \sum_m r(m)\,e^{-\mathrm{i}qm\Delta} \overset{(3)}{=} v_i \int g(x)\,e^{-\mathrm{i}qx}\,\mathrm{d}x \overset{(4)}{=} e^{-q^2\sigma^2/2}\,v_i. \tag{6.68}$$

(1) Matrix times vector.
(2) Substitute $m = i - j$; on the ring m runs over all M shifts, and $e^{\mathrm{i}qj\Delta} = e^{\mathrm{i}qi\Delta}e^{-\mathrm{i}qm\Delta}$.
(3) $r(m) \approx \Delta\, g(m\Delta)$, and a sum of Δ times a smooth function over a grid of spacing Δ is its integral, with $x = m\Delta$.
(4) The Fourier transform of a Gaussian density. Complete the square as in Example 1.43: $-x^2/2\sigma^2 - \mathrm{i}qx = -(x + \mathrm{i}q\sigma^2)^2/2\sigma^2 - q^2\sigma^2/2$; the shifted Gaussian still integrates to one.

A wave of wavenumber q is passed through with its amplitude multiplied by $\lambda_q = e^{-q^2\sigma^2/2}$. Long waves ($q\sigma \ll 1$) pass almost untouched; waves shorter than the resolution are crushed. This is the precise sense in which "smearing removes fine structure". The data contain the shrunken waves plus Poisson noise, and the noise is spread evenly over all waves: with $\mathbf{V} \approx \bar{\nu}\mathbb{1}$, each wave component of $\boldsymbol{n}$ has variance $\bar{\nu}$, as in the two-bin case. Inverting divides wave k by λ_k, so using $\mathbf{R}^{-1} = \sum_k \lambda_k^{-1}\boldsymbol{v}_k\boldsymbol{v}_k^\dagger$ in (6.64),

$$U_{jj} \overset{(1)}{=} \bar{\nu} \sum_k \lambda_k^{-2}\,|v_{k,j}|^2 \overset{(2)}{=} \frac{\bar{\nu}}{M} \sum_k e^{q_k^2\sigma^2}. \tag{6.69}$$

(1) $\mathbf{U} = \bar{\nu}\,\mathbf{R}^{-1}(\mathbf{R}^{-1})^\dagger = \bar{\nu}\sum_k \lambda_k^{-2}\boldsymbol{v}_k\boldsymbol{v}_k^\dagger$, because the waves are orthonormal and the λ_k are real.
(2) $|v_{k,j}|^2 = 1/M$ for every wave, and $\lambda_k^{-2} = e^{q_k^2\sigma^2}$.

The sum is dominated by the shortest wave. In our test case $\sigma = 1.5\Delta$, and on an interval with edges, rather than a ring, the natural waves are cosines with a whole number of half wavelengths, $q_k = \pi k/L$. The shortest has $q\sigma = \pi \times 19 \times 0.075 \approx 4.5$, so the formula predicts a gain of 4.447×10^{-5} for it. The singular values of our simulated $\mathbf{R}$ (the gains of its patterns; for a symmetric matrix they are the absolute eigenvalues) run from 0.976 down to 2.018×10^{-4}, a ratio of 4834, and follow the Gaussian formula over three decades (Figure 6.39); the last few lie above it because a finite matrix with edges is not exactly a ring. A wave that the detector shrinks to 2.018×10^{-4} of its size must be stretched back by the inverse of that factor, noise included. Averaged over the bins, the standard deviation of the naive inverse is 41200 events for a true content of about 1000.

Key idea: Why unfolding is hard

Smearing multiplies each pattern of the true histogram by a gain λ_k (for Gaussian resolution $\lambda_k = e^{-q_k^2\sigma^2/2}$), small for patterns finer than the resolution. The noise of the data is not shrunk. Inverting the response divides every pattern by its gain, so the fine patterns of the answer are noise amplified by $1/\lambda_k$. The inverse is unbiased and efficient, and useless whenever some gain is smaller than the relative noise.

The analysis also says what helps. Wider bins remove the shortest waves from the problem altogether, so the covariance of the inverse shrinks quickly once the bins are wider than the resolution [Cowan p. 163]. And if one keeps the naive inverse, it is still an honest result as long as the *full* covariance $\mathbf{U}$ goes with it: a χ^2 comparison with a theory, $(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0)^\mathsf{T}\mathbf{U}^{-1}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0)$, is valid, whereas the same χ^2 with only the diagonal of $\mathbf{U}$ is meaningless, because the off-diagonal elements are as large as the diagonal ones.

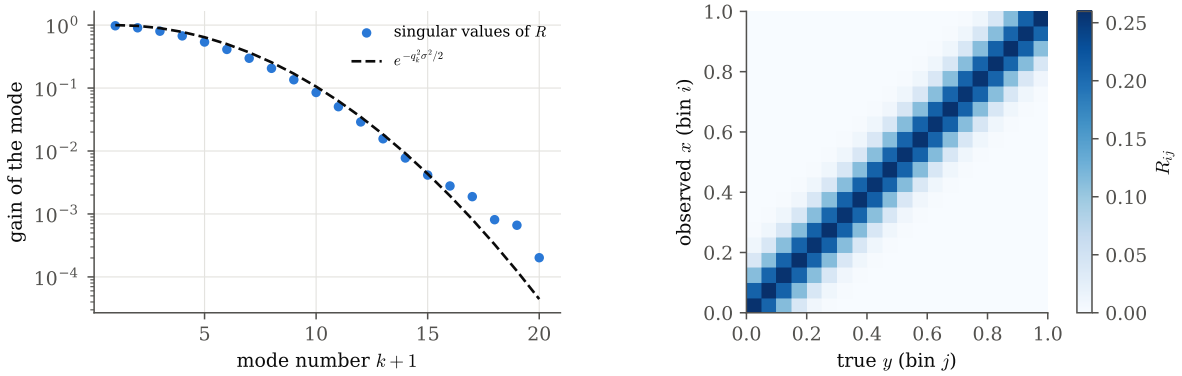


Figure 6.39: Right: the simulated response matrix of the test case; each column spreads the events of one true bin over about five measured bins, and the edge columns lose events. Left: its singular values, the factors by which it shrinks its patterns, from the smoothest to the roughest, against the Gaussian prediction $e^{-q_k^2\sigma^2/2}$ with $q_k = \pi k/L$ (dashed).

6.14.3 Trading bias for variance: regularisation

The naive inverse is unbiased and has the smallest variance of any unbiased estimator. So any estimator with a smaller variance must be *biased*. That is the whole art of unfolding: accept a bias that is small if some belief about the truth is right (usually, that it is smooth) in exchange for a large cut in variance [Cowan §11.4]. The tool is the mean squared error of Section 3.2.1, MSE = bias^2 + variance, which puts the two on one scale.

Smoothest among the acceptable. Cowan states the strategy in two steps. First, a whole region of true histograms fits the data acceptably: those whose log-likelihood is within some $\Delta \ln L$ of the maximum, $\ln L(\boldsymbol{\mu}) \geq \ln L_{\max} - \Delta \ln L$. The naive inverse is the single best fit, but its wild neighbours fit almost as well, which is exactly what the large $\mathbf{U}$ said. Second, among the acceptable histograms, pick the smoothest, as measured by a **regularisation function** $S(\boldsymbol{\mu})$ that is larger for smoother histograms. Maximising S subject to the likelihood constraint is, with a Lagrange multiplier α, maximising $\alpha \ln L(\boldsymbol{\mu}) + S(\boldsymbol{\mu})$ [Cowan eqs. 11.34–11.37]. A large α gives back the naive inverse; $\alpha \to 0$ ignores the data and returns the smoothest histogram there is.

Tikhonov regularisation. The most common smoothness measure is the mean squared curvature, Tikhonov's choice [Cowan §11.5.1]. For a histogram the second derivative becomes the second difference $\mu_i - 2\mu_{i+1} + \mu_{i+2}$, and

$$S(\boldsymbol{\mu}) = -\sum_{i=1}^{M-2} (\mu_i - 2\mu_{i+1} + \mu_{i+2})^2 = -|\mathbf{C}\boldsymbol{\mu}|^2 = -\boldsymbol{\mu}^{\mathsf{T}}\mathbf{G}\boldsymbol{\mu}, \qquad \mathbf{G} = \mathbf{C}^{\mathsf{T}}\mathbf{C}, \tag{6.70}$$

where row i of the $(M-2) \times M$ matrix $\mathbf{C}$ is $(\ldots, 0, 1, -2, 1, 0, \ldots)$ starting in column i. ($\mathbf{G}$ is Cowan's matrix in [Cowan eq. 11.48]: 6 on most of the diagonal, -4 and 1 beside it.) A straight line has zero second differences, so S does not penalise the overall level or slope at all; those are left entirely to the data, and a truth that is linear is unfolded without bias [Cowan p. 177].

With many events per bin the Poisson log-likelihood is close to the Gaussian $\ln L = -\frac{1}{2}\chi^2(\boldsymbol{\mu})$, $\chi^2 = (\boldsymbol{n} - \mathbf{R}\boldsymbol{\mu})^{\mathsf{T}}\mathbf{V}^{-1}(\boldsymbol{n} - \mathbf{R}\boldsymbol{\mu})$, taking $\boldsymbol{\beta} = 0$ for brevity. Maximising $\alpha \ln L + S$ is then minimising

$$\chi^2(\boldsymbol{\mu}) + \tau\, \boldsymbol{\mu}^{\mathsf{T}}\mathbf{G}\boldsymbol{\mu}, \qquad \tau = 2/\alpha, \tag{6.71}$$

and since this is quadratic in $\boldsymbol{\mu}$, the minimum has a closed form. Set the gradient to zero:

$$0 \overset{(1)}{=} -2\mathbf{R}^{\mathsf{T}}\mathbf{V}^{-1}(\boldsymbol{n} - \mathbf{R}\hat{\boldsymbol{\mu}}) + 2\tau\mathbf{G}\hat{\boldsymbol{\mu}}$$
$$\Rightarrow \quad \hat{\boldsymbol{\mu}}_\tau \overset{(2)}{=} \left(\mathbf{R}^{\mathsf{T}}\mathbf{V}^{-1}\mathbf{R} + \tau\mathbf{G}\right)^{-1}\mathbf{R}^{\mathsf{T}}\mathbf{V}^{-1}\,\boldsymbol{n} \equiv \mathbf{A}_\tau \boldsymbol{n}. \tag{6.72}$$

(1) The gradient of a quadratic form: $\nabla_\mu(\boldsymbol{a} - \mathbf{R}\boldsymbol{\mu})^{\mathsf{T}}\mathbf{V}^{-1}(\boldsymbol{a} - \mathbf{R}\boldsymbol{\mu}) = -2\mathbf{R}^{\mathsf{T}}\mathbf{V}^{-1}(\boldsymbol{a} - \mathbf{R}\boldsymbol{\mu})$ for symmetric $\mathbf{V}$, and $\nabla_{\boldsymbol{\mu}}\boldsymbol{\mu}^{\mathsf{T}}\mathbf{G}\boldsymbol{\mu} = 2\mathbf{G}\boldsymbol{\mu}$ for symmetric $\mathbf{G}$.
(2) Collect the terms in $\hat{\boldsymbol{\mu}}$ and multiply by the inverse of the bracket.

At $\tau = 0$ the bracket is $\mathbf{R}^{\mathsf{T}}\mathbf{V}^{-1}\mathbf{R}$ and $\mathbf{A}_0 = \mathbf{R}^{-1}$, the naive inverse. The estimator is linear in the data, so its bias and covariance are exact, by the same two steps as (6.64):

$$\boldsymbol{b}_\tau = \mathbb{E}[\hat{\boldsymbol{\mu}}_\tau] - \boldsymbol{\mu} \overset{(1)}{=} (\mathbf{A}_\tau\mathbf{R} - \mathbb{1})\,\boldsymbol{\mu}, \qquad \mathbf{U}_\tau \overset{(2)}{=} \mathbf{A}_\tau\mathbf{V}\mathbf{A}_\tau^{\mathsf{T}}. \tag{6.73}$$

(1) $\mathbb{E}[\mathbf{A}_\tau\boldsymbol{n}] = \mathbf{A}_\tau\boldsymbol{\nu} = \mathbf{A}_\tau\mathbf{R}\boldsymbol{\mu}$.
(2) Covariance of a linear map.

The bias vanishes when $\mathbf{A}_\tau\mathbf{R} = \mathbb{1}$, which happens only at $\tau = 0$; but it also vanishes for any truth with $\mathbf{G}\boldsymbol{\mu} = 0$, the straight lines. The bias formula contains the unknown $\boldsymbol{\mu}$, which is why the bias of an unfolding can be estimated but never known [Cowan §11.6].

The two bins again: how regularisation filters. Go back to the two-bin detector of (6.66) and penalise the difference between the bins, $\tau(\mu_1 - \mu_2)^2 = 2\tau d^2$ (with two bins there is no curvature; the first difference plays its role). In the patterns $\boldsymbol{u}_\pm$, with $\mathbf{V} = \bar{\nu}\mathbb{1}$, the objective (6.71) splits into a part for s and a part for d, and only the second is touched:

$$\frac{\partial}{\partial d}\left[\frac{(n_d - \epsilon d)^2}{\bar{\nu}} + 2\tau d^2\right] \overset{(1)}{=} -\frac{2\epsilon(n_d - \epsilon d)}{\bar{\nu}} + 4\tau d = 0$$
$$\Rightarrow \quad \hat{d}_\tau \overset{(2)}{=} \frac{\epsilon\, n_d}{\epsilon^2 + 2\tau\bar{\nu}} \overset{(3)}{=} F \cdot \frac{n_d}{\epsilon}, \qquad F = \frac{\epsilon^2}{\epsilon^2 + 2\tau\bar{\nu}}. \tag{6.74}$$

(1) The measured difference is $n_d = \epsilon d + \text{noise}$, so its χ^2 term is $(n_d - \epsilon d)^2/\bar{\nu}$.
(2) Solve the linear equation for d.
(3) Multiply and divide by ϵ: the regularised estimate is the naive one, n_d/ϵ, times a factor F between 0 and 1.

So regularisation does not smooth in some vague sense: it multiplies the badly measured pattern by a **filter factor** F, and leaves the well-measured total alone. Hoecker and Kartvelishvili show that this is the general structure [163]: after the data are rescaled to unit errors and $\mathbf{R}$ is decomposed into its patterns (its singular value decomposition), the regularised solution multiplies the k-th pattern of the naive inverse by $s_k^2/(s_k^2 + \tau)$, where s_k is that pattern's gain. Patterns with $s_k^2 \gg \tau$ pass; those with $s_k^2 \ll \tau$ are cut. It is a low-pass filter.

Which F is best? With $E[n_d] = \epsilon d$ and $\mathrm{Var}(n_d) = \bar{\nu}$, the estimate $\hat{d}_\tau$ has bias $(F-1)d$ and variance $F^2\bar{\nu}/\epsilon^2$, so

$$\mathrm{MSE}(F) \overset{(1)}{=} F^2\frac{\bar{\nu}}{\epsilon^2} + (1-F)^2 d^2, \qquad F^* \overset{(2)}{=} \frac{d^2}{d^2 + \bar{\nu}/\epsilon^2}, \qquad \mathrm{MSE}(F^*) \overset{(3)}{=} \frac{d^2\,\bar{\nu}/\epsilon^2}{d^2 + \bar{\nu}/\epsilon^2}. \tag{6.75}$$

(1) Variance plus squared bias (Section 3.2.1).
(2) Set $\mathrm{d}\,\mathrm{MSE}/\mathrm{d}F = 2F\bar{\nu}/\epsilon^2 - 2(1-F)d^2$ to zero.
(3) Insert F^*; both terms share the denominator $(d^2 + \bar{\nu}/\epsilon^2)^2$, and their numerators add to $d^2(\bar{\nu}/\epsilon^2)(d^2 + \bar{\nu}/\epsilon^2)$.

The best filter keeps the pattern in proportion to how much of it is signal: d^2 is the squared size of the true pattern and $\bar{v}/\epsilon^2$ the variance of its naive estimate. With $\bar{v} = 1000$, $\epsilon = 0.1$ and a true difference $\mu_1 - \mu_2 = 200$ ($d^2 = 20\,000$), the naive estimate has MSE $= 10^5$ (an error of 316 in d), setting the difference to zero costs $d^2 = 20\,000$ (an error of 141), and the best filter, $F^* = 1/6$, gets the error down to 129. The optimum strength follows from F^*: $2\tau\bar{v} = \epsilon^2(1/F^* - 1)$ gives $\tau^* = 1/(2d^2)$. It depends on the size of the true structure, which is unknown. This is why choosing τ is the hard part of every unfolding.

Other smoothness measures. The curvature is not the only choice. The entropy of the normalised histogram, $H = -\sum_i p_i \ln p_i$ with $p_i = \mu_i/\mu_{\text{tot}}$, is largest for a flat histogram, and it counts the number of ways $\Omega = \mu_{\text{tot}}!/(\mu_1! \cdots \mu_M!)$ of distributing the events over the bins: with Stirling's $\ln m! \approx m \ln m - m$, $\ln \Omega \approx \mu_{\text{tot}} \ln \mu_{\text{tot}} - \sum_i \mu_i \ln \mu_i = -\sum_i \mu_i \ln(\mu_i/\mu_{\text{tot}}) = \mu_{\text{tot}} H$ [Cowan eqs. 11.51–11.54]. Using $S = H$ is **maximum-entropy** unfolding. It keeps every bin positive, which Tikhonov does not, and it is popular for images with sharp features; read as a Bayesian prior $\pi(\boldsymbol{\mu}) \propto \Omega$ it pulls far too hard towards a flat histogram, so in practice its strength is tuned like τ. A cross-entropy relative to a reference histogram $\boldsymbol{q}$ pulls towards $\boldsymbol{q}$ instead of towards flatness [Cowan §§11.5.2–11.5.4]. Each choice is unbiased for a different kind of truth: curvature for straight lines, entropy for flat histograms, cross-entropy for $\boldsymbol{q}$.

6.14.4 Choosing the strength

In a test case we know the truth, so the exact bias and covariance (6.73) give the exact mean squared error averaged over the bins, $\text{MSE}(\tau) = \frac{1}{M}\sum_i \left(U_{\tau,ii} + b_{\tau,i}^2\right)$ [Cowan eq. 11.79], for every τ. Figure 6.40(b) shows it for our two-peak spectrum. As τ grows the variance falls from the naive inverse's 1.752×10^9 per bin and the squared bias rises; their sum has a minimum at $\tau = 1.413 \times 10^{-5}$, where the root-mean-square error per bin is 72.6 events: a bias of 53.1 and a standard deviation of 49.4. At the optimum the bias and the statistical error are about equal; that is what an optimal trade looks like. For scale, the Poisson error of a single bin with the mean content would be 31.6. Unfolding costs information, and no method gets it back.

Listing 6.16: Tikhonov unfolding as one linear solve (`code/ch06/30_unfolding.py`)

```
C = np.zeros((M - 2, M))
for i in range(M - 2):
    C[i, i:i + 3] = [1.0, -2.0, 1.0]     # second difference mu_i - 2 mu_{i+1} + mu_{i+2}
G = C.T @ C


def tikhonov_matrix(tau, Vdiag):
    """A_tau with mu-hat = A_tau n, for data covariance diag(Vdiag)."""
    W = R.T / Vdiag                       # R^T V^{-1}
    return np.linalg.solve(W @ R + tau * G, W)
```

Lines 119–122 build the second-difference matrix $\mathbf{C}$ of (6.70) and $\mathbf{G} = \mathbf{C}^\mathsf{T}\mathbf{C}$. The function on lines 125–128 returns $\mathbf{A}_\tau$ of (6.72): line 127 forms $\mathbf{R}^\mathsf{T}\mathbf{V}^{-1}$ for a diagonal $\mathbf{V}$, and line 128 solves $(\mathbf{R}^\mathsf{T}\mathbf{V}^{-1}\mathbf{R} + \tau\mathbf{G})\mathbf{A} = \mathbf{R}^\mathsf{T}\mathbf{V}^{-1}$ rather than forming an inverse.

Without the truth. A real analysis does not know $\boldsymbol{\mu}$, so it needs a rule that uses the data only. Three are common. (i) *Estimate the bias.* Replace $\boldsymbol{\mu}$ in (6.73) by the estimate itself, $\hat{\boldsymbol{b}} = (\mathbf{A}_\tau\mathbf{R} - \mathbb{1})\hat{\boldsymbol{\mu}}_\tau$, and lower τ until the estimated biases are no longer significant compared with their own statistical errors [Cowan eqs. 11.76–11.83]. (ii) *The effective rank.* Hoecker and Kartvelishvili rotate the rescaled data into the patterns of $\mathbf{R}$; the coefficients fall with the pattern number until they reach the noise level, which is one. If that happens at pattern k, they set $\tau = s_k^2$, so that the filter

passes exactly the patterns the data can see [163]. They also recommend checking the choice on a test distribution close to, but different from, the simulated one, which is what our toys do. (iii) *The L-curve.* Plot, for each τ, how badly the solution fits the data, $\chi^2(\hat{\boldsymbol{\mu}}_\tau)$, against how rough it is, $|\mathbf{C}\hat{\boldsymbol{\mu}}_\tau|^2$, both on log scales. Small τ gives tiny misfit and huge roughness, large τ the reverse, and in between the curve turns a corner; the corner, the point of largest curvature, is the compromise. For our one data set the corner sits at $\tau = 1.000 \times 10^{-6}$, close to the minimum-MSE value (Figure 6.40c).

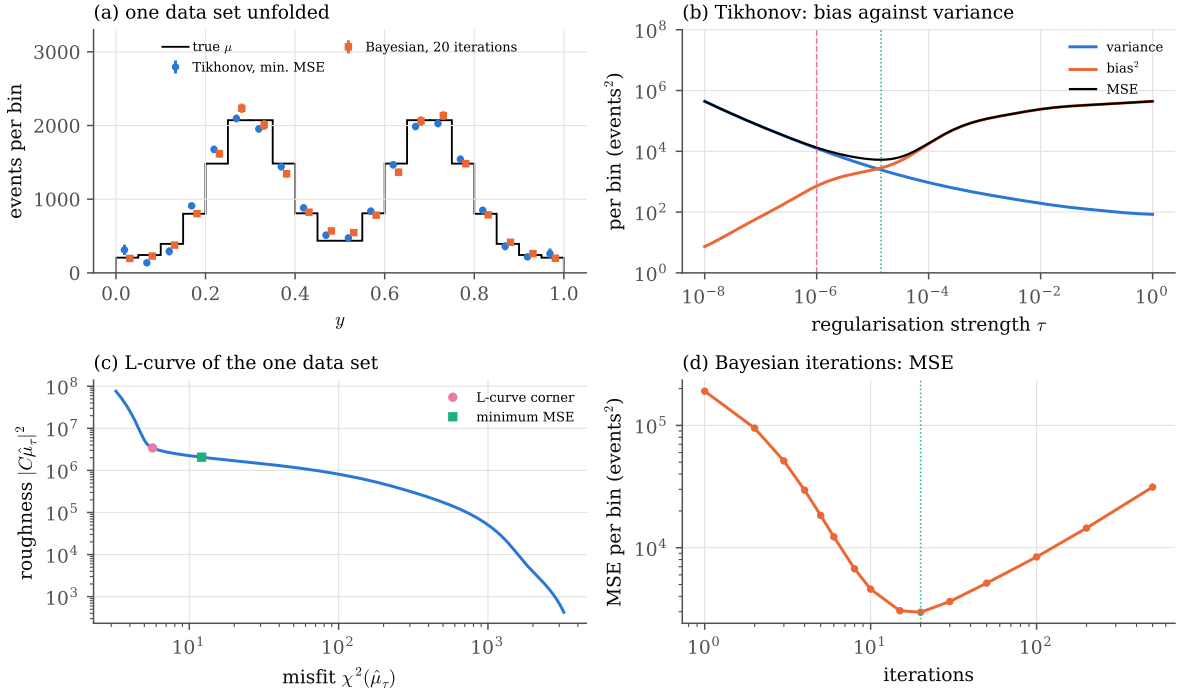


Figure 6.40: (a) The data set of Figure 6.38 unfolded with curvature regularisation at the minimum-MSE strength (blue) and with 20 Bayesian iterations (orange), on the truth (black). (b) Exact variance, squared bias and their sum, the MSE, per bin against τ; the dotted line marks the minimum, the dashed one the L-curve choice. (c) The L-curve of this data set: misfit against roughness as τ runs from small (upper left) to large (lower right). (d) For the Bayesian iterations, the MSE from 1000 toys against the number of iterations.

6.14.5 Unfolding with Bayes' theorem, iterated

D'Agostini's method starts from a different question [141]. *The event.* An event has been recorded in measured bin i. *What could have produced it?* An event created in any of the true bins $j = 1, \ldots, M$. These are the hypotheses. *How probable was the event under each hypothesis?* If the event was created in true bin j, the probability that it was recorded and landed in bin i is the response matrix element R_{ij}: that is exactly what (6.61) defined. *Only then the priors.* Before looking at this event, the probability that an event comes from true bin j is the fraction $p_j = \mu_j/\mu_{\text{tot}}$ of the truth in that bin. We do not know it, so we start from a guess $p_j^{(0)}$, flat for instance. *Normalise.* Bayes' theorem (Chapter 5) gives the probability that an event seen in bin i came from true bin j:

$$\mathrm{P}(j \mid i) = \frac{R_{ij}\, p_j^{(0)}}{\sum_l R_{il}\, p_l^{(0)}}. \tag{6.76}$$

Now share out the data. The n_i events of measured bin i are assigned to the true bins in these proportions, so true bin j receives $\sum_i n_i \, \mathrm{P}(j \mid i)$ recorded events. Recorded events are only the fraction ε_j of those created, so divide by the efficiency:

$$\hat{\mu}_j^{(1)} = \frac{1}{\varepsilon_j} \sum_i n_i \, \mathrm{P}(j \mid i) = \frac{\hat{\mu}_j^{(0)}}{\varepsilon_j} \sum_i R_{ij} \, \frac{n_i}{\hat{\nu}_i^{(0)}}, \qquad \hat{\nu}_i^{(0)} = \sum_l R_{il} \hat{\mu}_l^{(0)}. \tag{6.77}$$

The second form follows by inserting (6.76) with $p_j^{(0)} = \hat{\mu}_j^{(0)} / \hat{\mu}_{\mathrm{tot}}^{(0)}$; the total cancels between numerator and denominator. The result depends on the prior guess, which was the weak point of a Bayesian unfolding. D'Agostini's remedy is to iterate: use $\hat{\boldsymbol{\mu}}^{(1)}$ as the new prior, apply (6.77) again, and so on.

Where do the iterations go? At a fixed point, $\hat{\mu}_j^{(t+1)} = \hat{\mu}_j^{(t)}$, the second form of (6.77) says

$$\frac{1}{\varepsilon_j} \sum_i R_{ij} \frac{n_i}{\hat{\nu}_i} \overset{(1)}{=} 1 \quad \Longleftrightarrow \quad \sum_i R_{ij} \frac{n_i}{\hat{\nu}_i} \overset{(2)}{=} \sum_i R_{ij} \quad \Longleftrightarrow \quad \sum_i R_{ij} \left(\frac{n_i}{\hat{\nu}_i} - 1 \right) = 0, \tag{6.78}$$

(1) Divide the fixed-point condition by $\hat{\mu}_j$ (for a bin with $\hat{\mu}_j \neq 0$).
(2) The efficiency is the column sum, (6.62).

which is the maximum-likelihood condition (6.65). So iterating forever leads towards the maximum-likelihood solution, the naive inverse whenever that is positive, with all its noise. (The iteration is the expectation–maximisation algorithm for Poisson data, known in astronomy as Richardson–Lucy deconvolution.) Stopping early leaves the answer between the prior guess and the maximum-likelihood solution, and the number of iterations plays the role of τ. The smooth patterns, which have large gains, converge in a few iterations; the rough ones, with tiny gains, would take thousands. Stopping early is therefore a filter too.

Listing 6.17: Iterative Bayesian unfolding (`code/ch06/30_unfolding.py`)

```
172 def ibu(n, iters, prior=None):
173     """D'Agostini iterations; n may be (..., M).  Returns the estimate after `iters` steps."""
174     p = np.full(M, 1.0 / M) if prior is None else prior / prior.sum()
175     m_hat = np.broadcast_to(p * n.sum(axis=-1, keepdims=True), n.shape).copy()
176     for _ in range(iters):
177         nu_hat = m_hat @ R.T                                  # sum_j R_ij mu_j
178         m_hat = m_hat / eff * ((n / nu_hat) @ R)              # mu_j/eps_j sum_i R_ij n_i / nu_i
179     return m_hat
```

Line 175 sets the starting guess: the observed total spread flat over the true bins. In each iteration line 177 folds the current guess, $\hat{\nu}_i = \sum_j R_{ij} \hat{\mu}_j$, and line 178 is the second form of (6.77). Written this way it works on a whole stack of toy data sets at once.

Example: how many iterations? With 1000 toy data sets drawn from our $\boldsymbol{\nu}$, the mean squared error per bin falls steeply at first, while the iterations move the flat guess towards the peaks, reaches its minimum of 2969 (an error of 54.5 events per bin) at 20 iterations, and rises again as noise creeps in: after 500 iterations it is 31300 (Figure 6.40d). A small fixed number of iterations, chosen once and for all, would be far from optimal here, because our peaks are narrow and the flat start is far from them. The right number depends on the problem, and has to be chosen with the same care as τ. The error bars of an iterative unfolding cannot be read off a formula as simply as (6.73); we get them by unfolding 400 Poisson resamplings of the data and taking the spread.

The method of correction factors. The simplest unfolding of all, and perhaps the most widely used, has no matrix: multiply each measured bin by a factor from the simulation, $\hat{\mu}_i = C_i(n_i - \beta_i)$ with $C_i = \mu_i^{\mathrm{MC}}/\nu_i^{\mathrm{MC}}$, the ratio of true to measured content of bin i in the simulation [Cowan §11.3]. Its expectation is $C_i \nu_i^{\mathrm{sig}}$, with $\nu_i^{\mathrm{sig}} = \nu_i - \beta_i$ the expected signal in the bin, so its bias is

$$b_i = \left(\frac{\mu_i^{\mathrm{MC}}}{\nu_i^{\mathrm{MC}}} - \frac{\mu_i}{\nu_i^{\mathrm{sig}}} \right) \nu_i^{\mathrm{sig}},$$

zero only if the simulation has the same ratio of true to measured content as nature. When migrations are large this fails badly. With our flat simulation $C_i \approx 1$ in the interior, so the method returns essentially the smeared histogram: the peak bins come out −23.3% low and the central dip 87.4% high. The estimate is pulled towards the shape of the simulation, which may be the very model the measurement was meant to test.

6.14.6 Do the error bars cover? Toys

An error bar is a promise: in about 68% of repeated experiments, the interval $\hat{\mu}_j \pm \hat{\sigma}_j$ should contain the true μ_j (Section 4.12). For a regularised unfolding the promise is in doubt, because $\hat{\sigma}_j$ describes only the spread of $\hat{\mu}_j$, not its offset from the truth. We check it with toys, the Monte Carlo of everything in this chapter. Draw 2000 data sets $n_i \sim \mathrm{Poisson}(\nu_i)$ from the known truth. Unfold each one exactly as a real analysis would, with error bars computed from that data set alone: for the naive inverse and for Tikhonov from (6.64) and (6.73) with $\mathbf{V} = \mathrm{diag}(\boldsymbol{n})$, for the Bayesian iterations from 100 Poisson resamplings, and for the correction factors as $C_i\sqrt{n_i}$. Then count, bin by bin, the fraction of toys whose interval contains μ_j. The binomial error of each fraction is 1%.

The naive inverse covers 67.9% of the time averaged over the bins, against the nominal 68.3%: its enormous error bars are honest. The regularised methods do not keep the promise (Figure 6.41). Tikhonov at the minimum-MSE strength covers 46.1% on average and only 28.8% in the bin at the top of a peak; the worst bin reaches 19.7%. The Bayesian iterations cover 53.4% on average, down to 14.4% in the central dip, while the peak bin is fine (67.2%). The correction factors almost never cover (0.1%): their error bars are the small Poisson errors of the data, and their bias is tens of per cent.

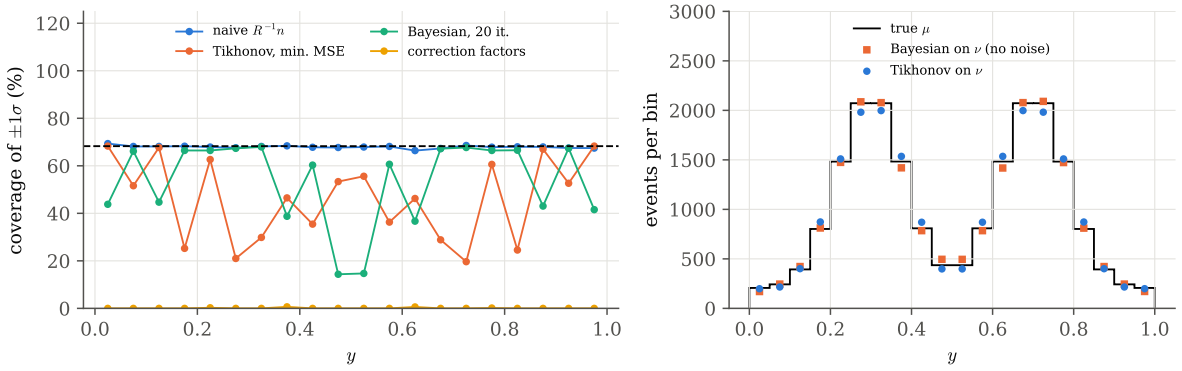


Figure 6.41: Left: the fraction of 2000 toy experiments in which the $\pm 1\sigma$ error bar of each bin contains the true value, for four unfoldings; the dashed line is the nominal 68.3%. Right: the two regularised methods applied to the noise-free expectation $\boldsymbol{\nu}$, which shows their bias alone. Curvature regularisation shaves the peaks; the Bayesian iterations, started flat, leave the dip partly filled.

Why does each method fail where it does? The right panel of Figure 6.41 unfolds the expectation $\boldsymbol{\nu}$ itself, with no noise, so that only the bias is left. Each method is biased towards

what it regards as simple. The curvature penalty prefers straight lines, so it cuts the tops of the peaks, where the curvature is largest. The Bayesian iterations remember their flat start, so they leave too many events in the dip between the peaks. The bias is a property of the method and the truth together, which is why it does not average away over toys and why the coverage fails in different bins for different methods.

Watch out: Regularised error bars are not distances to the truth

The covariance $\mathbf{A}_\tau \mathbf{V} \mathbf{A}_\tau^\mathsf{T}$ of a regularised unfolding is the spread of the estimate about its own biased mean. Quoting it as the uncertainty on the true spectrum undercovers: in our test case one bin is covered 19.7% of the time instead of 68.3%. Either check the coverage with toys and add the bias as a systematic error, or compare the theory with the data the other way round, by folding the theory through $\mathbf{R}$.

That last option is the one to prefer whenever it is possible. To test a theory, nothing needs to be unfolded: fold the theory's prediction through the response, $\boldsymbol{\nu} = \mathbf{R}\boldsymbol{\mu}_{\text{theory}} + \boldsymbol{\beta}$, and compare it with $\boldsymbol{n}$ using the Poisson likelihood. That comparison is unbiased, has the simple Poisson errors of the data, and is exactly the forward model of Example 1.43. Unfolding earns its cost when the result must be compared with other experiments or with theories not yet written, and then it must be published together with its full covariance and, ideally, the response matrix itself [Cowan pp. 153–154].

How this is used in real research: Unfolding in research

Verification is what our test case does: a known truth, a known detector, toys that check the bias, the variance and the coverage of every method. *Research* use looks the same with the truth unknown. Collider experiments unfold differential cross sections (a jet's transverse momentum, a top quark's rapidity) with a response matrix from an event generator and a GEANT4 detector simulation (Section 6.11.4), using either curvature/SVD regularisation [163] or the iterative Bayesian method [141], with the regularisation strength and the coverage checked on toys built from alternative truths. Cosmology meets the same linear algebra: a partial-sky power spectrum is the true C_ℓ multiplied by a mode-coupling matrix of the mask (Chapter 8), and the pseudo-C_ℓ method inverts that matrix only after binning in ℓ, the "wider bins" cure of Section 6.14.2, because the unbinned matrix is as ill-conditioned as a smearing matrix.

The lesson. Unfolding is the backward, inference direction of a forward chain that simulation lets us write down exactly, and a measurement can only return the patterns its detector did not crush, so every answer beyond those patterns is a choice of bias that has to be shown, checked with toys and reported.

6.15 Shrinking σ: variance reduction

The Monte Carlo error $\sigma/\sqrt{N}$ of Section 6.11.2 has two factors, and only one of them is ours to change. With independent random draws the $1/\sqrt{N}$ is fixed: a hundred times the computing time buys one more digit. The constant σ, the standard deviation of a single term, is not fixed. Importance sampling (Section 6.12) changes it by moving the draws to where the integrand lives. Three other cheap devices change it by arranging the draws more cleverly [HobsonBMC §3.1]. We

test each on the same integral,

$$I = \int_0^1 e^x \, \mathrm{d}x = e - 1 = 1.718,$$

where plain Monte Carlo with uniform U has single-term variance $\sigma^2 = \mathbb{E}[e^{2U}] - I^2 = \frac{1}{2}(e^2 - 1) - (e-1)^2 = 0.242$. Every estimator below uses the same number N of evaluations of e^x, so their variances can be compared directly.

The question. Without changing the integrand or the number of function evaluations, how can we make the average less noisy?

6.15.1 Antithetic pairs

If U is uniform, so is $1 - U$. Evaluate the integrand at both and average the pair. Why should this help? For a monotone integrand, a draw that lands high is paired with one that lands low, so their errors partly cancel. To see how much, compute the variance of one pair,

$$\operatorname{Var}\Big[\tfrac{1}{2}\big(h(U) + h(1-U)\big)\Big] \overset{(1)}{=} \tfrac{1}{4}\big[\operatorname{Var} h(U) + \operatorname{Var} h(1-U) + 2\operatorname{Cov}\big(h(U), h(1-U)\big)\big] \overset{(2)}{=} \tfrac{1}{2}\big(\sigma^2 + \gamma\big), \tag{6.79}$$

(1) The variance of a sum of two correlated terms, with the factor $\frac{1}{2}$ squared.
(2) $h(U)$ and $h(1-U)$ both have variance σ^2; write $\gamma = \operatorname{Cov}\big(h(U), h(1-U)\big)$.

With N evaluations we have $N/2$ pairs, and the average of $N/2$ independent pairs has variance $\frac{1}{2}(\sigma^2 + \gamma)/(N/2) = (\sigma^2 + \gamma)/N$. So the variance per evaluation is $\sigma^2 + \gamma = \sigma^2(1 + \rho)$, where $\rho = \gamma/\sigma^2$ is the correlation between the two members of a pair. A negative ρ helps, and for a monotone h it is always negative, because U and $1 - U$ move in opposite directions. For e^x, $\gamma = \mathbb{E}[e^U e^{1-U}] - I^2 = e - (e-1)^2 = -0.2342$, almost $-\sigma^2$, so the variance per evaluation drops from 0.242 to 0.007825, a predicted gain of 31. The simulation, 4000 repeats with $N = 1024$, measures a gain of 29.41 in variance.

6.15.2 Control variates

Suppose a second function $c(x)$ is correlated with the integrand and has a mean μ_c we know exactly. Estimate I by averaging $h(X) - \beta\,[c(X) - \mu_c]$ instead of $h(X)$. The added term has mean zero for every β, so the estimator stays unbiased, and its single-term variance is

$$\begin{aligned} \operatorname{Var}\big[h - \beta(c - \mu_c)\big] &\overset{(1)}{=} \sigma_h^2 - 2\beta \operatorname{Cov}(h, c) + \beta^2 \sigma_c^2 \\ &\overset{(2)}{=} \sigma_h^2 - \frac{\operatorname{Cov}(h,c)^2}{\sigma_c^2} \overset{(3)}{=} \sigma_h^2\big(1 - \rho_{hc}^2\big) \quad \text{at} \quad \beta^* = \frac{\operatorname{Cov}(h,c)}{\sigma_c^2}. \end{aligned} \tag{6.80}$$

(1) The variance of a difference; the constant $\beta\mu_c$ does not contribute.
(2) Minimise over β: the derivative $-2\operatorname{Cov}(h,c) + 2\beta\sigma_c^2$ vanishes at β^*; substitute it back.
(3) $\rho_{hc} = \operatorname{Cov}(h,c)/(\sigma_h \sigma_c)$ is the correlation coefficient.

We therefore use the draws to estimate only the *difference* between the integrand and a function whose integral we already know. The better c tracks h, the closer ρ_{hc}^2 is to one and the smaller the variance that is left.

For $h = e^x$ take $c(x) = x$, with $\mu_c = \frac{1}{2}$ and $\sigma_c^2 = \frac{1}{12}$ for x uniform on $[0,1]$. Then $\operatorname{Cov}(e^U, U) = \mathbb{E}[Ue^U] - \frac{1}{2}(e-1) = 1 - \frac{1}{2}(e-1) = \frac{1}{2}(3-e)$, using $\int_0^1 xe^x \mathrm{d}x = \big[(x-1)e^x\big]_0^1 = 1$, so

$\beta^* = 6(3-e) = 1.69$ and the residual variance is 0.00394, a predicted gain of 61; the simulation measures 59.7.

In physics the control variate is often a cheap approximation of an expensive simulation: a parametrised detector response averaged against the full GEANT4 simulation, or a leading-order cross section whose integral is known analytically.

6.15.3 Stratified sampling

Random draws clump: by chance, some parts of $[0,1]$ receive more points than others, and this clumping is part of σ. Stratification forbids it. Cut $[0,1]$ into K equal strata, draw N/K points uniformly inside each, and average the K stratum averages. Why does this lower the variance? The law of total variance [total variance = mean of the within-stratum variances + variance of the stratum means] splits σ^2 into two parts, and stratification removes the second one. With μ_k and σ_k^2 the mean and variance of h within stratum k,

$$\sigma^2 \overset{(1)}{=} \frac{1}{K}\sum_{k=1}^{K}\sigma_k^2 + \frac{1}{K}\sum_{k=1}^{K}(\mu_k - I)^2, \qquad \mathrm{Var}[\hat{I}_{\mathrm{strat}}] \overset{(2)}{=} \frac{1}{K^2}\sum_{k=1}^{K}\frac{\sigma_k^2}{N/K} \overset{(3)}{=} \frac{1}{N}\Big[\sigma^2 - \frac{1}{K}\sum_{k}(\mu_k - I)^2\Big] \le \frac{\sigma^2}{N}. \tag{6.81}$$

(1) The law of total variance, with the stratum chosen uniformly at random (each has probability $1/K$).
(2) $\hat{I}_{\mathrm{strat}} = \frac{1}{K}\sum_k \hat{\mu}_k$, the strata are sampled independently, and each stratum mean $\hat{\mu}_k$ of N/K draws has variance $\sigma_k^2/(N/K)$.
(3) Simplify and substitute the first equation.

The variation *between* strata, which plain sampling pays for through its random clumping, is gone; only the variation within each stratum is left.

How fast does the error then fall? Take one point per stratum ($K = N$) and a smooth integrand. Each stratum has width $1/N$, so h varies across it by about h'/N, and $\sigma_k^2 \approx h'(x_k)^2/(12N^2)$, the variance of a uniform spread over that range. Then $\mathrm{Var}[\hat{I}_{\mathrm{strat}}] \approx \frac{1}{N^2}\sum_k h'(x_k)^2/(12N^2) \approx \int_0^1 h'^2 \mathrm{d}x/(12N^3)$, so the error falls as $N^{-3/2}$ instead of $N^{-1/2}$. For e^x, $\int_0^1 e^{2x}\mathrm{d}x = \frac{1}{2}(e^2-1)$, which predicts a standard deviation $0.516\,N^{-3/2} = 1.6\times10^{-5}$ at $N = 1024$; the simulation measures 1.556×10^{-5}, a fitted slope of -1.504, and a gain in variance of 9.486×10^5 over plain sampling. The evenly spread quasi-random (Sobol) points of Section 6.7 push the same idea into many dimensions, where a full grid of strata would again cost n^d points.

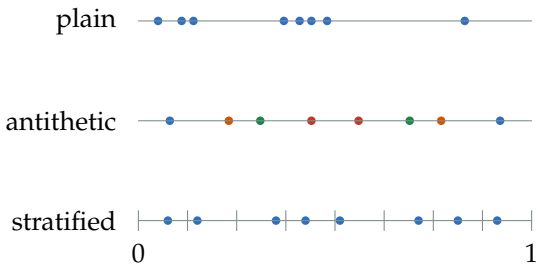


Figure 6.42: Eight evaluation points on $[0,1]$ three ways. Plain sampling can leave gaps and clumps. Antithetic sampling pairs each u with $1-u$ (same colour), so a high draw is always balanced by a low one. Stratified sampling puts exactly one point in each of eight equal strata (tick marks), which removes the clumping altogether.

Listing 6.18: Four estimators of $\int_0^1 e^x \mathrm{d}x$ with the same number of function evaluations (code/ch06/29_variance_reduction.py)

```
29 def plain(n):
30     return np.exp(rng.random((R, n))).mean(axis=1)
31
32
33 def antithetic(n):
34     u = rng.random((R, n // 2))
35     return (0.5 * (np.exp(u) + np.exp(1 - u))).mean(axis=1)
36
37
38 def control(n):
39     u = rng.random((R, n))
40     y = np.exp(u)
41     beta = 6 * (3 - np.e)                          # Cov(e^U, U) / Var(U) = [(3 - e)/2] /
       (1/12), exact
42     return (y - beta * (u - 0.5)).mean(axis=1)
43
44
45 def stratified(n):
46     u = (np.arange(n)[None, :] + rng.random((R, n))) / n
47     return np.exp(u).mean(axis=1)
```

Each function returns 4000 independent estimates, one per row of random numbers, so that the standard deviation of each estimator can be measured directly. Line 34 draws only $n/2$ uniforms and line 35 evaluates e^x at u and $1-u$. Line 42 subtracts the control variate with the exact β^* of (6.80). Line 46 places one uniform point inside each of the n strata $[k/n, (k+1)/n)$.

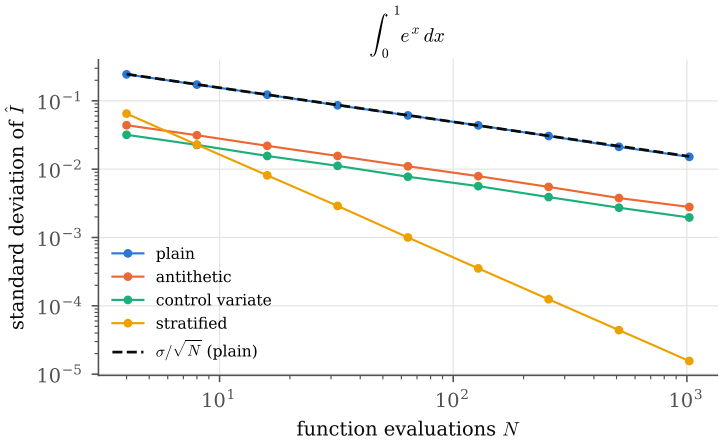


Figure 6.43: Standard deviation of four estimators of $\int_0^1 e^x\,\mathrm{d}x$ against the number of function evaluations, each from 4000 repeats. Plain Monte Carlo follows $\sigma/\sqrt{N}$ (black dashed). Antithetic pairs and the control variate keep the slope $-1/2$ but lower the line by factors of about 5.5 and 8. Stratification with one point per stratum changes the slope to $-3/2$.

Watch out: Error bars for correlated draws

All three devices make the evaluations dependent: pairs are anticorrelated, strata are not exchangeable. The naive error bar $s/\sqrt{N}$ computed from all N values as if they were independent is then wrong. Compute the error from the independent units: the $N/2$ pair averages, the per-stratum variances in (6.81), or, as in the script, from repeated independent runs.

6.16 Which sampler when

The choice of sampler is usually forced by what we know about the density. Every sampler above answers the same question, how to get draws or averages under a density we can describe, but each asks for something different about that density.

Taxonomy: Which sampler when

method	what it needs	cost per draw	fails when
inverse CDF	F^{-1} in closed form or fast to solve; any discrete table	one uniform	F not invertible (Gaussian, most posteriors)
Box–Muller, polar	nothing: Gaussians	two uniforms per two normals	never; beware approximate shortcuts
Cholesky	covariance $\mathbf{C}$ ($n \times n$)	$n^3/3$ once, n^2 per draw	n large: use a diagonalising basis ($a_{\ell m}$, Fourier)
rejection	values of f up to a constant, envelope g, $M = \sup f/g$	M candidates	f/g unbounded; high dimension (M grows exponentially)
plain Monte Carlo integration	draws from f	error $\sigma/\sqrt{N}$ in any dimension	h concentrated where f rarely goes (σ huge)
importance sampling	values of f (up to a constant for the self-normalised form), g like $\lvert h\rvert f$	weights f/g	g thinner-tailed than f (infinite variance)
variance reduction	structure: monotone h, known $\int c$, smoothness	same N, smaller σ	error bars computed as if draws were independent
MCMC (Chapter 9)	values of f up to a constant, nothing else	correlated draws	poor mixing, multimodality

Going down the table, we need to know less and less about f, until Markov chain Monte Carlo needs only the ratio of f at two points, which is the posterior without its evidence (Chapter 5). The price is that the draws become correlated, and the $1/\sqrt{N}$ law then holds with N replaced by an effective number of independent draws [Lambert §12.2].

Recap

- $F(X)$ is uniform for any continuous X; so $X = F^{-1}(U)$ has CDF F. Decay times: $-\tau \ln(1-U)$. Power laws, isotropic directions and discrete tables follow the same way.
- Two uniforms give two exact Gaussians (Box–Muller): $R^2 = -2\ln U_1$ is exponential and the angle is uniform, by the polar Jacobian. $\boldsymbol{\mu} + \mathbf{L}\boldsymbol{z}$ with $\mathbf{L}\mathbf{L}^{\mathsf{T}} = \mathbf{C}$ gives any Gaussian vector.
- Rejection: points uniform under Mg and kept under f are exact draws from f; acceptance $1/M$, geometric number of trials; needs f/g bounded.
- An integral is an expectation; its Monte Carlo estimate has error $\sigma/\sqrt{N}$ in any dimension,

beating grids ($N^{-2/d}$) above about four dimensions. The acceptance of a detector is the verified example; with a dead region the simulation is the answer.

- Importance sampling reweights by f/g; its variance contains $\int h^2 f^2/g$, so g must have thicker tails than f. Self-normalised weights handle unnormalised posteriors.
- A histogram of samples estimates the bin-averaged density with binomial error $\sqrt{Np_k(1-p_k)}/(N\Delta)$; the oscillator at random times reproduces the arcsine density.
- Unfolding: the measured histogram is $\boldsymbol{\nu} = \mathbf{R}\boldsymbol{\mu} + \boldsymbol{\beta}$ with $R_{ij} = \mathrm{P}(\text{observed } i \mid \text{true } j)$ from simulation. The inverse $\mathbf{R}^{-1}\boldsymbol{n}$ is unbiased but divides each pattern by its gain ($e^{-q^2\sigma^2/2}$ for Gaussian smearing), so its variance explodes. Tikhonov curvature and iterated Bayes trade bias for variance; choose the strength by MSE or the L-curve, and check the coverage with toys.
- Antithetic pairs, control variates and stratification shrink σ at fixed N.

Reading the literature

- **The main line.** Wasserman, *All of Statistics*, Chapter 24: basic Monte Carlo integration, importance sampling with the tail rule, and the start of MCMC [Wasserman pp. 403–416]; Section 11.4 on simulating the posterior [Wasserman p. 180].
- **The physicist's chapter.** Cowan, *Statistical Data Analysis*, Chapter 3: uniform generators, the transformation method, acceptance–rejection, the $1/\sqrt{n}$ argument against quadrature, event generators and detector simulation.
- **Derivations.** Casella and Berger, *Statistical Inference*, Section 5.6: the probability integral transform, Box–Muller, discrete generation, the Accept/Reject theorem with its proof and the geometric number of trials.
- **Why sampling.** Lambert, *A Student's Guide to Bayesian Statistics*, Chapter 12 (the intractable denominator, grids, quadrature, independent sampling); Kruschke, *Doing Bayesian Data Analysis*, Section 7.1 (a distribution represented by a representative sample); Hobson et al. (eds.), *Bayesian Methods in Cosmology*, Chapter 3 (Monte Carlo methods for parameter estimation).
- **Generators.** Marsaglia on the lattice structure of congruential generators [151]; L'Ecuyer on combined and well-chosen LCGs [150, 152]; the Mersenne Twister [154]; xorshift [153]; PCG [156]; counter-based generators [157]; the TestU01 battery [155]; Sobol's sequences [158].
- **Sampling methods and physics codes.** Box and Muller [159]; Marsaglia and Bray's polar method [160]; the Pythia 8.3 manual [161] and the Geant4 toolkit paper [162] for sampling at the scale of a collider experiment; Metropolis et al. [36] for where Chapter 9 goes next.

Chapter summary

Turning a probability model into numbers takes three pieces of machinery. First, uniform numbers: a computer is deterministic, so its "random" numbers are the output of a finite machine, and they must be tested like any other null hypothesis. Second, transformations: given good uniform numbers, a few exact transformations turn them into draws from any distribution we can describe. Third, averages: averages over those draws give integrals, with an error $\sigma/\sqrt{N}$ that does not depend on the dimension. In the pipeline from physical model to

inference,

$$\text{physical model} \to \text{field or dataset} \to \text{summary } S \to \boxed{P(S \mid \theta)} \to \text{infer physics},$$

Monte Carlo is the arrow into the sampling distribution: whenever the algebra of Chapters 2 and 3 gives out, we simulate the experiment many times and histogram the summary. Chapters 7 and 8 do this for fields and CMB maps, and Chapter 9 adapts it to densities we can only evaluate.

Random numbers on a deterministic machine

concept	operational meaning	used later in
Monty Hall as a test of the generator	The observed event is "Monty opened C". What made him open that door? The possible hypotheses are the three locations of the car, H_A, H_B, H_C. Under each we find the probability that Monty opens C: $\frac{1}{2}$ under H_A (he may open either goat door and tosses a coin), 1 under H_B (he must not open our door A and must not reveal the car), 0 under H_C (the car is there). Only then come the priors ($\frac{1}{3}$ each) and the normalisation, which give posterior probabilities $\frac{1}{3}, \frac{2}{3}, 0$. A simulation reproduces this only if its coin is fair, because a biased coin changes the probability of the event under H_A	every chapter
pseudorandom generator	a finite state, a transition function, an output function and a seed; reproducible, eventually periodic	every simulation
linear congruential generator	$x_{n+1} = (a x_n + c) \bmod m$; period at most m; low bits of a power-of-two modulus are poor	legacy codes
lattice structure, spectral test	t-tuples of an LCG lie on a lattice of parallel planes (RANDU: 15 planes in 3-D); the spectral test measures the largest gap between planes	Chapters 7 and 8
modern generators	bigger state and a nonlinear output function: xorshift, Mersenne Twister, PCG64 (numpy default), counter-based	every chapter
seeds and streams	SeedSequence hashes job identifiers into well-separated states; one stream per parallel job, never "seed = job number" by hand	Chapters 8 and 9
randomness tests	each test is a null hypothesis "the stream is i.i.d. uniform": χ^2 on histograms, serial correlation, runs; batteries such as TestU01	Chapter 4
quasi-random points	Sobol sequences fill the cube evenly; integration error close to $1/N$ instead of $1/\sqrt{N}$ for smooth integrands	Chapter 10

From uniform numbers to any distribution

concept	operational meaning	used later in
probability integral transform	$F(X) \sim \text{Uniform}(0,1)$ for continuous X; $F^{-1}(U)$ has CDF F	Chapters 4 and 8
inverse-CDF sampling	solve $F(x) = u$: decay times $-\tau \ln(1-u)$, power laws, isotropic directions ($\cos\theta = 2u - 1$), discrete tables by cumulative sums	every simulation
Box–Muller, polar method	$R = \sqrt{-2\ln U_1}$, $\Theta = 2\pi U_2$ give two exact independent normals; the polar form keeps a fraction $\pi/4$	Chapters 7 to 9
Cholesky colouring	$\boldsymbol{\mu} + \mathbf{L}\boldsymbol{z}$ with $\mathbf{L}\mathbf{L}^\mathsf{T} = \mathbf{C}$ has covariance $\mathbf{C}$; for isotropic fields the factor is $\sqrt{C_\ell}$ per mode	Chapters 7, 8 and 10
rejection sampling	keep $V \sim g$ if $U\,M g(V) < f(V)$; acceptance $1/M$; trials geometric; needs f/g bounded; hopeless in high dimension	Chapter 9

Integrals, errors and densities from samples

concept	operational meaning	used later in
Monte Carlo integration	write $I = \mathbb{E}_f[h]$, average h over draws from f, report $\hat{I} \pm s/\sqrt{N}$	every chapter
$1/\sqrt{N}$ law	$\mathrm{Var}\,\hat{I} = \sigma^2/N$ from independence, Gaussian by the CLT; the rate is dimension-independent, the constant σ is not; grids lose above $d \approx 4$	Chapters 8 to 10
probability as a mean of an indicator	simulate and count; binomial error $\sqrt{p(1-p)/N}$ (acceptance, $\Phi(2)$, p-values by simulation)	Chapters 4 and 8
verification vs research	detector acceptance: the simulation checks $\Omega/4\pi$, then is the answer with a dead strip; event generators and detector simulation	Chapter 8
importance sampling	average hf/g over draws from g; variance $\int h^2 f^2/g - I^2$; g must have thicker tails than f; effective sample size $(\sum w)^2/\sum w^2$	Chapter 9
self-normalised weights	posterior averages $\sum \theta_j w_j / \sum w_j$ with $w = \mathcal{L}\pi/g$; the evidence cancels; reweighting of simulated events	Chapter 9
histogram as density	$n_k/(N\Delta)$ estimates the bin average p_k/Δ with error $\sqrt{N p_k(1-p_k)}/(N\Delta)$; oscillator position from random times; summaries (mode, HDI) from draws	Chapters 7, 9 and 12
variance reduction	antithetic pairs $\sigma^2(1+\rho)$, control variates $\sigma^2(1-\rho^2)$, stratification removes between-strata variance ($N^{-3/2}$ for smooth h)	Chapter 8

Chapter 7

Random Fields and the Angular Power Spectrum

Where this chapter is going

A cosmological model predicts how strongly microwave temperatures in two directions a degree apart tend to be correlated. We can use that prediction even though the model leaves open which direction, Orion or Leo, happens to be hotter in our sky. Predictions for turbulent velocities, thermal ripples on a membrane and detector noise have the same character. Theory gives us the *statistics* of a field, while a measurement gives us a particular realisation from which to estimate them.

The object that carries the prediction is a random field: a random variable at every point of space, with a joint distribution for any finite set of points. When the statistics do not care where we stand or which way we look, the relation between two points depends only on their separation, and one function of distance, the correlation function $\xi(r)$, summarises it. For a Gaussian field that function is the whole story. Fourier modes turn the story into arithmetic: in a homogeneous field different wavevectors are uncorrelated, and the variance of each mode is the power spectrum $P(k)$, the Fourier transform of $\xi(r)$. Independent modes are also what make simulation cheap: colour white noise by $\sqrt{P(k)}$ and transform back. Because we live in one universe, the ensemble averages that define ξ and P must be replaced by averages over space, which is legitimate only when distant regions are nearly independent, and which leaves an irreducible scatter of $\sqrt{2/N_{\text{modes}}}$ on every measured power. Real maps are always smoothed by a beam, a pixel or a counting cell, and smoothing multiplies the power spectrum by $|W(k)|^2$. Finally, the power spectrum keeps the amplitude of each Fourier mode and throws away its phase. For a Gaussian field nothing is lost; for any other field the phases hold the shapes, and that is where the topological summaries of Chapter 12 will look.

On the book's pipeline (Chapter 0), where we decide what in a field carries the physics, compress it into a statistic S and work out how S scatters, $P(S \mid \theta)$, the two-point statistics $\xi(r)$, $P(k)$ and, on the sphere, C_ℓ are the first summaries of a field, and their sampling distributions can be derived exactly in the ideal case of a full, noise-free Gaussian field. Chapter 8 adds the mask, the beam and the noise.

7.1 A theory predicts the statistics of a field

Run CAMB with the Planck parameters (Chapter T1) and ask it for the temperature of the microwave sky in the direction of the Galactic north pole. It cannot answer. What it returns is the angular power spectrum C_ℓ: how much the temperature fluctuates, on average, on

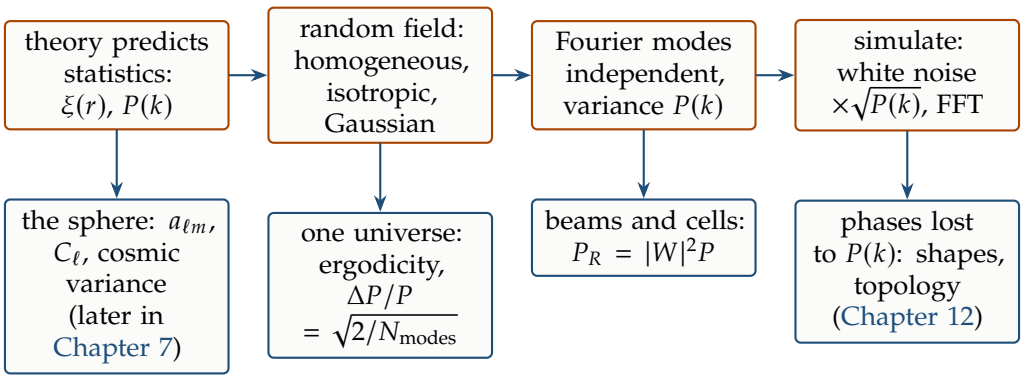


Figure 7.1: How the ideas of the chapter connect. The orange boxes are the backbone: the theory predicts statistics, a homogeneous field has independent Fourier modes, and independent modes make simulation cheap. The blue boxes are what happens when we confront one finite, smoothed realisation, and what the two-point summaries cannot see.

each angular scale. The particular pattern of hot and cold spots that the Planck satellite photographed was set by quantum fluctuations during inflation, and no theory predicts which way those fluctuations happened to fall. The same holds for the matter density of the universe (the theory gives $P(k)$, not the position of the Coma cluster), for the velocity field in a turbulent jet (Kolmogorov gives the spectrum of eddies, not the eddy behind the third bolt), and for the voltage noise in a detector (the noise spectral density, not the next sample). In each case the prediction is a probability distribution over *whole fields*, and the data are one draw from it.

So every inference problem about a field has the same shape. What we observed is one map. The candidate explanations are the possible parameter values, and each one fixes a probability distribution over maps. For each candidate we must be able to say how probable a map like ours would have been. Two things are needed before any of that: a way to write down a probability distribution over functions of position, and a way to compress such a distribution into a few numbers that we can compute from one map.

7.1.1 A random field is a random variable at every point

Think of a temperature map stored on a grid of n pixels. As far as statistics is concerned it is a list of n numbers, so a random map is a random vector $\big(\delta(\boldsymbol{x}_1), \dots, \delta(\boldsymbol{x}_n)\big)$, exactly the kind of object whose joint density, means and covariance matrix we met in Chapter 1 and whose Gaussian version we built in Chapter 2. A continuous field is the same idea with a random variable attached to every point of space. Any question we actually ask involves finitely many points at a time, so the definition only needs to say what happens at finite sets of points.

Working definition 7.1 (random field)

A **random field** $\delta(\boldsymbol{x})$ on a domain $\mathcal{M}$ (a line, a plane, a box, the sphere) is a collection of random variables, one for each point $\boldsymbol{x} \in \mathcal{M}$, together with the joint density $p\big(\delta(\boldsymbol{x}_1), \dots, \delta(\boldsymbol{x}_n)\big)$ for every finite set of points $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ [the n-point distributions]. These must be consistent: integrating $\delta(\boldsymbol{x}_n)$ out of the n-point density gives the $(n-1)$-point density of the remaining points. *Operationally:* a random field is a recipe that, run once, produces a whole map, and run many times produces many different maps with the same statistics.

The consistency condition is marginalisation (Chapter 1). To study the field at a few points, we average out the field at every other point, treating it as a nuisance variable; the answer must not depend on how many other points we started with. Throughout, δ is a *fluctuation*. For

matter it is the density contrast $\delta(\boldsymbol{x}) = \rho(\boldsymbol{x})/\bar{\rho} - 1$, the fractional excess of density over its mean $\bar{\rho}$; for the CMB it is $\Delta T/T_0$; for a detector it is the voltage minus its average. A fluctuation has zero mean by construction.

To picture a random field, think of the **ensemble**: an imaginary stack of universes, all made by the same physics but with different random draws (Figure 7.2). "The mean of δ at $\boldsymbol{x}$" means the average down the stack: fix the point $\boldsymbol{x}$ and average over all the universes. It does *not* mean the average over the points of our one map. Whether the two averages agree is a real question, and Section 7.5 answers it.

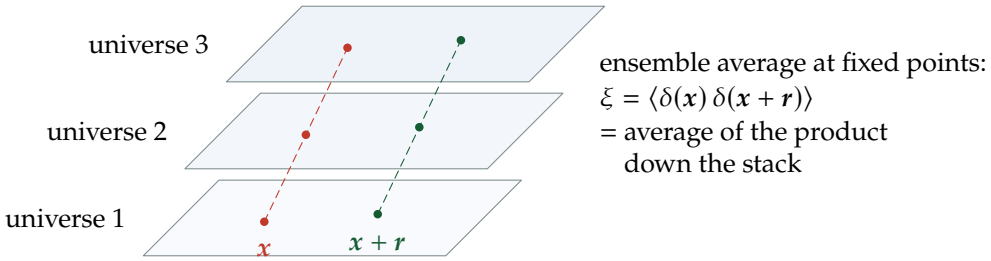


Figure 7.2: The ensemble. Each sheet is one universe, a complete realisation of the field. Statistical statements about the points $\boldsymbol{x}$ (red) and $\boldsymbol{x} + \boldsymbol{r}$ (green) are averages down the stack, along the dashed lines, not across one sheet.

The simplest summaries of a random field are the ones we know for random vectors, now indexed by position. The **mean** is $\mu(\boldsymbol{x}) = \mathbb{E}[\delta(\boldsymbol{x})]$, written $\langle\delta(\boldsymbol{x})\rangle$ in physics, with angle brackets meaning the ensemble average. The **covariance function** is the covariance of the field at two points,

$$C(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{Cov}\big(\delta(\boldsymbol{x}), \delta(\boldsymbol{y})\big) = \langle\delta(\boldsymbol{x})\delta(\boldsymbol{y})\rangle - \mu(\boldsymbol{x})\mu(\boldsymbol{y}), \tag{7.1}$$

the definition of covariance of [Cowan (1.46)] and [Wasserman p. 52, Def. 3.18] with the two random variables being the field at two places. On the diagonal it is the variance, $C(\boldsymbol{x}, \boldsymbol{x}) = \mathrm{Var}\,\delta(\boldsymbol{x})$. On a grid of n pixels, $C(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is simply the $n \times n$ covariance matrix $\mathbf{C}_{ij}$ of the map.

7.1.2 Statistical homogeneity and isotropy

A general covariance function depends on two points, which is far too much to measure from one map. A symmetry usually saves us. The laws that made the field are the same everywhere and in every direction, so the *statistics* of the field should be the same everywhere and in every direction too, even though any one realisation is lumpy.

Working definition 7.2 (statistical homogeneity and isotropy)

A random field is **statistically homogeneous** if all its n-point distributions are unchanged by a common shift of the points, $p\big(\delta(\boldsymbol{x}_1 + \boldsymbol{a}), \ldots, \delta(\boldsymbol{x}_n + \boldsymbol{a})\big) = p\big(\delta(\boldsymbol{x}_1), \ldots, \delta(\boldsymbol{x}_n)\big)$ for every $\boldsymbol{a}$ [the ensemble cannot tell where we are standing]. It is **statistically isotropic** if they are also unchanged by a common rotation of the points about any centre [the ensemble cannot tell which way we are looking]. *Test:* the mean is the same at every point, and the covariance of two points depends only on their separation vector (homogeneity), and only on its length (isotropy).

The test in the definition follows in two lines. Homogeneity applied to the one-point distribution says that $\delta(\boldsymbol{x} + \boldsymbol{a})$ has the same distribution as $\delta(\boldsymbol{x})$, so the mean is a constant, which for a fluctuation is zero. Applied to the two-point distribution with the shift $\boldsymbol{a} = -\boldsymbol{y}$:

$$C(\boldsymbol{x}, \boldsymbol{y}) \overset{(1)}{=} \langle\delta(\boldsymbol{x})\delta(\boldsymbol{y})\rangle \overset{(2)}{=} \langle\delta(\boldsymbol{x} - \boldsymbol{y})\,\delta(\boldsymbol{0})\rangle \overset{(3)}{=} \xi(\boldsymbol{x} - \boldsymbol{y}). \tag{7.2}$$

(1) The mean is zero, so the covariance (7.1) is the average of the product.
(2) Homogeneity: the pair of points $(\boldsymbol{x}, \boldsymbol{y})$ and the pair shifted by $-\boldsymbol{y}$, which is $(\boldsymbol{x}-\boldsymbol{y}, \boldsymbol{0})$, have the same joint distribution, so the average of the product is the same.
(3) The result depends on $\boldsymbol{x}$ and $\boldsymbol{y}$ only through the separation $\boldsymbol{r} = \boldsymbol{x}-\boldsymbol{y}$; we give that function its own name, $\xi(\boldsymbol{r})$.

Isotropy adds that rotating the separation vector does not change the answer, so ξ depends only on the distance $r = |\boldsymbol{r}|$. This is the **two-point correlation function**,

$$\xi(r) = \langle \delta(\boldsymbol{x})\, \delta(\boldsymbol{x}+\boldsymbol{r}) \rangle, \tag{7.3}$$

the same at every $\boldsymbol{x}$ and in every direction of $\boldsymbol{r}$. On the sky the separation is an angle, and (7.3) becomes $\langle \frac{\Delta T}{T_0}(\hat{\boldsymbol{x}}) \frac{\Delta T}{T_0}(\hat{\boldsymbol{y}}) \rangle = C(\arccos \hat{\boldsymbol{x}} \cdot \hat{\boldsymbol{y}})$, with isotropy embodied in the dependence on the angle alone [HobsonBMC p. 230, (10.1)].

The question ξ answers. Pick two points of the field a distance r apart. How strongly is the fluctuation at the first point related to the fluctuation at the second? If the field is above its mean at one point, how much does that tell us about whether it is above its mean at the other? The correlation function answers this with one number for each separation r.

A homogeneous field is not a smooth one. Each realisation in Figure 7.3 below is full of hot and cold patches. Homogeneity says only that a patch is equally likely to sit anywhere, so that, averaged over the ensemble, no place is special.

Properties every correlation function has. Four facts follow straight from (7.3), and they are the checks to run on any ξ someone hands us.

1. $\xi(0) = \langle \delta^2 \rangle = \sigma^2$, the variance of the field at a point.
2. $\xi(-\boldsymbol{r}) = \xi(\boldsymbol{r})$: put $\boldsymbol{x}' = \boldsymbol{x} + \boldsymbol{r}$ in (7.3), so $\langle \delta(\boldsymbol{x})\delta(\boldsymbol{x}+\boldsymbol{r}) \rangle = \langle \delta(\boldsymbol{x}'-\boldsymbol{r})\delta(\boldsymbol{x}') \rangle = \xi(-\boldsymbol{r})$.
3. $|\xi(r)| \le \xi(0)$. The Cauchy–Schwarz inequality for averages, $\langle XY \rangle^2 \le \langle X^2 \rangle \langle Y^2 \rangle$, with $X = \delta(\boldsymbol{x})$, $Y = \delta(\boldsymbol{x}+\boldsymbol{r})$ gives $\xi(r)^2 \le \sigma^2 \cdot \sigma^2$. Equivalently, the **correlation coefficient** $\rho(r) = \xi(r)/\xi(0)$ lies in $[-1, 1]$, as every correlation coefficient must [Wasserman p. 53, Thm. 3.19].
4. ξ is *positive definite*: for any points $\boldsymbol{x}_i$ and real weights a_i,

$$0 \overset{(1)}{\le} \operatorname{Var}\Big(\sum_i a_i\, \delta(\boldsymbol{x}_i) \Big) \overset{(2)}{=} \sum_{i,j} a_i a_j \operatorname{Cov}\big(\delta(\boldsymbol{x}_i), \delta(\boldsymbol{x}_j) \big) \overset{(3)}{=} \sum_{i,j} a_i a_j\, \xi(\boldsymbol{x}_i - \boldsymbol{x}_j). \tag{7.4}$$

   (1) A variance is never negative.
   (2) The variance of a linear combination is the double sum of covariances [Wasserman p. 53, Thm. 3.20].
   (3) Homogeneity, (7.2).

   So the matrix $\xi(\boldsymbol{x}_i - \boldsymbol{x}_j)$ built from any set of points is a valid covariance matrix. Not every even function bounded by $\xi(0)$ passes this test; Section 7.3 turns it into a one-line check in Fourier space.

Example: white noise. A CCD with independent read noise in each pixel, of standard deviation σ_n, is the simplest random field. Different pixels are independent, so $\langle \delta_i \delta_j \rangle = 0$ for $i \ne j$ and σ_n^2 for $i = j$:

$$\xi(\boldsymbol{r}) = \sigma_n^2\, \delta_{\boldsymbol{r},\boldsymbol{0}},$$

a spike at zero separation and nothing elsewhere ($\delta_{r,\mathbf{0}}$ is the Kronecker delta, 1 if $\boldsymbol{r} = \mathbf{0}$ and 0 otherwise). It is homogeneous and isotropic, and knowing one pixel tells us nothing about its neighbour.

Example: a detector that averages. Suppose instead each recorded sample is the average of m consecutive white-noise values, $\delta_i = \frac{1}{m}\sum_{j=0}^{m-1} w_{i+j}$, with w_j independent, mean zero and variance σ_w^2. This is what an integrating amplifier, or a one-dimensional beam of width m pixels, does. For a lag $r \geq 0$,

$$\xi(r) = \langle \delta_i\, \delta_{i+r}\rangle \overset{(1)}{=} \frac{1}{m^2}\sum_{j=0}^{m-1}\sum_{l=0}^{m-1}\langle w_{i+j}\, w_{i+r+l}\rangle \overset{(2)}{=} \frac{\sigma_w^2}{m^2}\,\#\{(j,l):\ j = r+l,\ 0\leq j,l\leq m-1\}$$
$$\overset{(3)}{=} \frac{\sigma_w^2}{m^2}\,(m-r) \quad (0\leq r<m), \qquad \xi(r) = 0 \quad (r\geq m). \tag{7.5}$$

(1) Insert the definition of both samples and use linearity of the average.
(2) $\langle w_a w_b\rangle$ is σ_w^2 when $a = b$ and zero otherwise, so the double sum counts the pairs with $i + j = i + r + l$, that is $j = r + l$.
(3) For each $l = 0, \ldots, m-1-r$ there is exactly one $j = r + l \leq m - 1$, so there are $m - r$ such pairs; when $r \geq m$ there are none.

The correlation function is a triangle: $\xi(0) = \sigma_w^2/m$ (averaging m independent values divides the variance by m), it falls linearly, and it vanishes beyond the averaging length, where two samples share no input. For $m = 4$ and $\sigma_w = 1$: $\xi = 0.25,\ 0.1875,\ 0.125,\ 0.0625,\ 0$ at $r = 0, \ldots, 4$. The lesson holds for any smoothing: averaging over a length makes the field correlated over that length (Section 7.6).

Example: when the symmetries fail. Real data can break both symmetries, even when the underlying field has them. Homogeneity first. A galaxy survey observes the sky to different depths in different directions, so the noise added to the density map has a variance $\sigma_n^2(\boldsymbol{x})$ that changes with position. The covariance of the observed map is $\xi(\boldsymbol{x}-\boldsymbol{y}) + \sigma_n^2(\boldsymbol{x})\delta_{\boldsymbol{xy}}$, with $\delta_{\boldsymbol{xy}}$ the Kronecker delta (noise in different pixels is independent). This depends on where the pair is, not only on its separation, so the observed field is *not* homogeneous even if the underlying density is. Isotropy fails in a subtler way. Galaxy distances are inferred from redshifts. Galaxies falling into dense regions have extra (peculiar) velocities along our line of sight, and these squash structures along that direction. So ξ measured in redshift space depends on the angle between $\boldsymbol{r}$ and the line of sight [108]. Both effects are real, and both are handled by modelling them on top of a homogeneous, isotropic field.

Seeing the ensemble. The script `code/ch07/01_ensemble_two_points.py` makes $M = 3000$ universes: two-dimensional fields on 128×128 unit cells with the Gaussian correlation function $\xi(r) = e^{-r^2/2\ell^2}$, $\ell = 4$ cells. (Section 7.4 shows how such fields are made; here we only look at them.) The question is the one ξ answers: does the value at one point tell us about the value a distance r away? At a fixed point $\boldsymbol{x}_0$ the script records $\delta(\boldsymbol{x}_0)$ and $\delta(\boldsymbol{x}_0 + \boldsymbol{r})$ in every universe, for $r = 2$, 6 and 15 cells, and plots one against the other (Figure 7.3). At $r = 2$ the cloud is a narrow ellipse: a hot $\boldsymbol{x}_0$ almost always has a hot neighbour. At $r = 6$ the ellipse is fat, and at $r = 15$ it is round. The ensemble average of the product is 0.8555, 0.3074 and −0.01811, against $\xi = 0.8825$, 0.3247 and ≈ 0.001. Each estimate is within one or two standard errors of the truth. (For two unit-variance Gaussians with correlation ρ, one product has mean ρ and mean square $1 + 2\rho^2$, by (7.9) below, so its variance is $1 + \rho^2$; the average of M products then has standard error $\sqrt{(1+\rho^2)/M}$, between $1/\sqrt{M} \approx 0.01826$ and 0.024 here.) Repeating the measurement at a

different point $\boldsymbol{x}_1$ gives 0.8934, 0.3203 and −0.006406: the same function of r, as homogeneity demands.

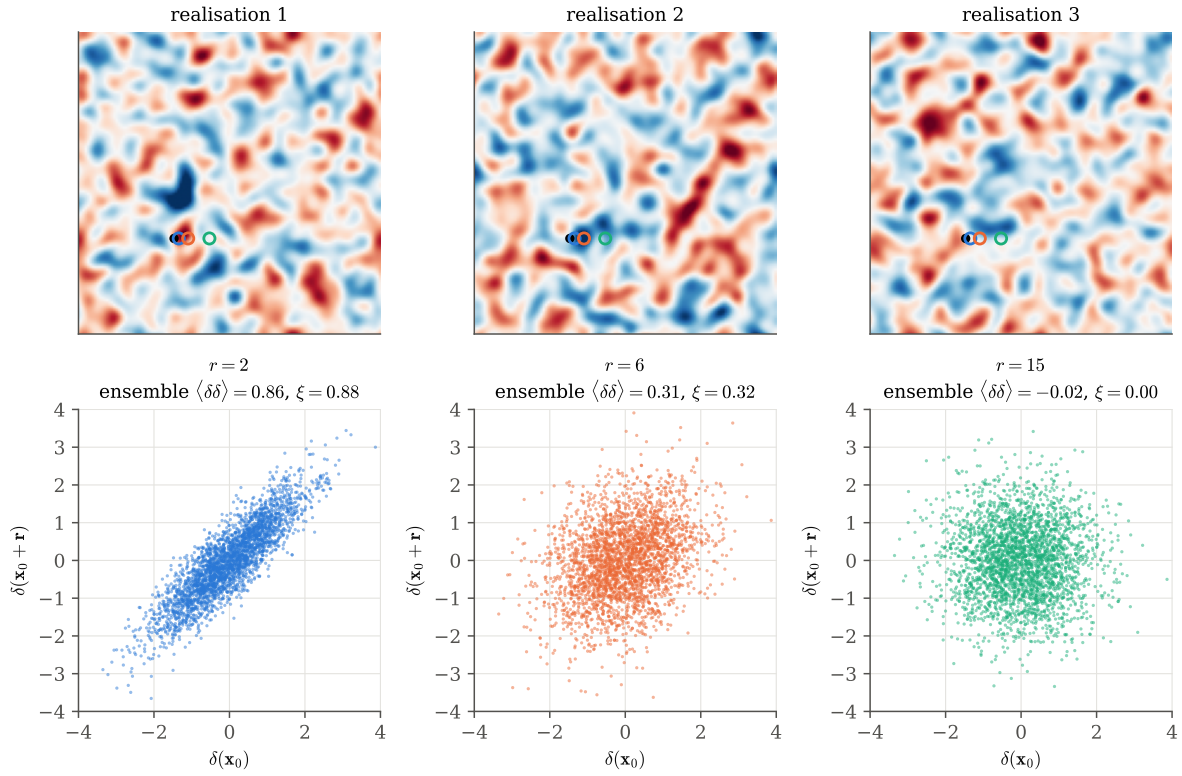


Figure 7.3: Top: three universes drawn from the same Gaussian random field. The black dot is the fixed point $\boldsymbol{x}_0$; the coloured circles are the points 2, 6 and 15 cells to its right. Bottom: the pair $\big(\delta(\boldsymbol{x}_0), \delta(\boldsymbol{x}_0 + \boldsymbol{r})\big)$ collected over 3000 universes. Close points rise and fall together, distant points are unrelated, and the average of the product traces $\xi(r)$.

Watch out: A correlation function is not a property of one map

Each realisation in Figure 7.3 has its own hot and cold spots, and in none of them is the field "homogeneous" in the everyday sense. Homogeneity, isotropy and $\xi(r)$ are properties of the ensemble. Measuring them from one map requires an extra assumption, ergodicity, and carries an irreducible error (Section 7.5).

The correlation function squeezes a whole distribution over maps into one function of one variable. How much of the distribution survives? In general, very little: the n-point distributions can do almost anything that an average over pairs does not see. But there is one large class of fields, and the most important one in physics, for which ξ captures everything.

7.2 Gaussian random fields: when two points tell the whole story

A random field is specified by all of its n-point distributions, for every n and every placement of the points. Measuring this entire hierarchy would require infinitely much information. The near-Gaussianity of the microwave background motivates a much simpler model for its temperature fluctuations [HobsonBMC p. 229]. For a Gaussian field with the symmetries assumed here, the mean and the correlation function $\xi(r)$ fix every n-point distribution. This is what makes the power spectrum such an effective summary for CMB cosmology.

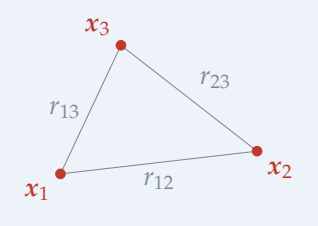
Working definition 7.3 (Gaussian random field)

A random field is **Gaussian** if for every finite set of points $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ the vector $\boldsymbol{\delta} = (\delta(\boldsymbol{x}_1), \dots, \delta(\boldsymbol{x}_n))$ has a multivariate Gaussian density [the density of Chapter 2, fixed by a mean vector and a covariance matrix]. For a homogeneous, isotropic, zero-mean Gaussian field the covariance matrix is read off the correlation function, $\mathbf{C}_{ij} = \xi(|\boldsymbol{x}_i - \boldsymbol{x}_j|)$, so

$$p(\boldsymbol{\delta}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det \mathbf{C}}} \exp\!\left(-\tfrac{1}{2}\,\boldsymbol{\delta}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{\delta}\right). \tag{7.6}$$

Consequence: $\xi(r)$ determines every n-point distribution.

Figure 7.4 shows the construction for three points. Write down the three separations, look up ξ at each, fill the symmetric matrix, and (7.6) is the joint density of the three field values. Nothing else about the field enters. Consistency, the marginalisation condition in the definition of a random field, holds automatically: integrating one variable out of a multivariate Gaussian leaves the Gaussian of the remaining ones with the corresponding sub-block of $\mathbf{C}$ (Chapter 2), which is exactly the matrix we would have built for the smaller set of points.



$$\mathbf{C} = \begin{pmatrix} \xi(0) & \xi(r_{12}) & \xi(r_{13}) \\ \xi(r_{12}) & \xi(0) & \xi(r_{23}) \\ \xi(r_{13}) & \xi(r_{23}) & \xi(0) \end{pmatrix}$$

$$\big(\delta(\boldsymbol{x}_1), \delta(\boldsymbol{x}_2), \delta(\boldsymbol{x}_3)\big) \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$$

Figure 7.4: For a Gaussian field the joint distribution at any set of points is built from the separations alone. Three points give three distances; the correlation function at those distances fills the covariance matrix, and the multivariate Gaussian with that matrix is the joint law of the three values.

Why so many physical fields are Gaussian. Three mechanisms recur, one at a time. *Many small causes.* When the value at a point is the sum of many small independent contributions, as the voltage across a resistor sums the kicks of many electrons, the central limit theorem (Chapter 2) makes it Gaussian. The same argument applied to the values at any finite set of points makes them jointly Gaussian. *Quantum vacuum.* Quantum fluctuations of a free field in its vacuum state are Gaussian, and the simplest models of inflation stretch exactly such fluctuations to cosmological size. *Linear evolution.* A linear map $\mathbf{A}$ applied to a Gaussian vector gives another Gaussian vector, with covariance $\mathbf{A}\mathbf{C}\mathbf{A}^{\mathsf{T}}$ (Chapter 1). The CMB fluctuations have amplitude 10^{-5}, so the equations that evolve them are linear, and an initially Gaussian field stays Gaussian [HobsonBMC p. 230]. Gaussianity is lost when evolution is non-linear. Gravitational collapse turns the nearly Gaussian early matter field into a skewed web of halos and filaments; this is the regime where the linear and halofit spectra of Chapter T1 part company.

7.2.1 What Gaussianity predicts

Because the whole distribution is fixed by ξ, a Gaussian field makes sharp predictions that we can check on a map. We need three of them repeatedly.

The pair distribution. For two points a distance r apart, (7.6) with $n = 2$ is a bivariate Gaussian. After dividing each value by σ (so both have unit variance), its correlation coefficient

is $\rho(r) = \xi(r)/\sigma^2$, with $\sigma^2 = \xi(0)$ the variance at a point. Its contours are ellipses along the diagonal, narrow when ρ is near 1 and circular when $\rho = 0$ (Figure 7.5, left). These are the clouds of Figure 7.3, now drawn from the formula.

The field around a hot spot. *What we observed.* A point where the field is high, $\delta(\boldsymbol{x}) = \nu\sigma$: the field sits ν standard deviations above its mean. *The question.* What should we expect a distance r away? *The idea.* Split the second value into a part that the first value predicts and a remainder Z that it says nothing about,

$$\delta(\boldsymbol{x}+\boldsymbol{r}) = \frac{\xi(r)}{\sigma^2}\,\delta(\boldsymbol{x}) + Z.$$

This equation defines Z, and the coefficient $\xi(r)/\sigma^2$ is chosen so that Z is uncorrelated with $\delta(\boldsymbol{x})$: $\operatorname{Cov}(Z, \delta(\boldsymbol{x})) = \xi(r) - \frac{\xi(r)}{\sigma^2}\sigma^2 = 0$. Z is a linear combination of the Gaussian pair, so Z and $\delta(\boldsymbol{x})$ are jointly Gaussian, and for jointly Gaussian variables uncorrelated means independent. Knowing $\delta(\boldsymbol{x})$ therefore tells us nothing about Z, and

$$\begin{aligned}
\mathbb{E}\big[\delta(\boldsymbol{x}+\boldsymbol{r}) \,\big|\, \delta(\boldsymbol{x}) = \nu\sigma\big] &\overset{(1)}{=} \frac{\xi(r)}{\sigma^2}\,\nu\sigma + \mathbb{E}[Z] \overset{(2)}{=} \nu\sigma\,\frac{\xi(r)}{\sigma^2}, \\
\operatorname{Var}\big[\delta(\boldsymbol{x}+\boldsymbol{r}) \,\big|\, \delta(\boldsymbol{x}) = \nu\sigma\big] &\overset{(3)}{=} \operatorname{Var} Z \overset{(4)}{=} \sigma^2 - \frac{\xi(r)^2}{\sigma^2}.
\end{aligned} \tag{7.7}$$

(1) Conditioning on $\delta(\boldsymbol{x}) = \nu\sigma$ fixes the first term; Z is independent of $\delta(\boldsymbol{x})$, so its distribution is unchanged.
(2) Z is a combination of zero-mean variables.
(3) The first term is fixed by the condition, so only Z fluctuates.
(4) $\operatorname{Var} Z = \operatorname{Var}\delta(\boldsymbol{x}+\boldsymbol{r}) - \frac{\xi^2}{\sigma^4}\operatorname{Var}\delta(\boldsymbol{x}) = \sigma^2 - \xi^2/\sigma^2$, using the independence just shown.

What it means. The mean profile around a point of height $\nu\sigma$ is the correlation function itself, scaled by ν/σ: a hot spot is surrounded by warm field out to the correlation length. This is the seed of the theory of peaks and clusters in [111]. If instead of a single height we select all points above a threshold, $\delta > \nu\sigma$, we average (7.7) over the heights of the selected points (the tower property of expectation). The conditional mean is proportional to the height, so the average simply replaces $\nu\sigma$ by the mean height $\bar{\nu}\sigma$ of the selected points.

The script `code/ch07/02_gaussian_checks.py` tests this on one 512×512 map with $\xi = e^{-r^2/2\ell^2}$, $\ell = 4$. A fraction 0.02405 of the pixels exceed 2σ (the Gaussian tail gives 0.0228), and their mean height is $\bar{\nu} = 2.379$ (the Gaussian value is $\varphi(2)/[1-\Phi(2)] = 2.373$, with φ and Φ the standard normal density and CDF). Four cells to their right the average field is 1.478, against $\bar{\nu}\,\xi(4)/\sigma^2 = 1.443$, and the whole profile follows (7.7) (Figure 7.5, middle). The small excess at large r is noise: the selected pixels belong to a few hundred hot spots, so the average beyond the correlation length scatters by about $1/\sqrt{\text{a few hundred}} \approx 0.05$.

Four-point moments from two-point ones. A Gaussian field carries no new information at higher order: every moment is built from ξ. *The question.* What is the average of a product of four field values? *The tool.* The moment generating function of a zero-mean Gaussian vector (Chapters 1 and 2), $\mathbb{E}\big[e^{\boldsymbol{t}^{\mathsf{T}}\boldsymbol{\delta}}\big] = e^{\frac{1}{2}\boldsymbol{t}^{\mathsf{T}}\mathbf{C}\boldsymbol{t}}$. Expand both sides in powers of $\boldsymbol{t}$: on the left each coefficient is a moment, on the right it is built from $\mathbf{C}$ alone. Matching coefficients gives the answer. For four field values $\delta_1, \dots, \delta_4$ (not necessarily at distinct points) and $C_{ij} = \langle \delta_i \delta_j \rangle$:

$$\begin{aligned}
\langle \delta_1\delta_2\delta_3\delta_4 \rangle &\overset{(1)}{=} \Big[\text{coefficient of } t_1t_2t_3t_4 \text{ in } \mathbb{E}\; e^{\boldsymbol{t}^{\mathsf{T}}\boldsymbol{\delta}}\Big] \overset{(2)}{=} \Big[\text{coefficient of } t_1t_2t_3t_4 \text{ in } \tfrac{1}{2!}\big(\tfrac{1}{2}\textstyle\sum_{ij} C_{ij}t_it_j\big)^2\Big] \\
&\overset{(3)}{=} \tfrac{1}{8}\cdot 8\,\big(C_{12}C_{34} + C_{13}C_{24} + C_{14}C_{23}\big) \overset{(4)}{=} C_{12}C_{34} + C_{13}C_{24} + C_{14}C_{23}.
\end{aligned} \tag{7.8}$$

(1) Expanding $e^{\boldsymbol{t}^\mathsf{T}\boldsymbol{\delta}}$, the term $t_1t_2t_3t_4$ appears only from $\frac{1}{4!}(\boldsymbol{t}^\mathsf{T}\boldsymbol{\delta})^4$, where it occurs 4! times with coefficient $\delta_1\delta_2\delta_3\delta_4$; so its coefficient in the average is the four-point moment.
(2) In $e^{\frac{1}{2}\boldsymbol{t}^\mathsf{T}\mathbf{C}\boldsymbol{t}} = 1 + \frac{1}{2}\boldsymbol{t}^\mathsf{T}\mathbf{C}\boldsymbol{t} + \frac{1}{2!}(\frac{1}{2}\boldsymbol{t}^\mathsf{T}\mathbf{C}\boldsymbol{t})^2 + \dots$ only the quadratic power of the exponent has terms of fourth order in $\boldsymbol{t}$.
(3) $(\sum_{ij} C_{ij}t_it_j)^2 = \sum_{ijkl} C_{ij}C_{kl}t_it_jt_kt_l$. The term $t_1t_2t_3t_4$ arises when (i,j,k,l) is an ordering of $(1,2,3,4)$: 24 orderings, which split the four indices into one of the three pairings, each pairing reached in $2\times2\times2 = 8$ ways (which pair comes first, and the order inside each pair). The prefactor is $\frac{1}{2}\cdot\frac{1}{4} = \frac{1}{8}$.
(4) Simplify.

This is **Isserlis' theorem**, called Wick's theorem in field theory: the four-point function is the sum over the three ways of pairing the points, and every odd moment vanishes (the exponent is even in $\boldsymbol{t}$). Setting $\delta_3 = \delta_1$ and $\delta_4 = \delta_2$ at two points a distance r apart,

$$\langle \delta(\boldsymbol{x})^2\, \delta(\boldsymbol{x}+\boldsymbol{r})^2 \rangle = \sigma^4 + 2\,\xi(r)^2, \tag{7.9}$$

and at $r = 0$ this is the familiar $\langle\delta^4\rangle = 3\sigma^4$ of a single Gaussian. Equation (7.9) gives a test of Gaussianity that needs nothing but the map: measure both sides and compare. For the Gaussian map the ratio of the two sides is 0.9949 at $r = 0$ and 1.009 at $r = 6$. For a lognormal field, $y = e^{0.8\,\delta}$ standardised to zero mean and unit variance, a common model for the non-linear matter density, the same ratio is 9.061 at $r = 0$ and 2.823 at $r = 6$ (Figure 7.5, right). The lognormal field has a perfectly good two-point function, but that function no longer predicts its four-point function: the lognormal field carries information beyond ξ.

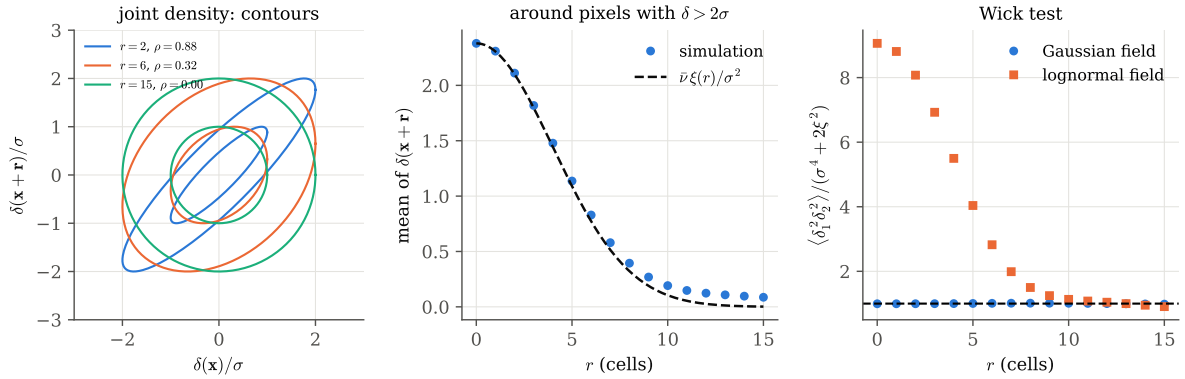


Figure 7.5: Three predictions of Gaussianity. Left: the 1σ and 2σ contours of the joint density of the field at two points, for three separations. Middle: the mean field at distance r from pixels above 2σ (dots) follows the scaled correlation function (dashed). Right: the four-point moment $\langle\delta_1^2\delta_2^2\rangle$ divided by its Gaussian prediction $\sigma^4 + 2\xi^2$ is 1 for a Gaussian map and far from 1 for a lognormal map.

Key idea: For a Gaussian field, $\xi(r)$ is everything

A homogeneous, isotropic Gaussian random field is completely specified by its mean and its correlation function. All n-point distributions follow from (7.6), all higher moments are sums over pairings (7.8), and the conditional field around any feature is a linear function of ξ. A summary that captures ξ (or its Fourier transform) loses nothing. For a non-Gaussian field the same summary is incomplete.

The part of the four-point function that is *not* the sum over pairings, $\langle\delta_1\delta_2\delta_3\delta_4\rangle - (C_{12}C_{34} + C_{13}C_{24} + C_{14}C_{23})$, is called the connected four-point function; its Fourier transform is the trispectrum, and the Fourier transform of the three-point function is the bispectrum. They

are the "higher-order" row of the taxonomy in Section 7.7, and they vanish identically for a Gaussian field.

From here on the working model is a homogeneous, isotropic Gaussian field, whose one free function is $\xi(r)$. In pixel variables its covariance matrix $\mathbf{C}_{ij} = \xi(|\boldsymbol{x}_i - \boldsymbol{x}_j|)$ is full: every pixel is correlated with its neighbours. The natural question is whether some other set of variables makes that matrix diagonal, so that the variables are independent.

7.3 Fourier modes and the power spectrum

A Gaussian map of a million pixels is a Gaussian vector with a million-by-million covariance matrix. To simulate it with the Cholesky factor of Chapter 6 costs about $n^3/3 \approx 3 \times 10^{17}$ operations, and to evaluate its likelihood (7.6) needs the inverse and the determinant of the same matrix. Both are hopeless. Yet the matrix is far from arbitrary: by homogeneity, every entry depends only on the separation of the two pixels. That structure is enough to diagonalise it without computing anything.

The question. In which variables is a homogeneous field a set of *uncorrelated* random numbers, and what are their variances?

The general tool. Any set of correlated random variables x_j with a symmetric covariance matrix V can be rotated into uncorrelated ones, $y_i = \sum_j A_{ij} x_j$: take the rows of A to be the eigenvectors of V. The variances of the y_i are then the eigenvalues of V [Cowan §1.7, (1.61)–(1.65)]. For a general matrix the eigenvectors must be found numerically. *The homogeneous case.* Here we know the eigenvectors in advance: they are plane waves. The reason is the one familiar from physics. A Hamiltonian that is unchanged by translations commutes with the translation operator, so both can be diagonalised together, by momentum eigenstates $e^{i\boldsymbol{k}\cdot\boldsymbol{x}}$. For the same reason the normal modes of a uniform chain of masses and springs are waves. A homogeneous covariance is also unchanged by translations, so its eigenvectors are plane waves too.

7.3.1 Four pixels by hand

Take a ring of $n = 4$ pixels, a periodic line so that pixel 4 is pixel 0 again. Homogeneity says the covariance of pixels i and j is ξ of their distance around the ring, so with $\xi_0 = \xi(0)$, $\xi_1 = \xi(1)$ and $\xi_2 = \xi(2)$,

$$\mathbf{C} = \begin{pmatrix} \xi_0 & \xi_1 & \xi_2 & \xi_1 \\ \xi_1 & \xi_0 & \xi_1 & \xi_2 \\ \xi_2 & \xi_1 & \xi_0 & \xi_1 \\ \xi_1 & \xi_2 & \xi_1 & \xi_0 \end{pmatrix}.$$

Each row is the row above shifted one place to the right, with wrap-around: a **circulant** matrix. Try the plane wave $e^{(m)}_j = e^{2\pi i\, jm/4}$, $j = 0, \dots, 3$, for an integer wavenumber m. Its first component after multiplication by $\mathbf{C}$ is

$$(\mathbf{C}e^{(m)})_0 = \xi_0 + \xi_1 e^{2\pi i m/4} + \xi_2 e^{2\pi i\, 2m/4} + \xi_1 e^{2\pi i\, 3m/4} \overset{(1)}{=} \xi_0 + 2\xi_1 \cos\frac{\pi m}{2} + \xi_2 \cos(\pi m),$$

(1) $e^{2\pi i\, 3m/4} = e^{-2\pi i m/4}$, so the two ξ_1 terms add to $2\xi_1 \cos(\pi m/2)$; and $e^{i\pi m} = \cos \pi m$ is real.

and every other row gives the same number times $e^{(m)}_j$, because row j is row 0 shifted by j and the plane wave picks up exactly the factor $e^{2\pi i jm/4}$ under that shift. So $\mathbf{C}e^{(m)} = \lambda_m e^{(m)}$ with

$$\lambda_0 = \xi_0 + 2\xi_1 + \xi_2, \qquad \lambda_1 = \lambda_3 = \xi_0 - \xi_2, \qquad \lambda_2 = \xi_0 - 2\xi_1 + \xi_2.$$

With $\xi_0 = 1$, $\xi_1 = 0.5$, $\xi_2 = 0.2$ the eigenvalues are 2.2, 0.8, 0.2, 0.8. They are the variances of the four Fourier coefficients of the map: large for the smooth mode $m = 0$, small for the zig-zag mode $m = 2$, because neighbouring pixels are positively correlated and the zig-zag fights that correlation.

The eigenvalues also tell us whether a proposed ξ is possible at all. Try $\xi_0 = 1$, $\xi_1 = 0.9$, $\xi_2 = 0$: neighbours very strongly correlated, next-neighbours not at all. It looks harmless, and $|\xi| \le \xi_0$ holds. But $\lambda_2 = 1 - 1.8 = -0.8 < 0$, a negative variance, so no random field has this correlation function. For example, if pixels 0, 1 are correlated at 0.9 and pixels 1, 2 are too, then pixels 0 and 2 cannot be uncorrelated. The eigenvalues are the positivity test (7.4) made explicit.

Figure 7.6 summarises the construction, and `code/ch07/03_fourier_diagonal.py` repeats it on a ring of $n = 64$ pixels with $\xi(r) = e^{-r^2/2\ell^2}$, $\ell = 3$. Rotating the 64×64 covariance matrix into the Fourier basis leaves it diagonal: the largest off-diagonal entry is 2.402×10^{-14}, rounding error. The diagonal agrees with the discrete Fourier transform of the first row of $\mathbf{C}$, and with the eigenvalues of $\mathbf{C}$ from a general-purpose eigensolver, to 6.217×10^{-15}. As a statistical check the script then draws $M = 20000$ rings with the Cholesky factor, a method that knows nothing about Fourier modes, Fourier transforms each one, and estimates the correlation coefficients between different modes. Over the 25 modes that carry appreciable power their r.m.s. is 0.00727, which is the $1/\sqrt{M} = 0.007071$ expected for truly uncorrelated variables (Figure 7.7).

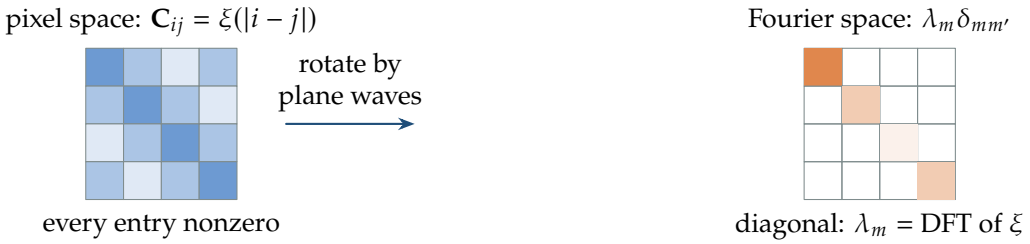


Figure 7.6: A homogeneous covariance on a ring of four pixels. In the pixel basis (left), darker entries indicate larger correlations ξ between pixels. Changing to the plane-wave basis (right) diagonalises this same covariance matrix. Its diagonal entries give the Fourier-mode variances and are the discrete Fourier transform of the correlation function.

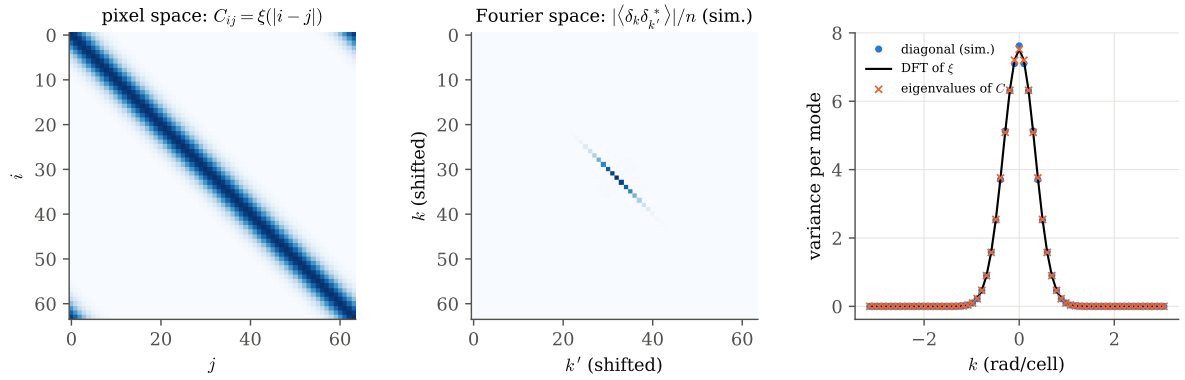


Figure 7.7: The same statement for 64 pixels. Left: the covariance matrix in pixel space, a band of width a few correlation lengths. Middle: the covariance of the Fourier coefficients, estimated from 20 000 simulated rings, is diagonal. Right: the diagonal (dots) equals the discrete Fourier transform of ξ (line), and both equal the eigenvalues of $\mathbf{C}$ (crosses).

7.3.2 The continuum: why Fourier modes are uncorrelated

The same calculation works in any dimension d and in the continuum. We fix a Fourier convention for the whole chapter:

$$\delta_{\boldsymbol{k}} = \int \mathrm{d}^d x\; \delta(\boldsymbol{x})\, e^{-i\boldsymbol{k}\cdot\boldsymbol{x}}, \qquad \delta(\boldsymbol{x}) = \int \frac{\mathrm{d}^d k}{(2\pi)^d}\; \delta_{\boldsymbol{k}}\, e^{i\boldsymbol{k}\cdot\boldsymbol{x}}, \tag{7.10}$$

with the useful identity $\int \mathrm{d}^d x\, e^{-i(\boldsymbol{k}-\boldsymbol{k}')\cdot\boldsymbol{x}} = (2\pi)^d \delta_D(\boldsymbol{k}-\boldsymbol{k}')$, where δ_D is the Dirac delta. Because $\delta(\boldsymbol{x})$ is real, complex conjugating the first relation gives the **reality condition**

$$\delta^*_{\boldsymbol{k}} = \delta_{-\boldsymbol{k}}, \tag{7.11}$$

so the modes $\boldsymbol{k}$ and $-\boldsymbol{k}$ are not independent numbers: one is the complex conjugate of the other [HobsonBMC p. 132]. Now compute the covariance of two Fourier coefficients of a homogeneous field:

$$\begin{aligned}
\langle \delta_{\boldsymbol{k}}\, \delta^*_{\boldsymbol{k}'} \rangle &\overset{(1)}{=} \int \mathrm{d}^d x \int \mathrm{d}^d y\; \langle \delta(\boldsymbol{x})\delta(\boldsymbol{y})\rangle\, e^{-i\boldsymbol{k}\cdot\boldsymbol{x}+i\boldsymbol{k}'\cdot\boldsymbol{y}} \\
&\overset{(2)}{=} \int \mathrm{d}^d x \int \mathrm{d}^d y\; \xi(\boldsymbol{x}-\boldsymbol{y})\, e^{-i\boldsymbol{k}\cdot\boldsymbol{x}+i\boldsymbol{k}'\cdot\boldsymbol{y}} \\
&\overset{(3)}{=} \int \mathrm{d}^d y\; e^{-i(\boldsymbol{k}-\boldsymbol{k}')\cdot\boldsymbol{y}} \int \mathrm{d}^d r\; \xi(\boldsymbol{r})\, e^{-i\boldsymbol{k}\cdot\boldsymbol{r}} \\
&\overset{(4)}{=} (2\pi)^d\, \delta_D(\boldsymbol{k}-\boldsymbol{k}')\, P(\boldsymbol{k}), \qquad P(\boldsymbol{k}) \equiv \int \mathrm{d}^d r\; \xi(\boldsymbol{r})\, e^{-i\boldsymbol{k}\cdot\boldsymbol{r}}.
\end{aligned} \tag{7.12}$$

(1) Insert (7.10) for $\delta_{\boldsymbol{k}}$ and for $\delta^*_{\boldsymbol{k}'}$ (whose exponent changes sign), and take the ensemble average inside the integrals, where only the fields are random.
(2) Homogeneity: the average of the product depends only on $\boldsymbol{x}-\boldsymbol{y}$, (7.2).
(3) Change variables from $(\boldsymbol{x},\boldsymbol{y})$ to $(\boldsymbol{r}=\boldsymbol{x}-\boldsymbol{y},\boldsymbol{y})$, a shift with unit Jacobian. The exponent becomes $-i\boldsymbol{k}\cdot(\boldsymbol{r}+\boldsymbol{y}) + i\boldsymbol{k}'\cdot\boldsymbol{y} = -i\boldsymbol{k}\cdot\boldsymbol{r} - i(\boldsymbol{k}-\boldsymbol{k}')\cdot\boldsymbol{y}$, which factorises.
(4) The $\boldsymbol{y}$ integral is the Dirac-delta identity; the $\boldsymbol{r}$ integral is a function of $\boldsymbol{k}$ alone, which we name $P(\boldsymbol{k})$.

Two Fourier modes with different wavevectors are uncorrelated. The reason is the $\boldsymbol{y}$ integral, the integral over where the pair of points sits: it averages the phase $e^{-i(\boldsymbol{k}-\boldsymbol{k}')\cdot\boldsymbol{y}}$ to zero unless $\boldsymbol{k}=\boldsymbol{k}'$. Homogeneity is used in exactly one place, step (2); without it the $\boldsymbol{y}$ integral would not split off. The variance of each mode is the **power spectrum** $P(\boldsymbol{k})$, the Fourier transform of ξ. Using the reality condition to replace $\delta^*_{\boldsymbol{k}'}$ by $\delta_{-\boldsymbol{k}'}$, the same result reads $\langle \delta_{\boldsymbol{k}}\delta_{\boldsymbol{k}'}\rangle = (2\pi)^d\delta_D(\boldsymbol{k}+\boldsymbol{k}')P(k)$, the form in which cosmology usually defines the matter power spectrum. Inverting the Fourier transform in (7.12) gives ξ back.

Key idea: Wiener–Khinchin: the power spectrum is the Fourier transform of ξ

For a statistically homogeneous field, different Fourier modes are uncorrelated,

$$\langle \delta_{\boldsymbol{k}}\, \delta^*_{\boldsymbol{k}'} \rangle = (2\pi)^d\, \delta_D(\boldsymbol{k}-\boldsymbol{k}')\, P(\boldsymbol{k}),$$
$$P(\boldsymbol{k}) = \int \mathrm{d}^d r\; \xi(\boldsymbol{r})\, e^{-i\boldsymbol{k}\cdot\boldsymbol{r}}, \qquad \xi(\boldsymbol{r}) = \int \frac{\mathrm{d}^d k}{(2\pi)^d}\; P(\boldsymbol{k})\, e^{i\boldsymbol{k}\cdot\boldsymbol{r}}.$$

Isotropy makes P depend only on $k=|\boldsymbol{k}|$. The power spectrum is the variance per Fourier mode, so $P(k)\geq 0$, and a function is a valid correlation function if and only if its Fourier transform is non-negative.

The last sentence is Bochner's theorem; the four-pixel example above is its smallest case. It has two directions. *Valid ξ implies $P \geq 0$:* this is immediate from (7.12), since P is a variance. *$P \geq 0$ implies a valid ξ:* here the proof is a construction, the simulation recipe of Section 7.4, which takes any $P \geq 0$ and builds a field that has it.

A remark on conventions. Some cosmology texts use the symmetric convention, with $(2\pi)^{-d/2}$ on each side of the transform; then $\langle \delta_{\boldsymbol{k}} \delta^*_{\boldsymbol{k}'} \rangle = \delta_D(\boldsymbol{k} - \boldsymbol{k}')P(k)$ with the *same* $P(k)$, because the relation between P and ξ is unchanged (Chapter T1). This is the $P(k)$ that CAMB returns.

How the variance splits among the modes. Setting $\boldsymbol{r} = 0$ in the inverse transform,

$$\sigma^2 = \xi(0) = \int \frac{\mathrm{d}^d k}{(2\pi)^d} P(k), \tag{7.13}$$

so the variance of the field is the sum of the variances of its modes, and $P(k)\,\mathrm{d}^d k/(2\pi)^d$ is the share of the variance carried by the modes in $\mathrm{d}^d k$. In three dimensions with isotropy, $\mathrm{d}^3 k = 4\pi k^2 \mathrm{d}k$, and

$$\sigma^2 = \int_0^\infty \frac{\mathrm{d}k}{k}\, \Delta^2(k), \qquad \Delta^2(k) \equiv \frac{k^3 P(k)}{2\pi^2},$$

the **dimensionless power spectrum**: the variance per logarithmic interval of k (in two dimensions the same reasoning gives $\Delta^2 = k^2 P/2\pi$). A spectrum with constant Δ^2 puts equal variance in every octave of scale. The primordial curvature perturbation of inflation is close to this "scale-invariant" case.

Isotropic transforms in three and two dimensions. For an isotropic field the angular part of the Fourier integral can be done once and for all. In three dimensions, put the k_z axis along $\boldsymbol{r}$, so $\boldsymbol{k} \cdot \boldsymbol{r} = kr\cos\theta$:

$$\begin{aligned} \xi(r) &\overset{(1)}{=} \int_0^\infty \frac{k^2 \mathrm{d}k}{(2\pi)^3} P(k) \int_0^{2\pi} \mathrm{d}\phi \int_{-1}^{1} \mathrm{d}(\cos\theta)\, e^{ikr\cos\theta} \overset{(2)}{=} \int_0^\infty \frac{k^2 \mathrm{d}k}{(2\pi)^3} P(k)\, 2\pi\, \frac{2\sin kr}{kr} \\ &= \int_0^\infty \frac{k^2 \mathrm{d}k}{2\pi^2} P(k) \frac{\sin kr}{kr}, \end{aligned} \tag{7.14}$$

(1) Spherical coordinates in $\boldsymbol{k}$-space, $\mathrm{d}^3 k = k^2 \mathrm{d}k\, \mathrm{d}\phi\, \mathrm{d}(\cos\theta)$, and P depends only on k.

(2) $\int_{-1}^{1} e^{ikru}\mathrm{d}u = \frac{e^{ikr} - e^{-ikr}}{ikr} = \frac{2\sin kr}{kr}$.

the standard relation between ξ and P in cosmology. In two dimensions, with polar angle ϕ between $\boldsymbol{k}$ and $\boldsymbol{r}$,

$$\xi(r) = \int_0^\infty \frac{k\,\mathrm{d}k}{(2\pi)^2} P(k) \int_0^{2\pi} e^{ikr\cos\phi}\mathrm{d}\phi = \int_0^\infty \frac{k\,\mathrm{d}k}{2\pi} P(k)\, J_0(kr), \tag{7.15}$$

using the integral representation of the Bessel function, $J_0(x) = \frac{1}{2\pi}\int_0^{2\pi} e^{ix\cos\phi}\mathrm{d}\phi$. The inverse relations are the same with the roles of P and ξ exchanged and the factors of 2π moved: $P(k) = 4\pi \int_0^\infty r^2 \mathrm{d}r\, \xi(r) \frac{\sin kr}{kr}$ in 3D and $P(k) = 2\pi \int_0^\infty r\,\mathrm{d}r\, \xi(r) J_0(kr)$ in 2D. On the sphere the analogue is the Legendre series $C(\theta) = \sum_\ell \frac{2\ell+1}{4\pi} C_\ell P_\ell(\cos\theta)$ [HobsonBMC (10.4)], with the multipole ℓ playing the role of k.

Example: the Gaussian pair. Take $\xi(r) = \sigma^2 e^{-r^2/2\ell^2}$ in d dimensions. The Fourier integral factorises over Cartesian components:

$$P(k) \overset{(1)}{=} \sigma^2 \prod_{a=1}^{d} \int_{-\infty}^{\infty} \mathrm{d}r_a\, e^{-r_a^2/2\ell^2 - ik_a r_a} \overset{(2)}{=} \sigma^2 \prod_{a=1}^{d} \sqrt{2\pi\ell^2}\, e^{-k_a^2\ell^2/2} \overset{(3)}{=} \sigma^2 (2\pi\ell^2)^{d/2}\, e^{-k^2\ell^2/2}. \tag{7.16}$$

(1) $r^2 = \sum_a r_a^2$ and $\boldsymbol{k}\cdot\boldsymbol{r} = \sum_a k_a r_a$, so the exponential is a product of one-dimensional factors.
(2) Complete the square: $-\frac{r_a^2}{2\ell^2} - ik_a r_a = -\frac{(r_a + ik_a\ell^2)^2}{2\ell^2} - \frac{k_a^2\ell^2}{2}$, and the shifted Gaussian integral is $\sqrt{2\pi\ell^2}$.
(3) Multiply the d factors; $\sum_a k_a^2 = k^2$.

A Gaussian of width ℓ in real space is a Gaussian of width $1/\ell$ in Fourier space: a field correlated over ℓ has its power at $k \lesssim 1/\ell$. The 2D version, $P = 2\pi\ell^2\sigma^2 e^{-k^2\ell^2/2}$, is the spectrum used in Figures 7.3 and 7.5. As a check of (7.13), $\int \frac{\mathrm{d}^2k}{(2\pi)^2} P = \frac{2\pi\ell^2\sigma^2}{(2\pi)^2}\cdot\frac{2\pi}{\ell^2} = \sigma^2$.

Example: the exponential in two dimensions. For $\xi(r) = \sigma^2 e^{-r/\ell}$ the 2D inverse relation needs $\int_0^\infty r\, e^{-ar} J_0(kr)\,\mathrm{d}r$ with $a = 1/\ell$. Start from the Laplace transform of J_0, $\int_0^\infty e^{-ar} J_0(kr)\,\mathrm{d}r = (a^2 + k^2)^{-1/2}$, and differentiate both sides with respect to a:

$$-\int_0^\infty r\, e^{-ar} J_0(kr)\,\mathrm{d}r = -\frac{a}{(a^2+k^2)^{3/2}} \quad\Rightarrow\quad P(k) = 2\pi\sigma^2\frac{a}{(a^2+k^2)^{3/2}} = \frac{2\pi\ell^2\sigma^2}{(1+k^2\ell^2)^{3/2}}.$$

The kink of $e^{-r/\ell}$ at $r = 0$ shows up as a power-law tail $P \propto k^{-3}$ at large k: a correlation function that is not smooth at the origin means a field with structure on all small scales (Figure 7.8, panel b). The script `code/ch07/04_pk_xi_pairs.py` evaluates (7.15) numerically up to $k_{\max} = 10$. For the Gaussian pair it reproduces the analytic ξ to 8.335×10^{-8}. For the exponential the largest error is 0.02499, at $r = 0$, and it has a clean explanation: the truncated tail carries the variance $\int_{k_{\max}}^\infty \frac{k\,\mathrm{d}k}{2\pi}\frac{2\pi\ell^2}{(k\ell)^3} = \frac{1}{\ell k_{\max}} = 0.025$ for $\ell = 4$ cells and $\sigma = 1$.

Example: a preferred wavelength. If $P(k)$ is a narrow bump around k_0, nearly all the variance sits in waves of wavelength $2\pi/k_0$, and (7.15) gives $\xi(r) \approx \sigma^2 J_0(k_0 r)$ times a slowly decaying envelope. The correlation function oscillates and goes *negative*: a hot spot is surrounded by a cold ring at distance $r \approx 3.83/k_0$, the first minimum of J_0. For $k_0 = 0.5$ the minimum sits at 7.664 cells, and the numerical transform puts it at 7.519 with $\xi = -0.3749$. The field looks like a pattern of cells and stripes, the look of convection rolls or of a Turing pattern (Figure 7.8, panel c). The baryon acoustic oscillations are a weak version of the same thing: wiggles in $P(k)$ become a bump in $\xi(r)$ at the sound horizon (Chapter T1).

Example: power laws. For $P(k) = Ak^n$ the substitution $u = kr$ in (7.14) gives $\xi(r) = \frac{A}{2\pi^2} r^{-(n+3)} \int_0^\infty u^{n+2}\frac{\sin u}{u}\mathrm{d}u$, so $\xi \propto r^{-(n+3)}$, and in d dimensions $\xi \propto r^{-(n+d)}$. The u integral converges only for $-3 < n < -1$ in 3D. Near $u = 0$ the integrand is about u^{n+2}, which needs $n > -3$; at large u it oscillates with amplitude u^{n+1}, which needs $n < -1$. For $n \le -3$ the trouble is at small k, the largest scales; for $n \ge -1$ it is at large k, the smallest scales. Either way the spectrum must be cut off, by the size of the survey or by smoothing. Steep spectra (very negative n) give fields dominated by large waves; shallow ones look like noise (Figure 7.10 below).

7.3.3 A Gaussian field in Fourier space

Different modes. For a Gaussian field each $\delta_{\boldsymbol{k}}$ is a linear combination of Gaussian variables, so the Fourier modes are jointly Gaussian too. Jointly Gaussian variables that are uncorrelated are independent, so different modes are independent.

Inside one mode. Each $\delta_{\boldsymbol{k}} = a + ib$ is complex, so we also ask about its real part a and imaginary part b. Two facts fix them. First, the variance of the mode is $\langle|\delta_{\boldsymbol{k}}|^2\rangle = \langle a^2\rangle + \langle b^2\rangle$, proportional

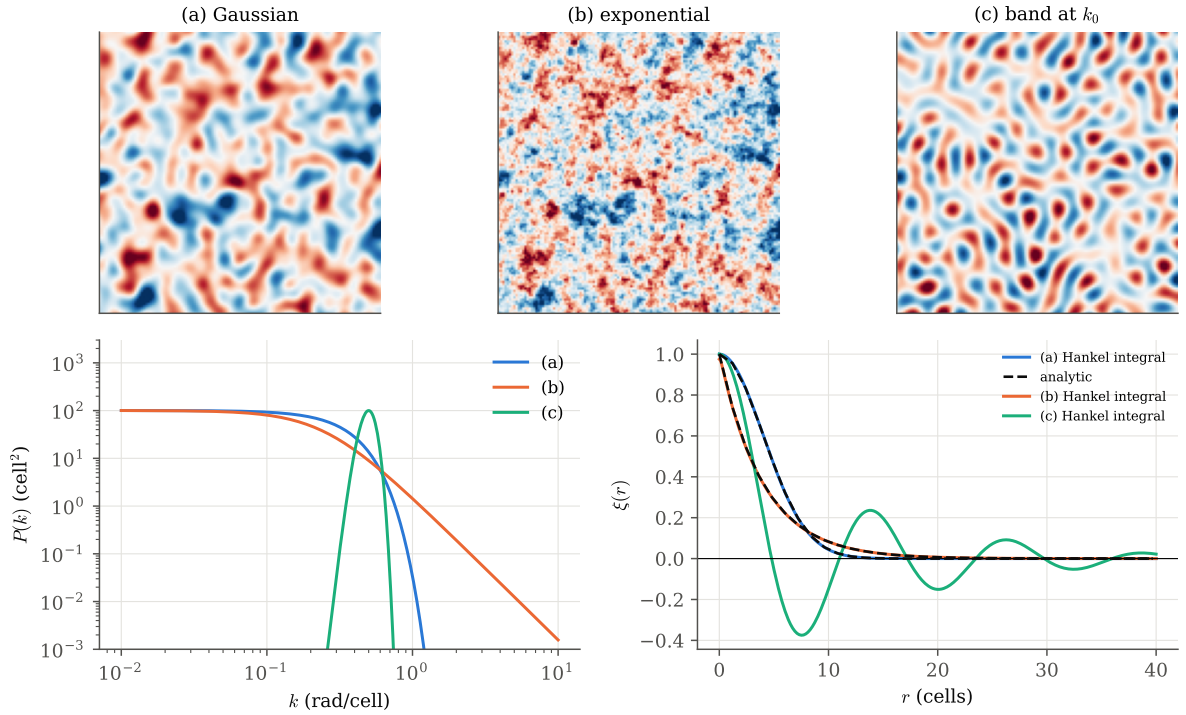


Figure 7.8: Three spectra coloured onto the same white noise. Top: (a) the Gaussian pair, smooth blobs of one size; (b) the exponential, whose k^{-3} tail adds structure on every small scale; (c) a narrow band of wavenumbers, which makes a cellular pattern. Bottom: the spectra and their correlation functions; the analytic ξ (dashed) lies on the numerical Hankel transform, and the band spectrum gives an oscillating ξ with a negative first lobe.

to $P(k)$. Second, for $\boldsymbol{k} \neq \mathbf{0}$ (so that $\boldsymbol{k} \neq -\boldsymbol{k}$) the reality condition gives $\langle \delta_{\boldsymbol{k}} \delta_{\boldsymbol{k}} \rangle = \langle \delta_{\boldsymbol{k}} \delta^*_{-\boldsymbol{k}} \rangle$. That is the covariance of two *different* modes, $\boldsymbol{k}$ and $-\boldsymbol{k}$, so it vanishes by (7.12). Writing it out,

$$0 = \langle (a + ib)^2 \rangle = \langle a^2 \rangle - \langle b^2 \rangle + 2i \langle ab \rangle .$$

The real part says $\langle a^2 \rangle = \langle b^2 \rangle$; the imaginary part says $\langle ab \rangle = 0$. So a and b are independent Gaussians with equal variance, each carrying half of the mode's power.

Amplitude and phase. Then $|\delta_{\boldsymbol{k}}|^2 = a^2 + b^2$ is a sum of two squared Gaussians, a scaled χ^2_2, which is an exponential distribution. In polar form, $\delta_{\boldsymbol{k}} = |\delta_{\boldsymbol{k}}| e^{i\varphi_{\boldsymbol{k}}}$, the amplitude has the Rayleigh distribution and the phase $\varphi_{\boldsymbol{k}}$ is uniform on $[0, 2\pi)$, independent of the amplitude. This is the same polar decomposition of two independent Gaussians that gives the Box–Muller method (Chapter 6).

What it means. A homogeneous Gaussian field is nothing but a sum of plane waves with independent Rayleigh amplitudes, whose mean square is set by $P(k)$, and independent, uniformly random phases. The random amplitudes are why a measured power spectrum scatters about the true one (Section 7.5). The random phases are exactly what $P(k)$ ignores (Section 7.7).

7.4 Simulating a Gaussian random field with the FFT

Every later chapter needs fake universes: hundreds of CMB skies to calibrate an estimator in Chapter 8, mock maps to test a sampler in Chapter 9, Gaussian fields as the null hypothesis for the topological summaries of Chapter 12. *The old way.* The Cholesky factor of Chapter 6 makes a correlated Gaussian vector by colouring white noise: $\boldsymbol{\delta} = \mathbf{L}\boldsymbol{z}$, with $\boldsymbol{z}$ independent standard

normals and $\mathbf{L}\mathbf{L}^{\mathsf{T}} = \mathbf{C}$. But it costs $n^3/3$ operations for n pixels, so it stops at a few thousand pixels. *The way out.* In the Fourier basis the covariance of a homogeneous field is diagonal (Section 7.3), so its "Cholesky factor" is diagonal too: multiply each Fourier mode of white noise by $\sqrt{P(k)}$. The fast Fourier transform does the change of basis in about $n \log n$ operations.

The question. Given a power spectrum $P(k) \geq 0$, how do we produce, on a grid, a real Gaussian field whose Fourier modes have exactly the variance $P(k)$, and with which normalisation?

7.4.1 The box and the grid

A computer holds a field on a finite grid, and the FFT treats the grid as periodic. So we work in a cubic box of side L in d dimensions, of volume $V = L^d$, with periodic boundary conditions. The box is divided into N cells per side, each of size $\Delta = L/N$; there are N^d cells in all, with centres $\boldsymbol{x}_{\boldsymbol{n}} = \boldsymbol{n}\Delta$ for an integer vector $\boldsymbol{n}$.

The box and the grid each limit which waves exist. Periodicity allows only the wavevectors $\boldsymbol{k} = \frac{2\pi}{L}\boldsymbol{m}$ with integer $\boldsymbol{m}$, so there is one allowed $\boldsymbol{k}$ per cell of volume $(2\pi/L)^d$ in $\boldsymbol{k}$-space. The grid allows only $|k_a| \leq \pi/\Delta$ along each axis a. The two ends of that range have names. The **fundamental mode** $k_f = 2\pi/L$ is the longest wave that fits in the box. The **Nyquist wavenumber** $k_{\mathrm{Ny}} = \pi N/L$ is the shortest wave the grid can represent, one that alternates sign from cell to cell. In a finite box the Fourier series replaces the Fourier integral (7.10):

$$\delta_{\boldsymbol{k}} = \int_V \mathrm{d}^d x\; \delta(\boldsymbol{x})\, e^{-i\boldsymbol{k}\cdot\boldsymbol{x}}, \qquad \delta(\boldsymbol{x}) = \frac{1}{V}\sum_{\boldsymbol{k}} \delta_{\boldsymbol{k}}\, e^{i\boldsymbol{k}\cdot\boldsymbol{x}}, \tag{7.17}$$

and the calculation (7.12) goes through with one change: the $\boldsymbol{y}$ integral over the box gives $\int_V \mathrm{d}^d y\, e^{-i(\boldsymbol{k}-\boldsymbol{k}')\cdot\boldsymbol{y}} = V\delta_{\boldsymbol{k}\boldsymbol{k}'}$ (the integral of a whole number of periods of a wave vanishes). Hence

$$\langle \delta_{\boldsymbol{k}}\, \delta^*_{\boldsymbol{k}'} \rangle = V\, P(k)\, \delta_{\boldsymbol{k}\boldsymbol{k}'}. \tag{7.18}$$

The Dirac delta $(2\pi)^d\delta_D(\boldsymbol{k}-\boldsymbol{k}')$ of the infinite volume has become the Kronecker delta times the volume, which is how one should read "$(2\pi)^d\delta_D(\boldsymbol{0}) = V$". Every mode has variance $VP(k)$; a bigger box has more modes, each of larger variance, and the variance per unit volume is the same.

On the grid the integral in (7.17) becomes a sum over cells, $\delta_{\boldsymbol{k}} \approx \Delta^d \sum_{\boldsymbol{n}} \delta_{\boldsymbol{n}} e^{-i\boldsymbol{k}\cdot\boldsymbol{x}_{\boldsymbol{n}}} = \Delta^d F_{\boldsymbol{k}}$, where $F_{\boldsymbol{k}}$ is exactly what `numpy.fft.fftn` returns. The inverse is exact: $\delta_{\boldsymbol{n}} = \frac{1}{V}\sum_{\boldsymbol{k}} \Delta^d F_{\boldsymbol{k}} e^{i\boldsymbol{k}\cdot\boldsymbol{x}_{\boldsymbol{n}}} = \frac{1}{N^d}\sum_{\boldsymbol{k}} F_{\boldsymbol{k}} e^{i\boldsymbol{k}\cdot\boldsymbol{x}_{\boldsymbol{n}}}$, which is `numpy.fft.ifftn`, because $V/\Delta^d = N^d$. Since $F_{\boldsymbol{k}} = \delta_{\boldsymbol{k}}/\Delta^d$, the target (7.18) in numpy's units is

$$\langle |F_{\boldsymbol{k}}|^2 \rangle = \frac{V\, P(k)}{\Delta^{2d}} = \frac{N^d\, P(k)}{\Delta^d}. \tag{7.19}$$

7.4.2 The recipe, and why it is right

The recipe has four steps (Figure 7.9). Draw one independent standard normal $w_{\boldsymbol{n}}$ per cell. Fourier transform it, $W_{\boldsymbol{k}} = \sum_{\boldsymbol{n}} w_{\boldsymbol{n}} e^{-i\boldsymbol{k}\cdot\boldsymbol{x}_{\boldsymbol{n}}}$. Multiply each mode by $\sqrt{P(k)/\Delta^d}$. Transform back. To see that the result has the right covariance, first compute the covariance of the transformed white noise:

$$\langle W_{\boldsymbol{k}} W^*_{\boldsymbol{k}'} \rangle \overset{(1)}{=} \sum_{\boldsymbol{n}}\sum_{\boldsymbol{m}} \langle w_{\boldsymbol{n}} w_{\boldsymbol{m}} \rangle\, e^{-i\boldsymbol{k}\cdot\boldsymbol{x}_{\boldsymbol{n}} + i\boldsymbol{k}'\cdot\boldsymbol{x}_{\boldsymbol{m}}} \overset{(2)}{=} \sum_{\boldsymbol{n}} e^{-i(\boldsymbol{k}-\boldsymbol{k}')\cdot\boldsymbol{x}_{\boldsymbol{n}}} \overset{(3)}{=} N^d\, \delta_{\boldsymbol{k}\boldsymbol{k}'}. \tag{7.20}$$

(1) Definition of $W_{\boldsymbol{k}}$ and of its complex conjugate; average inside the sums.

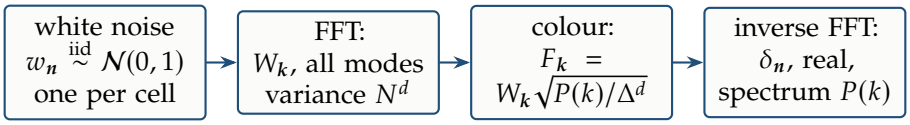


Figure 7.9: The FFT recipe for a Gaussian random field. White noise has the same variance in every Fourier mode; multiplying each mode by the square root of the desired power spectrum, and transforming back, gives a field whose modes have exactly that variance. It is the Cholesky construction of Chapter 6 carried out in the basis where the covariance is diagonal.

(2) The noise is independent with unit variance, $\langle w_n w_m \rangle = \delta_{nm}$, which kills one sum.
(3) For $\boldsymbol{k} = \boldsymbol{k}'$ each of the N^d terms is 1. For $\boldsymbol{k} \neq \boldsymbol{k}'$ the terms are the N-th roots of unity along at least one axis, which sum to zero.

White noise is a field with the same variance, N^d, in every mode and no correlation between modes: a flat power spectrum, which is where the name comes from. The colouring then gives, for $F_k = W_k\sqrt{P(k)/\Delta^d}$,

$$\langle F_k F^*_{k'} \rangle = \sqrt{\tfrac{P(k)}{\Delta^d}}\sqrt{\tfrac{P(k')}{\Delta^d}}\,\langle W_k W^*_{k'} \rangle = \frac{N^d P(k)}{\Delta^d}\,\delta_{kk'},$$

which is the target (7.19). Three more checks complete the argument. *The output is Gaussian,* because it is a linear function of the Gaussian vector $\boldsymbol{w}$. *It is real.* The noise w is real, so $W_{-k} = W^*_k$. Multiplying by $\sqrt{P(|\boldsymbol{k}|)}$, which is real and the same at $\boldsymbol{k}$ and $-\boldsymbol{k}$, keeps $F_{-k} = F^*_k$. That is the reality condition (7.11), so the inverse transform has no imaginary part. *Its variance is right.* By Parseval,

$$\sigma^2 = \langle \delta_n^2 \rangle = \frac{1}{V^2}\sum_{\boldsymbol{k}} \langle |\delta_k|^2 \rangle = \frac{1}{V}\sum_{\boldsymbol{k}} P(k) \;\approx\; \int \frac{\mathrm{d}^d k}{(2\pi)^d} P(k), \tag{7.21}$$

the box version of (7.13): the allowed $\boldsymbol{k}$ have density $V/(2\pi)^d$, so $\frac{1}{V}\sum_{\boldsymbol{k}}$ becomes $\int \mathrm{d}^d k/(2\pi)^d$. Finally, the $\boldsymbol{k} = 0$ mode is the mean of the box. We set it to zero, so every simulated box has mean exactly zero.

Key idea: Colouring white noise in Fourier space

To simulate a homogeneous Gaussian field with power spectrum $P(k)$ on a periodic grid: FFT a map of independent standard normals, multiply mode $\boldsymbol{k}$ by $\sqrt{P(k)/\Delta^d}$, inverse FFT. The cost is $O(N^d \log N^d)$ instead of the $O(N^{3d})$ of a Cholesky factorisation, and the result is exact for the band of wavevectors $k_f \le |k_a| \le k_{\mathrm{Ny}}$ that the box and the grid can hold.

The implementation in the shared library is ten lines.

Listing 7.1: The FFT simulator of a Gaussian random field (`code/ch07/lib_fields.py`)

```
    rng = np.random.default_rng() if rng is None else rng
    D = L / N
    kmag, _ = kgrid(N, L, d)
    with np.errstate(divide="ignore", invalid="ignore"):
        amp = np.sqrt(np.where(kmag > 0, P(np.where(kmag > 0, kmag, 1.0)), 0.0) / D ** d)
    shape = ([nsim] if nsim else []) + [N] * d
    axes = tuple(range(-d, 0))
    w = rng.standard_normal(shape)                      # unit white noise, one number per
    cell
    wk = np.fft.rfftn(w, axes=axes)                     # its FFT: <|wk|^2> = N^d, Hermitian
    return np.fft.irfftn(wk * amp, s=[N] * d, axes=axes)  # colour it, transform back (real)
```

Line 53 is the cell size Δ, line 54 the grid of $|\boldsymbol{k}|$ in the layout of `rfftn`, which stores only the half of $\boldsymbol{k}$-space with $k_d \geq 0$ because the other half is fixed by the reality condition. Line 56 is the colouring factor $\sqrt{P(k)/\Delta^d}$, set to zero at $\boldsymbol{k} = 0$. Lines 59–61 are the recipe: white noise, its FFT, colour, transform back. When `nsim` is given, the first axis counts realisations and the transforms act on the remaining d axes, so a stack of independent fields costs one call. The library also contains the two estimators we use to check the output, and the sphere version of the same recipe.

Estimating $P(k)$ and $\xi(r)$ from a map. Each mode has $\langle|\delta_{\boldsymbol{k}}|^2\rangle = VP(k)$ by (7.18), so $|\delta_{\boldsymbol{k}}|^2/V$ is a one-mode estimate of P. The natural estimate of the power at wavenumber k averages it over the N_{modes} modes in a thin shell $k - \frac{\Delta k}{2} < |\boldsymbol{k}| \leq k + \frac{\Delta k}{2}$,

$$\hat{P}(k) = \frac{1}{N_{\text{modes}}} \sum_{\boldsymbol{k}\in\text{shell}} \frac{|\delta_{\boldsymbol{k}}|^2}{V}, \qquad \langle\hat{P}(k)\rangle = \frac{1}{N_{\text{modes}}} \sum_{\boldsymbol{k}\in\text{shell}} P(|\boldsymbol{k}|), \tag{7.22}$$

so $\hat{P}$ is unbiased for the average of P over the shell. The correlation function could be estimated directly, by averaging $\delta_{\boldsymbol{n}}\delta_{\boldsymbol{n}+\boldsymbol{r}}$ over all cells, but that costs N^d operations for each separation $\boldsymbol{r}$. The Fourier route is faster: it is the discrete form of Wiener–Khinchin. With $\delta_{\boldsymbol{n}}$ real,

$$\begin{aligned}\hat{\xi}(\boldsymbol{r}) \equiv \frac{1}{N^d}\sum_{\boldsymbol{n}} \delta_{\boldsymbol{n}}\delta_{\boldsymbol{n}+\boldsymbol{r}} &\overset{(1)}{=} \frac{1}{N^d}\sum_{\boldsymbol{n}} \frac{1}{N^{2d}} \sum_{\boldsymbol{k},\boldsymbol{k}'} F^*_{\boldsymbol{k}} F_{\boldsymbol{k}'}\, e^{-i\boldsymbol{k}\cdot\boldsymbol{x}_{\boldsymbol{n}} + i\boldsymbol{k}'\cdot(\boldsymbol{x}_{\boldsymbol{n}}+\boldsymbol{r})} \\ &\overset{(2)}{=} \frac{1}{N^{2d}} \sum_{\boldsymbol{k}} |F_{\boldsymbol{k}}|^2 e^{i\boldsymbol{k}\cdot\boldsymbol{r}} \overset{(3)}{=} \frac{1}{N^d}\,\mathrm{IFFT}\big(|F|^2\big)(\boldsymbol{r}),\end{aligned} \tag{7.23}$$

(1) Write $\delta_{\boldsymbol{n}} = \delta^*_{\boldsymbol{n}}$ with the inverse FFT, and $\delta_{\boldsymbol{n}+\boldsymbol{r}}$ with the inverse FFT.
(2) The sum over cells of $e^{i(\boldsymbol{k}'-\boldsymbol{k})\cdot\boldsymbol{x}_{\boldsymbol{n}}}$ is $N^d\delta_{\boldsymbol{k}\boldsymbol{k}'}$, as in (7.20).
(3) The remaining sum is N^d times the inverse FFT of $|F_{\boldsymbol{k}}|^2$.

Its expectation, by (7.19), is $\langle\hat{\xi}(\boldsymbol{r})\rangle = \frac{1}{V}\sum_{\boldsymbol{k}} P(k)e^{i\boldsymbol{k}\cdot\boldsymbol{r}} \equiv \xi_{\text{box}}(\boldsymbol{r})$: the correlation function of the field the box can actually hold, periodic, with no power below k_f or above k_{Ny}. It equals the continuum $\xi(r)$ only for $r \ll L$ and only if P is negligible outside that band. Averaging (7.23) over shells of $|\boldsymbol{r}|$ gives the isotropic $\hat{\xi}(r)$.

7.4.3 Two dimensions: power laws

The script `code/ch07/05_simulate_2d.py` colours one white-noise map of 512×512 unit cells with the power laws $P = Ak^n$, $n = -1, -2, -3$, choosing A so that (7.21) gives $\sigma^2 = 1$ in every case (Figure 7.10). The look of the maps is set by the slope. For $n = -1$ most of the variance is at small scales and the map looks like noise; for $n = -3$, where $\Delta^2 = k^2P/2\pi \propto k^{-1}$ grows towards large scales, a few box-sized waves dominate. In each case the measured $\hat{P}(k)$ follows the input, and averaged over all shells and 20 independent maps the ratio $\hat{P}/\langle\hat{P}\rangle$ is 1.002, 0.9963 and 1.017. The measured $\hat{\xi}(r)$ of the single map scatters about ξ_{box}, and how far it can stray depends on the slope. In this map the variance $\hat{\xi}(0)$ is 1.012, 1.207 and 1.74 for $n = -1, -2, -3$, although each was designed to be 1 on average. The 20 independent maps show how large such strays are: from map to map $\hat{\xi}(0)$ scatters by 0.004822, 0.09247 and 0.3368. Nothing is wrong with the simulation. When the variance lives in the few longest waves, one box contains only a handful of those waves, so it holds only a handful of independent pieces of information about the variance, and the measured variance scatters accordingly. Section 7.5 makes this precise.

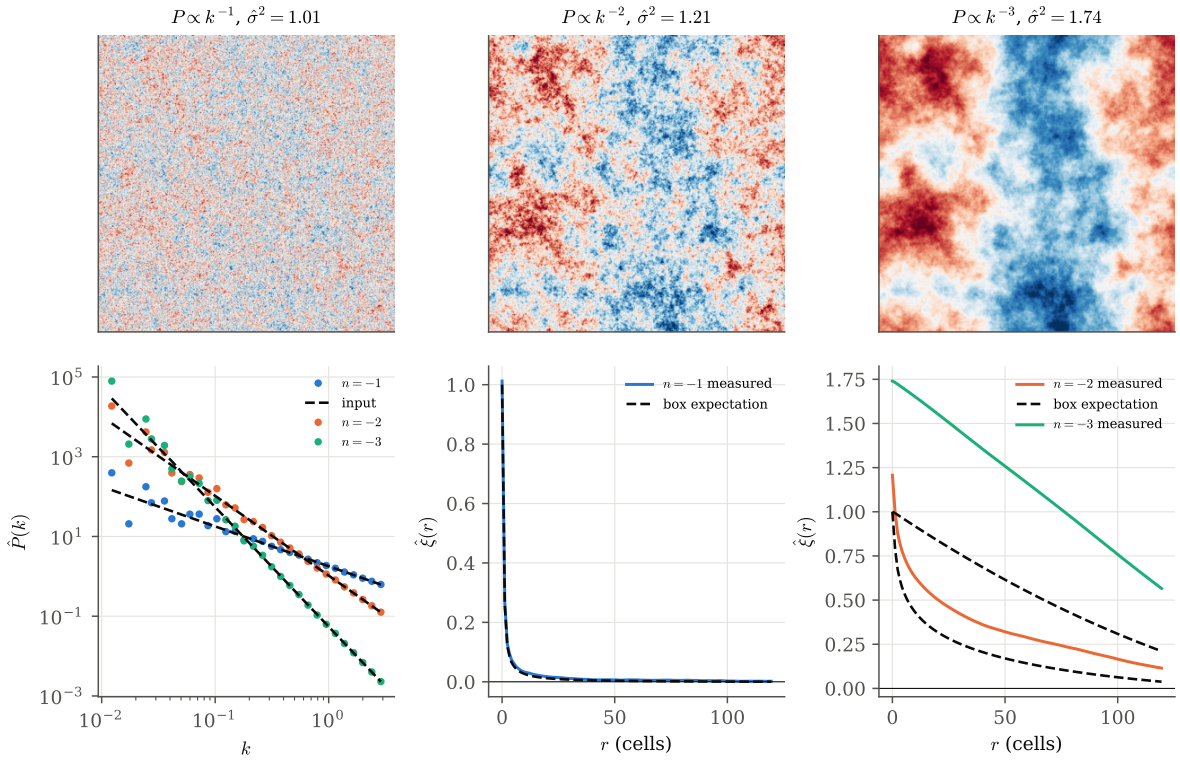


Figure 7.10: Gaussian fields with power-law spectra $P \propto k^n$, coloured from the same white noise. Top: $n = -1, -2, -3$; the steeper the spectrum, the larger the dominant structures. Bottom left: the shell-averaged power of each map (dots) against the input averaged over the same modes (dashed). Bottom middle and right: the measured correlation functions of these maps against their exact box expectations.

7.4.4 Three dimensions: the matter field and the BAO bump

The script `code/ch07/06_simulate_3d_bao.py` takes the linear matter power spectrum of the fiducial Planck 2018 model at $z = 0$ from CAMB (the spectrum of Chapter T1, in $(h^{-1}\text{Mpc})^3$) and fills 4 periodic boxes of side $L = 2000\,h^{-1}\text{Mpc}$ with 256^3 cells of $7.812\,h^{-1}\text{Mpc}$. The box holds the wavenumbers from $k_f = 0.003142$ to $k_{\text{Ny}} = 0.4021\,h\,\text{Mpc}^{-1}$, which covers the turnover of $P(k)$ and the baryon acoustic oscillations. The measured spectra lie on CAMB's, with a mean ratio 0.982 over the shells and visibly larger scatter in the first shells, which contain only 6 modes each (Figure 7.11). In $r^2\hat{\xi}(r)$ every box shows the acoustic bump, which the continuum transform (7.14) puts at $r = 101.3\,h^{-1}\text{Mpc}$; the four boxes scatter about the theory by about $1.301\,(h^{-1}\text{Mpc})^2$ in $r^2\xi$, against a bump height of order ten. A box of $8\,(h^{-1}\text{Gpc})^3$ is larger than any survey yet completed, and a real survey also has a mask, noise and galaxies that are biased tracers of the matter; all of that makes the real measurement harder (Chapter 8).

Watch out: A Gaussian density field can go below -1

The cell-to-cell standard deviation of the simulated matter field is 1.291, so a sizeable fraction of cells have $\delta < -1$, a negative density. The Gaussian model is a statement about the statistics of the linear field on large scales, not a model of the density in $8\,h^{-1}\text{Mpc}$ cells today. Mock catalogues that must be positive use, for example, a lognormal transformation of a Gaussian field, which is non-Gaussian (Figure 7.5) and, for the same $P(k)$, has different higher-order statistics.

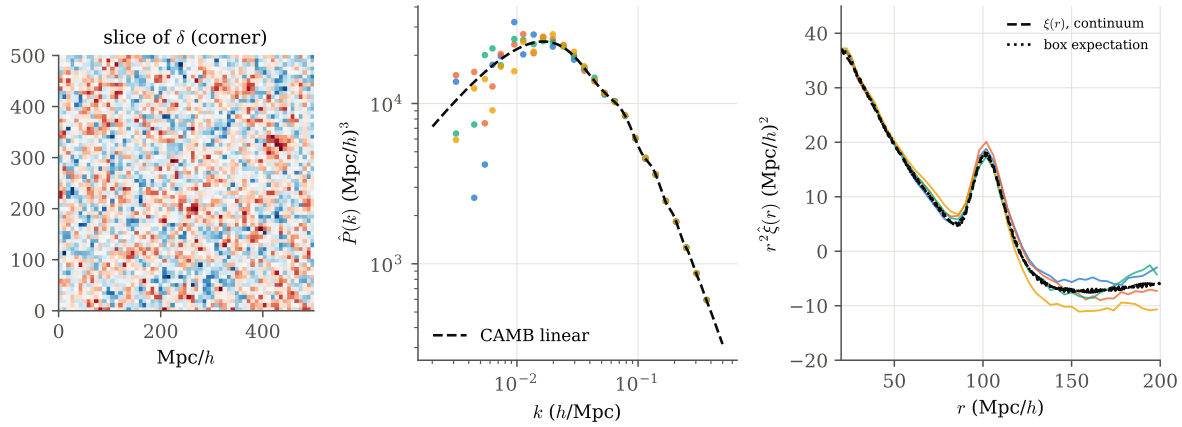


Figure 7.11: A Gaussian matter field with the linear CAMB spectrum. Left: a 500 h^{-1}Mpc corner of one slice. Middle: the shell-averaged power of four independent boxes against the input spectrum; the scatter is largest at small k, where shells contain few modes. Right: $r^2\hat{\xi}(r)$ of the four boxes, the continuum prediction (dashed) and the exact box expectation (dotted); the acoustic bump near 100 h^{-1}Mpc is recovered in every box.

The sphere. On the sky the plane waves are replaced by spherical harmonics and the Fourier coefficients by $a_{\ell m}$, which are independent with variance C_ℓ for a statistically isotropic Gaussian field [HobsonBMC (10.2)–(10.3)]. The library function `gaussian_field_sphere` applies the same recipe: one Gaussian per (ℓ, m), scaled by $\sqrt{C_\ell}$, followed by a spherical-harmonic transform on a HEALPix grid [7], using our own reproducible random stream. The full-sky estimator $\widehat{C}_\ell$ and its cosmic variance, the sphere's version of the scatter just seen in the first k-shells, are derived for the CMB in the second half of Chapter 7.

7.5 One universe: ergodicity and sample variance

Every definition so far is an average down the stack of universes of Figure 7.2: $\xi(r)$ is the ensemble average of $\delta(\boldsymbol{x})\delta(\boldsymbol{x}+\boldsymbol{r})$ at fixed points, $P(k)$ the ensemble variance of a mode. But we have one universe. What we can actually compute is an average *across* our one map: for example the volume average $\hat{\xi}(r) = \frac{1}{V}\int_V \mathrm{d}^d x\, \delta(\boldsymbol{x})\delta(\boldsymbol{x}+\boldsymbol{r})$ over a region of volume V, or the shell average $\hat{P}(k)$ of (7.22). Using the second kind of average in place of the first is the **ergodic hypothesis**. It is an assumption about the field, it can fail, and even when it holds it leaves a finite-volume error that no instrument can remove.

The questions. (1) When does an average over one large volume converge to the ensemble average? (2) For a finite volume, how far from the ensemble value is the spatial average likely to be, for the mean and for the power spectrum?

7.5.1 The variance of a spatial average

Start with the simplest spatial average, the mean of the field over a region of volume V, $\bar{\delta}_V = \frac{1}{V}\int_V \mathrm{d}^d x\, \delta(\boldsymbol{x})$. Its ensemble mean is the ensemble mean of the field, which is zero. The question is how much it scatters from one universe to the next, that is, its variance:

$$\operatorname{Var}\bar{\delta}_V \overset{(1)}{=} \frac{1}{V^2}\int_V \mathrm{d}^d x \int_V \mathrm{d}^d y\, \langle\delta(\boldsymbol{x})\delta(\boldsymbol{y})\rangle \overset{(2)}{=} \frac{1}{V^2}\int_V \mathrm{d}^d x \int_V \mathrm{d}^d y\, \xi(\boldsymbol{x}-\boldsymbol{y})$$

$$\overset{(3)}{\approx} \frac{1}{V^2}\int_V \mathrm{d}^d x \int \mathrm{d}^d r\; \xi(\boldsymbol{r}) \overset{(4)}{=} \frac{1}{V}\int \mathrm{d}^d r\; \xi(\boldsymbol{r}) \overset{(5)}{=} \frac{P(0)}{V}. \tag{7.24}$$

(1) Square the average and take the expectation inside both integrals; it is the continuum version of the variance of a sum, a double sum of covariances [Wasserman p. 53, Thm. 3.20].
(2) Homogeneity, (7.2).
(3) For fixed $\boldsymbol{x}$, substitute $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{x}$. If the region is much larger than the range of ξ, then for almost every $\boldsymbol{x}$ all the separations where ξ is non-negligible keep $\boldsymbol{y}$ inside the region, and the $\boldsymbol{r}$ integral may be extended over all space. Only a shell of thickness about one correlation length at the edge violates this.
(4) The $\boldsymbol{r}$ integral no longer depends on $\boldsymbol{x}$, and $\int_V \mathrm{d}^d x = V$.
(5) By (7.12) at $\boldsymbol{k} = 0$, $P(0) = \int \mathrm{d}^d r\; \xi(\boldsymbol{r})$.

What it means. A correlated field over a volume V is worth as much as a certain number of independent measurements, one per *correlation volume*. To see this, recall that the mean of n independent measurements of variance σ^2 has variance σ^2/n (Chapter 3), and write (7.24) in the same form:

$$\operatorname{Var}\bar{\delta}_V \approx \frac{\sigma^2}{N_{\mathrm{eff}}}, \qquad N_{\mathrm{eff}} = \frac{V}{V_c}, \qquad V_c \equiv \frac{1}{\sigma^2}\int \mathrm{d}^d r\; \xi(r). \tag{7.25}$$

The volume V acts like N_{eff} independent measurements, one per correlation volume V_c, because neighbouring points repeat each other's information. A survey of volume V contains about V/L^3 statistically independent patches of size L, and the scatter between surveys that this causes is called cosmic variance.

When the average converges: ergodicity. Equation (7.24) answers question (1) for the mean. The spatial average converges to the ensemble mean, in the sense that its variance goes to zero, whenever $\frac{1}{V^2}\int_V \int_V \xi \to 0$ as $V \to \infty$. It is enough that $\xi(r) \to 0$ as $r \to \infty$: then $\int_V \xi$ grows more slowly than V, and dividing by V^2 sends the variance to zero. A field for which spatial averages converge to ensemble averages is called **ergodic**.

For estimating ξ itself, the same argument applies with the field replaced by the products $\delta(\boldsymbol{x})\delta(\boldsymbol{x}+\boldsymbol{r})$. It now needs the covariance of two such products, a four-point function, to die away at large distances. For a Gaussian field Wick's theorem (7.8) writes that four-point function as products of ξ's, so a correlation function that decays fast enough makes $\hat{\xi}$ converge as well [164].

Example: a field that is not ergodic. Add to a Gaussian field $g(\boldsymbol{x})$ a constant $A \sim \mathcal{N}(0, \sigma_A^2)$, drawn once per universe and the same at every point: $\delta(\boldsymbol{x}) = A + g(\boldsymbol{x})$. The result is homogeneous, isotropic and Gaussian, with $\xi_\delta(r) = \sigma_A^2 + \xi_g(r)$. But $\xi_\delta \to \sigma_A^2 \neq 0$ at large r, and (7.24) becomes $\operatorname{Var}\bar{\delta}_V \to \sigma_A^2 + P_g(0)/V \to \sigma_A^2$. However large the volume, the spatial mean converges to A, the offset of *our* universe, and never to the ensemble mean 0. This is not a mathematical curiosity. To a survey, fluctuation modes with wavelengths longer than the survey are exactly such an offset: they shift its mean density, and no measurement inside the survey can tell them apart from a different background.

Checking it on simulated maps. The script `code/ch07/07_sample_variance.py` draws 200 maps of 256×256 cells with the Gaussian $\xi = e^{-r^2/2\ell^2}$, $\ell = 4$, cuts each into square patches of side s, and measures the variance of the patch means (Figure 7.13, right). For this field $P(0) = 2\pi\ell^2 = 100.5$ cells^2, so the correlation area is about 100 cells. At $s = 64$ the measured variance is $0.02169 \pm 6.082 \times 10^{-4}$ (the error from the scatter between the 200 independent maps), the exact expression on the first line of (7.24), evaluated with the box modes, gives 0.02064, and

the large-patch limit $P(0)/s^2$ gives 0.02454. The limit overestimates slightly because it ignores the edge shell of step (3), where pairs straddling the boundary are lost. With the extra offset of standard deviation 0.5 the variance at $s = 64$ is 0.2792, and it flattens at $\sigma_A^2 = 0.25$ instead of falling: the non-ergodic field never averages down.

Watch out: Ergodicity is an assumption about the ensemble

No test on one realisation can prove that a field is ergodic, because the non-ergodic part (the offset A above, or modes larger than the survey) looks like a constant in every map. Ergodicity is part of the model, justified by the physics (here, by the decay of ξ predicted by the theory), and its failure on the largest scales is one reason those scales are the hardest to measure.

7.5.2 The sample variance of the power spectrum

The power spectrum is estimated from one map by averaging $|\delta_{\boldsymbol{k}}|^2/V$ over a shell of wavevectors, (7.22). *The question.* How far from $P(k)$ is that average likely to be? For a Gaussian field the answer is exact, because we know the distribution of every mode (Section 7.3.3). *One mode first.* For a wavevector that is not its own negative, $\delta_{\boldsymbol{k}} = a + ib$ with a and b independent Gaussians, each of variance $VP/2$ (half of the mode variance VP of (7.18)), so

$$x_{\boldsymbol{k}} \equiv \frac{|\delta_{\boldsymbol{k}}|^2}{VP(k)} = \frac{a^2+b^2}{VP} \sim \tfrac{1}{2}\chi_2^2 = \text{Exponential}(1), \qquad \mathbb{E}\,x = 1, \quad \operatorname{Var} x = 1. \tag{7.26}$$

A single mode is a terrible estimate of the power: its standard deviation equals its mean. *Then the shell.* Averaging helps, but only over *independent* modes, and here the reality condition matters. The shell contains N_{modes} wavevectors, but they come in pairs $\pm\boldsymbol{k}$ with $|\delta_{-\boldsymbol{k}}|^2 = |\delta_{\boldsymbol{k}}^*|^2 = |\delta_{\boldsymbol{k}}|^2$, so only $N_{\text{modes}}/2$ of the values are independent (Figure 7.12). Then

$$\operatorname{Var}\hat{P} \overset{(1)}{=} \operatorname{Var}\Big(\frac{2}{N_{\text{modes}}}\sum_{\text{pairs}}\frac{|\delta_{\boldsymbol{k}}|^2}{V}\Big) \overset{(2)}{=} \frac{4}{N_{\text{modes}}^2}\cdot\frac{N_{\text{modes}}}{2}\cdot P^2\operatorname{Var} x \overset{(3)}{=} \frac{2P^2}{N_{\text{modes}}}, \tag{7.27}$$

(1) Group the shell sum into $N_{\text{modes}}/2$ pairs of equal terms; each pair contributes $2|\delta_{\boldsymbol{k}}|^2/V$.
(2) The pairs are independent, so the variance of the sum is the sum of $N_{\text{modes}}/2$ equal variances, each $\operatorname{Var}(|\delta_{\boldsymbol{k}}|^2/V) = P^2\operatorname{Var} x$ (we take P constant across the thin shell).
(3) $\operatorname{Var} x = 1$ by (7.26).

Key idea: Sample variance of the power spectrum

For a Gaussian field measured without noise over a volume V, the shell-averaged power has fractional error

$$\frac{\Delta\hat{P}}{P} = \sqrt{\frac{2}{N_{\text{modes}}}}, \qquad N_{\text{modes}} = \begin{cases} \dfrac{V\,k^2\Delta k}{2\pi^2} & \text{(3D)}, \\[2ex] \dfrac{A\,k\,\Delta k}{2\pi} & \text{(2D, area } A\text{)}. \end{cases}$$

The "2" comes from the reality condition: $\pm\boldsymbol{k}$ carry the same information. In fact $N_{\text{modes}}\hat{P}/P \sim \chi^2_{N_{\text{modes}}}$, a sum of $N_{\text{modes}}/2$ independent χ_2^2 variables.

The mode count is density times volume. The allowed wavevectors sit on a lattice with one point per cell of volume $(2\pi/L)^d$ in $\boldsymbol{k}$-space (Section 7.4.1), a density of $V/(2\pi)^d$ points. A thin shell has volume $4\pi k^2\Delta k$ in 3D and $2\pi k\Delta k$ in 2D. Multiplying gives the two cases of N_{modes}

above. On the sphere the same counting gives $2\ell+1$ real degrees of freedom at multipole ℓ, and the full-sky estimator has $\mathrm{Var}\,\widehat{C}_\ell/C_\ell^2 = 2/(2\ell+1)$, the cosmic variance of the CMB derived later in Chapter 7.

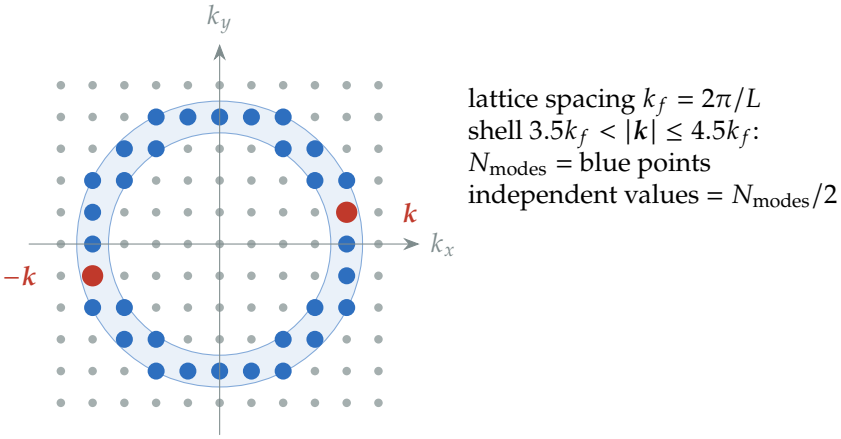


Figure 7.12: Counting modes in two dimensions. The allowed wavevectors form a square lattice of spacing $k_f = 2\pi/L$. The power at $|\boldsymbol{k}| \approx 4k_f$ is estimated from the blue points in the shell. The red pair $\boldsymbol{k}, -\boldsymbol{k}$ carries a single independent value of $|\delta_{\boldsymbol{k}}|^2$, because $\delta_{-\boldsymbol{k}} = \delta_{\boldsymbol{k}}^*$.

Checking it. Figure 7.13 tests both steps on simulated maps. Over 45900 modes of 50 maps, the normalised power $x_{\boldsymbol{k}}$ has mean 1.003 and variance 1.01, and its histogram is the exponential e^{-x} over three decades (left). Over 500 maps of 128×128 cells and 60 shells, the scatter of $\hat{P}$ follows $\sqrt{2/N_{\rm modes}}$ (middle): the first shell, with 4 modes, scatters by 0.7424 against $\sqrt{2/4} = 0.7071$, the tenth, with 20 modes, by 0.3181 against 0.3162, and averaged over all shells the ratio of measured to predicted scatter is 1.01. The scatter of the steep-spectrum maps in Figure 7.10 and of the first k-shells in Figure 7.11 is this effect: there, most of the variance sat in shells with a handful of modes.

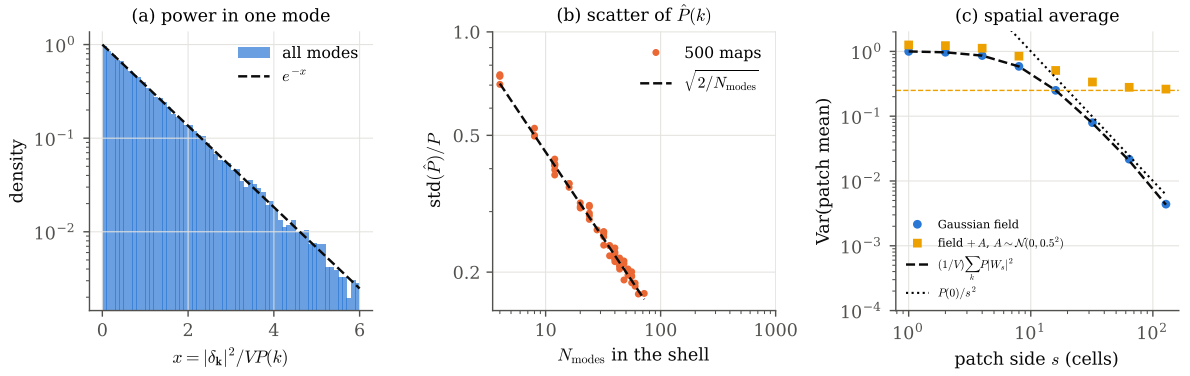


Figure 7.13: One universe, finite volume. Left: the power in a single Fourier mode, in units of its expectation, is exponentially distributed (log scale). Middle: the fractional scatter of the shell-averaged $\hat{P}(k)$ over 500 maps falls as $\sqrt{2/N_{\rm modes}}$. Right: the variance of the mean over an $s\times s$ patch falls as $P(0)/s^2$ for the Gaussian field (blue), following the exact box prediction (dashed); with a random offset per universe (orange) it stalls at the offset's variance.

Example: how well can a galaxy survey measure $P(k)$? Take an idealised survey of volume $V = 1\,(h^{-1}\mathrm{Gpc})^3 = 10^9\,(h^{-1}\mathrm{Mpc})^3$, no noise, and shells of width $\Delta k = 0.01\,h\,\mathrm{Mpc}^{-1}$ at $k =$

$0.1\,h\,\mathrm{Mpc}^{-1}$:

$$N_{\mathrm{modes}} = \frac{V\,k^2 \Delta k}{2\pi^2} = \frac{10^9 \times 10^{-2} \times 10^{-2}}{19.74} = 5066, \qquad \frac{\Delta \hat{P}}{P} = \sqrt{\frac{2}{5066}} = 2.0\%.$$

At $k = 0.01\,h\,\mathrm{Mpc}^{-1}$ with $\Delta k = 0.005\,h\,\mathrm{Mpc}^{-1}$ the same survey has $N_{\mathrm{modes}} = 10^9 \times 10^{-4} \times 0.005/19.74 = 25$ and $\Delta\hat{P}/P = 28\%$. Large scales are always measured worst, because a fixed volume holds few long waves. The only remedy is a bigger volume; a better detector does not help.

Watch out: Sample variance is not noise

The error $\sqrt{2/N_{\mathrm{modes}}}$ is present in a perfect, noise-free measurement of the whole field. It comes from the randomness of the field itself: our universe is one draw. Instrument noise adds to it (white noise of power N multiplies $\sqrt{2/N_{\mathrm{modes}}}$ by $(P+N)/P$), but removing the noise never removes it.

In a periodic box the $\hat{P}$ in different shells are uncorrelated; in a real survey they are not. In the box, different shells contain different modes, and different modes are independent. A survey does not observe a periodic box. It multiplies the field by a window (the survey footprint), and multiplication in real space is convolution in Fourier space. The convolution mixes neighbouring modes and so correlates the shells [HobsonBMC p. 136, (6.5)]. Undoing that mixing on the sphere is the pseudo-C_ℓ problem of Chapter 8.

7.6 Smoothing: beams, pixels and window functions

No instrument records a field at a point. A telescope sees the sky through a beam a few arcminutes wide; a detector pixel integrates over its area; a galaxy survey counts galaxies in cells, or spheres, of finite size. Analysts also smooth on purpose: the amplitude of matter fluctuations is quoted as σ_8, the standard deviation of the density averaged over spheres of radius $8\,h^{-1}$Mpc, and the topological summaries of Chapter 12 are always computed on a field smoothed at a chosen scale. In each case the observed field is a weighted average of the true field over a neighbourhood. The averaging leaves the field homogeneous and Gaussian, and its effect on the power spectrum is a single multiplication.

Working definition 7.4 (window function)

A **window function** $W_R(\boldsymbol{x})$ is a weight centred on the origin with characteristic width R and $\int \mathrm{d}^d x\, W_R(\boldsymbol{x}) = 1$. The field **smoothed** on scale R is the convolution

$$\delta_R(\boldsymbol{x}) = \int \mathrm{d}^d y\; W_R(\boldsymbol{x}-\boldsymbol{y})\,\delta(\boldsymbol{y}) \tag{7.28}$$

[the weighted average of δ around $\boldsymbol{x}$, with weights W_R]. Its Fourier transform $W(\boldsymbol{k})$ is also called the window function; the normalisation means $W(\boldsymbol{0}) = 1$. *Examples:* a Gaussian beam, a top-hat sphere (counts in cells), a square pixel.

The question. If we know the power spectrum $P(k)$ of the true field, what are the power spectrum, the correlation function and the variance of the smoothed field?

7.6.1 The convolution theorem does all the work

Fourier transform (7.28) with the convention (7.10):

$$\delta_{R,\boldsymbol{k}} \overset{(1)}{=} \int \mathrm{d}^d x\, e^{-i\boldsymbol{k}\cdot\boldsymbol{x}} \int \mathrm{d}^d y\, W_R(\boldsymbol{x}-\boldsymbol{y})\,\delta(\boldsymbol{y}) \overset{(2)}{=} \int \mathrm{d}^d y\, \delta(\boldsymbol{y})\, e^{-i\boldsymbol{k}\cdot\boldsymbol{y}} \int \mathrm{d}^d u\, W_R(\boldsymbol{u})\, e^{-i\boldsymbol{k}\cdot\boldsymbol{u}}$$
$$\overset{(3)}{=} W(\boldsymbol{k})\,\delta_{\boldsymbol{k}}. \tag{7.29}$$

(1) Definitions (7.10) and (7.28).
(2) Swap the order of integration and, for fixed $\boldsymbol{y}$, substitute $\boldsymbol{u} = \boldsymbol{x}-\boldsymbol{y}$; then $e^{-i\boldsymbol{k}\cdot\boldsymbol{x}} = e^{-i\boldsymbol{k}\cdot\boldsymbol{y}} e^{-i\boldsymbol{k}\cdot\boldsymbol{u}}$ and the two integrals separate.
(3) Each integral is a Fourier transform.

Smoothing acts mode by mode: each Fourier mode is multiplied by a number, $W(\boldsymbol{k})$. Multiplying each mode by its own number cannot make two different modes correlated, so the covariance of the smoothed modes follows at once from (7.12):

$$\langle \delta_{R,\boldsymbol{k}}\, \delta^*_{R,\boldsymbol{k}'} \rangle = W(\boldsymbol{k}) W^*(\boldsymbol{k}') \langle \delta_{\boldsymbol{k}} \delta^*_{\boldsymbol{k}'} \rangle = (2\pi)^d \delta_D(\boldsymbol{k}-\boldsymbol{k}')\, |W(\boldsymbol{k})|^2 P(k). \tag{7.30}$$

Key idea: Smoothing multiplies the power spectrum by $|W|^2$

A smoothed homogeneous field is homogeneous, with

$$P_R(k) = |W(k)|^2 P(k), \qquad \sigma_R^2 = \langle \delta_R^2 \rangle = \int \frac{\mathrm{d}^d k}{(2\pi)^d}\, |W(k)|^2 P(k).$$

If the field is Gaussian, so is the smoothed field, since (7.28) is linear. Smoothing removes power at $k \gtrsim 1/R$ and leaves larger scales alone.

7.6.2 The common windows

Gaussian. $W_R(\boldsymbol{x}) = (2\pi R^2)^{-d/2} e^{-x^2/2R^2}$. Its transform is (7.16) divided by the normalising constant $(2\pi R^2)^{d/2}$,

$$W(k) = e^{-k^2R^2/2}. \tag{7.31}$$

It is the standard model of a telescope beam; a beam of full width at half maximum θ_{FWHM} has $R = \theta_{\mathrm{FWHM}}/\sqrt{8\ln 2}$.

Top-hat sphere. In three dimensions, $W_R(\boldsymbol{x}) = \frac{3}{4\pi R^3}$ for $|\boldsymbol{x}| < R$ and zero outside: the average over a ball, which is what counting galaxies in spheres does. It is isotropic, so the 3D radial transform (the inverse of (7.14)) applies:

$$W(k) \overset{(1)}{=} \frac{3}{4\pi R^3} \int_0^R 4\pi r^2 \frac{\sin kr}{kr}\,\mathrm{d}r \overset{(2)}{=} \frac{3}{R^3 k}\left[\frac{\sin kr}{k^2} - \frac{r\cos kr}{k}\right]_0^R \overset{(3)}{=} \frac{3\,(\sin x - x\cos x)}{x^3}, \quad x = kR. \tag{7.32}$$

(1) $W(k) = 4\pi \int r^2 \mathrm{d}r\, W_R(r) \frac{\sin kr}{kr}$ with the constant weight inside the ball.
(2) Cancel 4π and one power of r, and integrate $r \sin kr$ by parts: $\int r \sin kr\,\mathrm{d}r = \frac{\sin kr}{k^2} - \frac{r\cos kr}{k}$.
(3) Evaluate at $r = R$ (the lower limit gives 0) and write $x = kR$.

For small x, $\sin x - x\cos x = \frac{x^3}{3} - \frac{x^5}{30} + \ldots$, so $W \to 1$ as it must. This is the W in the σ_8 integral, $\sigma_8^2 = \int_0^\infty \frac{k^2 \mathrm{d}k}{2\pi^2} P(k)\, W^2(k \cdot 8\,h^{-1}\mathrm{Mpc})$, evaluated with the CAMB spectrum in Chapter T1.

Disc. In two dimensions the average over a disc of radius R has, by the 2D radial transform and the identity $\frac{\mathrm{d}}{\mathrm{d}x}[xJ_1(x)] = xJ_0(x)$,

$$W(k) = \frac{1}{\pi R^2}\int_0^R 2\pi r\, J_0(kr)\,\mathrm{d}r = \frac{2}{R^2}\cdot\frac{R\,J_1(kR)}{k} = \frac{2J_1(kR)}{kR}. \tag{7.33}$$

Square patch. The patch means of Section 7.5.1 (the field averaged over square patches of side s) are a smoothing too. The average over a square has $W(\boldsymbol{k}) = \prod_a \frac{\sin(k_a s/2)}{k_a s/2}$, the product of the transforms of one-dimensional boxes, one per axis a. So the variance of the patch mean is σ_R^2 with this window. On the grid of the simulation it is the box sum $\frac{1}{V}\sum_{\boldsymbol{k}} P(k)|W(\boldsymbol{k})|^2$, which gave the dashed curve of Figure 7.13. The same formula also explains the large-patch limit $P(0)/V$ of (7.24). For a large patch, $|W|^2$ is a narrow spike around $\boldsymbol{k}=0$, of area $(2\pi)^d/V$ with $V = s^d$ the patch volume. Over so narrow a spike P is nearly constant, equal to $P(0)$, so the integral is $P(0)\cdot\frac{(2\pi)^d}{V}\cdot\frac{1}{(2\pi)^d} = P(0)/V$.

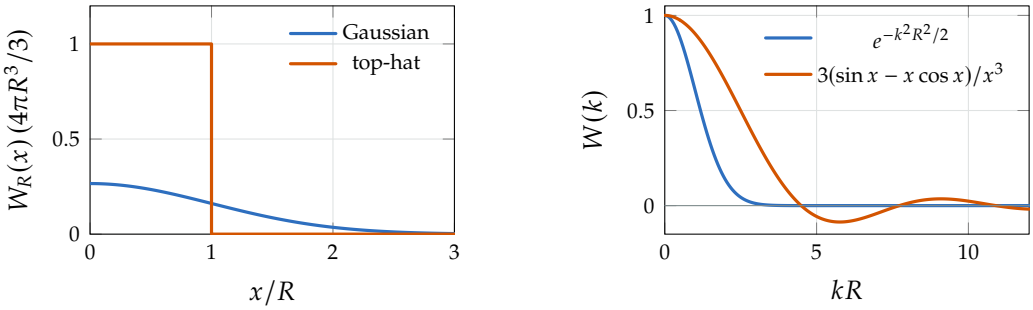


Figure 7.14: The two standard three-dimensional windows. Left: in real space, the Gaussian (scaled to the top-hat's height convention) and the top-hat ball of radius R. Right: their Fourier transforms. Both are 1 at $k = 0$ and cut off near $kR \approx 2$; the sharp edge of the top-hat leaves oscillating side lobes that let through some power at $kR \gg 1$, which the Gaussian suppresses completely.

Example: smoothing white noise. Pixel noise of variance σ_n^2 in cells of area Δ^2 has $\xi = \sigma_n^2\delta_{\boldsymbol{r},\boldsymbol{0}}$, so $P(k) = \sum_{\boldsymbol{r}}\Delta^2\xi(\boldsymbol{r})e^{-i\boldsymbol{k}\cdot\boldsymbol{r}} = \sigma_n^2\Delta^2$, flat. Smooth it with a 2D Gaussian of width $R \gg \Delta$:

$$\sigma_R^2 \overset{(1)}{=} \int\frac{\mathrm{d}^2k}{(2\pi)^2}\,\sigma_n^2\Delta^2\,e^{-k^2R^2} \overset{(2)}{=} \frac{\sigma_n^2\Delta^2}{2\pi}\int_0^\infty k\,e^{-k^2R^2}\,\mathrm{d}k \overset{(3)}{=} \frac{\sigma_n^2\Delta^2}{4\pi R^2}. \tag{7.34}$$

(1) The variance formula for σ_R^2 with $|W|^2 = e^{-k^2R^2}$ from (7.31).
(2) Polar coordinates, $\mathrm{d}^2k = 2\pi k\,\mathrm{d}k$.
(3) $\int_0^\infty k\,e^{-k^2R^2}\mathrm{d}k = \frac{1}{2R^2}$.

The noise variance falls as if we had averaged $4\pi R^2/\Delta^2$ independent pixels: a Gaussian beam of width R has an effective area $4\pi R^2$. For $R = 2$ pixels the noise r.m.s. drops by a factor $\sqrt{4\pi\cdot 2^2} = 2\sqrt{\pi}\times 2 \approx 7.1$. This is the trade-off of every smoothed map: less noise, less resolution.

7.6.3 Gradients and spectral moments

The variance is the first of a family. Because ∇ acting on $e^{i\boldsymbol{k}\cdot\boldsymbol{x}}$ brings down $i\boldsymbol{k}$, the gradient of the smoothed field has Fourier modes $i\boldsymbol{k}\,\delta_{R,\boldsymbol{k}}$, and by (7.30)

$$\langle|\nabla\delta_R|^2\rangle = \int\frac{\mathrm{d}^dk}{(2\pi)^d}\,k^2|W(k)|^2P(k) \equiv \sigma_1^2, \qquad \sigma_j^2 \equiv \int\frac{\mathrm{d}^dk}{(2\pi)^d}\,k^{2j}|W(k)|^2P(k), \tag{7.35}$$

the **spectral moments** of [111], with $\sigma_0 = \sigma_R$.

What σ_1/σ_0 measures. The ratio is an inverse length. Its inverse, σ_0/σ_1, is the typical distance over which the field changes by its own r.m.s., and so the typical spacing of its hot and cold spots. *Why it matters.* The expected number of peaks of a Gaussian field, and the expected Euler characteristic of its excursion sets, depend on the power spectrum only through such moments [110, 111]. Those counts are the topological summaries of Chapter 12. For a Gaussian field they carry no information beyond $P(k)$; their value lies in what they reveal when the field is *not* Gaussian.

Without smoothing, the moments of a power law can diverge. With $P = A/k$ in 2D, $\sigma_1^2 \propto \int k^2 \cdot \frac{A}{k} \cdot k\,\mathrm{d}k$, which grows like k_{Ny}^3 as the upper limit k_{Ny} grows. So the unsmoothed field is rough on the scale of the pixel, and its count of hot spots is set by the grid, not by the physics. A Gaussian window cures this:

$$\sigma_0^2 = \frac{A}{2\pi}\int_0^\infty e^{-k^2R^2}\mathrm{d}k = \frac{A}{2\pi}\frac{\sqrt{\pi}}{2R}, \qquad \sigma_1^2 = \frac{A}{2\pi}\int_0^\infty k^2 e^{-k^2R^2}\mathrm{d}k = \frac{A}{2\pi}\frac{\sqrt{\pi}}{4R^3}, \qquad \frac{\sigma_1}{\sigma_0} = \frac{1}{\sqrt{2}\,R},$$

using $\int_0^\infty e^{-k^2R^2}\mathrm{d}k = \frac{\sqrt{\pi}}{2R}$ and $\int_0^\infty k^2 e^{-k^2R^2}\mathrm{d}k = \frac{\sqrt{\pi}}{4R^3}$. Every count of features in a smoothed map must therefore say on which scale R it was smoothed.

Checking it. The script `code/ch07/08_smoothing.py` smooths one 512×512 map with $P = A/k$ and unit raw variance by Gaussian and disc windows (Figure 7.15). The smoothed power spectra lie on $|W|^2P$, including the side lobes of the disc's J_1. At $R = 4$ cells the Gaussian-smoothed map has $\sigma_R = 0.2505$ against the predicted 0.2466, the disc-smoothed one 0.3464 against 0.3439; the disc keeps more variance because it averages over a smaller effective area (πR^2 rather than $4\pi R^2$). Both variances fall as $1/R$, as the integral for σ_0^2 above predicts. The gradient ratio σ_1/σ_0 at $R = 4$, measured by spectral differentiation of the map, is 0.1766, against $1/(\sqrt{2}\cdot 4) = 0.177$ in the continuum and 0.1799 from the box modes.

7.6.4 What a beam does to the acoustic peaks

The temperature spectrum of the microwave sky (Chapter T1) has a flat plateau at the largest angles and then a train of acoustic peaks, the first at $\ell \approx 220$, the next at 536, 813, 1126, 1421 and 1726 in the fiducial model. Here ℓ is the multipole, the sphere's version of a wavenumber: a pattern of multipole ℓ has an angular wavelength of about $2\pi/\ell$ radians, so $\ell = 200$ means spots about a degree across. Its power is C_ℓ, plotted as $D_\ell = \ell(\ell+1)C_\ell/2\pi$ (Chapter T1). Every satellite that mapped this sky saw it through a beam, and the beams differed by a factor of eighty: about 7° for COBE, between 0.9° and 0.2° for the channels of WMAP, and about 5′ for the best Planck channels. *The question.* Which of the peaks can an experiment with a given beam see?

The beam window. On a patch of sky small enough to be flat, ℓ is the length of a two-dimensional wave vector, and C_ℓ plays the role of $P(k)$. A Gaussian beam is the Gaussian window (7.31) with $R = \sigma_b = \mathrm{FWHM}/\sqrt{8\ln 2}$, measured in radians. So each multipole of the sky is multiplied by $B_\ell = e^{-\ell^2\sigma_b^2/2}$, and, since smoothing multiplies the power by $|W|^2$ (Section 7.6.1), the observed spectrum is $B_\ell^2 C_\ell$. On the full sphere the exact form replaces ℓ^2 by $\ell(\ell+1)$, $B_\ell^2 = e^{-\ell(\ell+1)\sigma_b^2}$, which Chapter 8 derives from the Legendre expansion of the beam.

Where does the beam cut? It leaves the large scales alone and removes power above some

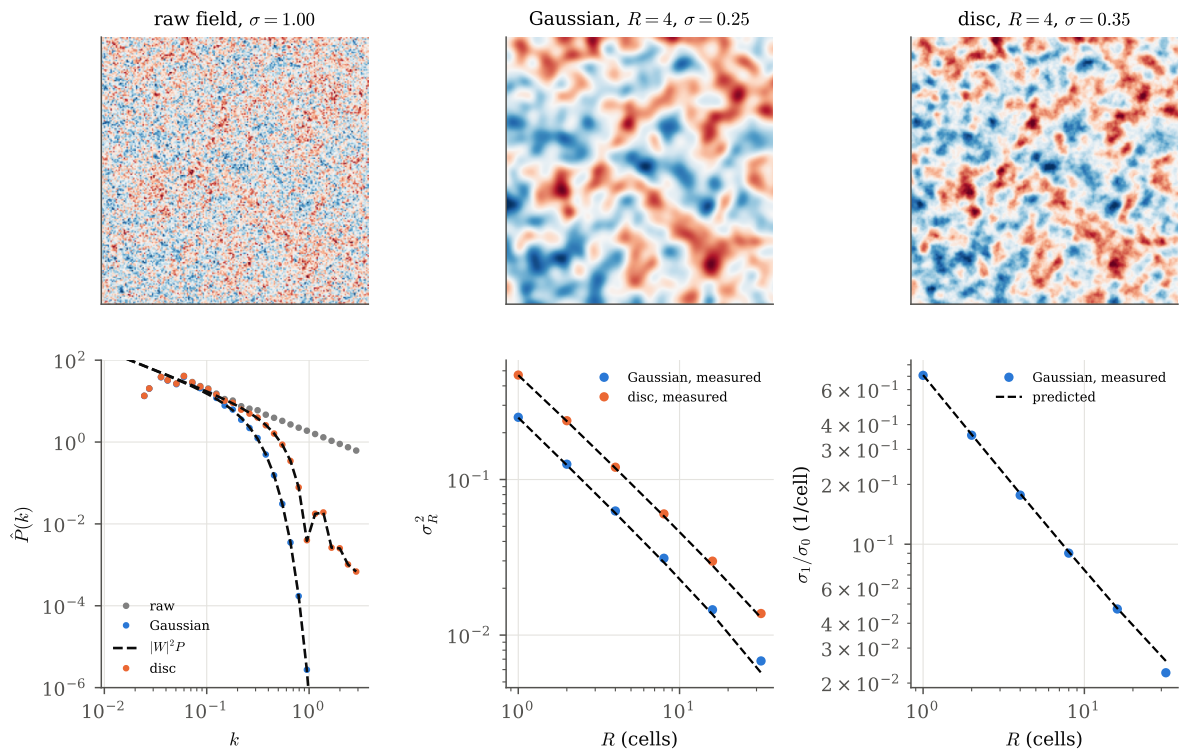


Figure 7.15: Smoothing a field with $P \propto k^{-1}$. Top: a corner of the raw map, and the same map smoothed with a Gaussian and with a disc of radius 4 cells, each shown in units of its own standard deviation. Bottom left: the smoothed spectra follow $|W|^2 P$. Bottom middle: σ_R^2 falls as $1/R$ for both windows. Bottom right: the ratio σ_1/σ_0, the inverse feature spacing, falls as $1/(\sqrt{2}R)$.

multipole. A natural marker is the multipole where it has removed half the power:

$$e^{-\ell(\ell+1)\sigma_b^2} = \tfrac{1}{2} \overset{(1)}{=} \ell(\ell+1)\,\sigma_b^2 = \ln 2 \overset{(2)}{\Longrightarrow} \ell_{1/2} \approx \frac{\sqrt{\ln 2}}{\sigma_b}$$
$$\overset{(3)}{=} \frac{\sqrt{8\,\ln 2}}{\mathrm{FWHM}} \approx \frac{1.96}{\mathrm{FWHM\,[rad]}} \overset{(4)}{=} \frac{6740}{\mathrm{FWHM\,[arcmin]}}. \tag{7.36}$$

(1) Take minus the logarithm of both sides.

(2) For $\ell \gg 1$, $\ell(\ell+1) \approx \ell^2$. The exact positive root, $\ell = \left(-1 + \sqrt{1 + 4\ln 2/\sigma_b^2}\right)/2$, is smaller by less than $1/2$.

(3) Insert $\sigma_b = \mathrm{FWHM}/\sqrt{8\ln 2}$: $\sqrt{\ln 2} \cdot \sqrt{8\ln 2} = \sqrt{8}\,\ln 2 = 1.96$.

(4) One arcminute is 2.909×10^{-4} rad.

The amplitude B_ℓ (rather than the power B_ℓ^2) falls to one half at $\sqrt{2}$ times this multipole, $\sqrt{2\ln 2}/\sigma_b$. A beam twice as narrow reaches twice as far in ℓ. For the five beams of Figure 7.16, script `code/ch07/10_beam_peaks.py` gives:

FWHM	7°	1°	30′	10′	5′
σ_b [arcmin]	178	25.5	12.7	4.25	2.12
$\ell_{1/2}$	16	112	224	673	1347
peaks with $B_\ell^2 > 1/2$	0	0	1	2	4
deconvolved noise = sky at ℓ	62	363	654	1479	2014

The exact roots and the approximation (7.36) agree to one unit in every case (1348 against 1347 for the 5′ beam). The left panel of Figure 7.16 shows the beamed spectra $D_\ell B_\ell^2$. The suppression

is exponential in ℓ^2, so a beam always takes the highest peaks first: the 10′ beam keeps the first peak at 93% of its power and the fifth at only 5%.

Undoing the beam, and its price. The beam is known, so it can be divided out: $\widehat{C}_\ell/B_\ell^2$ is again an estimate of the true C_ℓ. The price is noise. The detector noise is added *after* the beam, so dividing by B_ℓ^2 also multiplies the noise power by $1/B_\ell^2 = e^{\ell(\ell+1)\sigma_b^2}$. White noise has a flat spectrum, as in the example of smoothing white noise above: pixel noise of variance σ_n^2 in cells of solid angle Ω has $N_\ell = \sigma_n^2\Omega$ for every ℓ, and $\sqrt{N_\ell}$ is the map depth, quoted in μK arcmin. So the deconvolved spectrum is $C_\ell + N_\ell/B_\ell^2$ on average, and the second term climbs like $e^{\ell^2\sigma_b^2}$. The right panel of Figure 7.16 shows it for the same depth, 30 μK arcmin, in every case: each curve follows the sky until the noise wall rises through it, at the multipole in the last row of the table. Beyond that point no amount of deconvolution recovers the peaks; Chapter 8 turns this into an error bar.

What each experiment could see. A 7° beam, COBE's, keeps half the power only up to $\ell \approx 16$. That is the Sachs–Wolfe plateau, the flat part of D_ℓ at $\ell \lesssim 20$ that carries the amplitude and the tilt of the primordial fluctuations, and nothing of the peaks: at the first peak, $\ell = 220$, the beam leaves $B_\ell^2 = e^{-\ell(\ell+1)\sigma_b^2} \approx 10^{-57}$ of the power. WMAP's sharpest channel had a beam of about 13′, so $\ell_{1/2} \approx 518$: the first peak keeps 88% of its power, the second 48% and the third only 18%. With its noise wall between those of the 10′ and 30′ columns, such a beam measures the first two peaks well and the third at the edge of its reach. A 5′ beam, Planck's, keeps half the power up to $\ell \approx 1347$, so the first four peaks pass almost untouched, and the deconvolved noise of a 30 μK arcmin map overtakes the sky only at $\ell \approx 2014$: all six peaks of the figure, and the start of the damping tail beyond them, are within reach. The beam width and the reach in multipoles are two ways of stating one number: $\ell_{1/2} \approx 6740/\mathrm{FWHM}[']$.

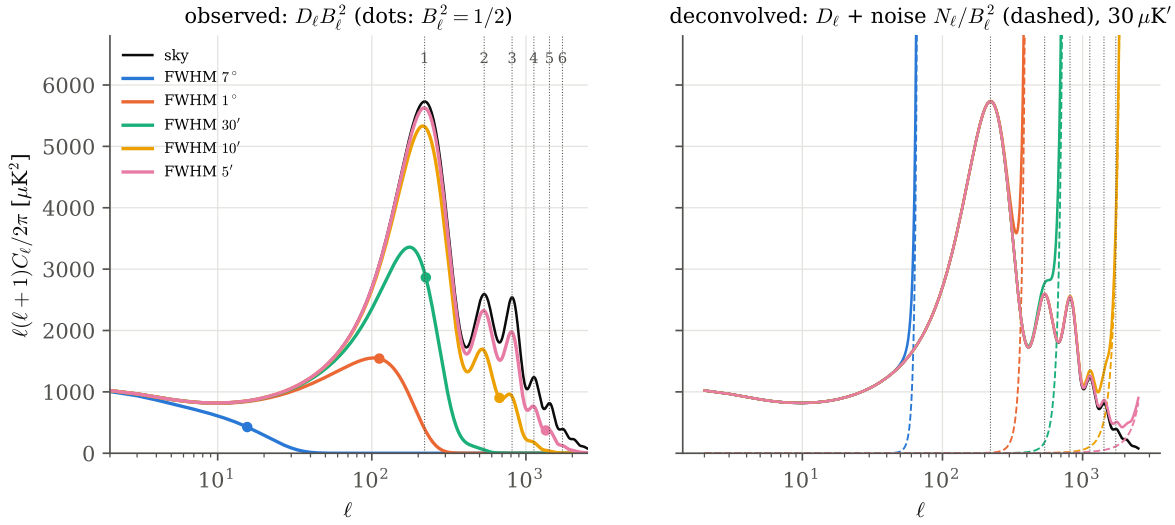


Figure 7.16: What a beam does to the acoustic peaks. Left: the fiducial spectrum D_ℓ (black) seen through Gaussian beams from 7° to 5′; the dotted lines mark the six peaks and each dot the multipole $\ell_{1/2}$ where the beam has removed half the power. The higher peaks go first. Right: the beam divided back out, for white noise of the same depth in every case. The sky is recovered (solid curves lie on the black one) until the deconvolved noise N_ℓ/B_ℓ^2 (dashed) rises through it like a wall.

Smoothing is the first example of an observational effect acting on a summary: the instrument applies a known linear operation to the field, and the power spectrum changes in a known way.

Chapter 8 adds two effects that are harder to undo: noise, which adds power, and the mask, which couples modes.

7.7 What the power spectrum throws away

Imagine a universe in which a first-order phase transition left behind thin shells, the walls of expanding bubbles, scattered at random. Its density map is a field of rings. Now take that map, Fourier transform it, keep the amplitude $|\delta_{\boldsymbol{k}}|$ of every mode exactly, replace every phase by a random number, and transform back. The result has precisely the same power spectrum as the rings, mode by mode, and it looks like nothing but Gaussian noise (Figure 7.17, panels a and b). Any analysis that measures only $P(k)$ would declare the two universes identical. The physics that made the rings is invisible to it.

The reason is in the definition of the estimator. Write each mode in polar form, $\delta_{\boldsymbol{k}} = |\delta_{\boldsymbol{k}}| e^{i\varphi_{\boldsymbol{k}}}$, with amplitude $|\delta_{\boldsymbol{k}}|$ and phase $\varphi_{\boldsymbol{k}}$. The estimator $\hat{P}(k)$ of (7.22) uses $|\delta_{\boldsymbol{k}}|^2$ and nothing else, so it is unchanged by any change of phases $\varphi_{\boldsymbol{k}} \to \varphi_{\boldsymbol{k}} + \alpha_{\boldsymbol{k}}$ with $\alpha_{-\boldsymbol{k}} = -\alpha_{\boldsymbol{k}}$ (the restriction keeps the reality condition (7.11), so the map stays real). Half of the numbers that describe the map, one phase per independent mode, are discarded.

The question. When is discarding the phases harmless, and what do the phases carry when it is not?

For a Gaussian field the phases are pure noise. In Section 7.3.3 we found that a homogeneous Gaussian field has independent, uniformly distributed phases, independent also of the amplitudes. There is nothing to learn from a uniform random number that the theory already told us was uniform. This is the Fourier-space form of the result of Section 7.2: for a Gaussian field the power spectrum is the whole story.

For a non-Gaussian field the phases carry the shapes. In a non-Gaussian field the phases of different modes are *correlated*. A thin ring is a sum of many plane waves whose crests line up along the ring. That alignment, a relation between the phases of different $\boldsymbol{k}$, is what makes a sharp, connected structure instead of a random interference pattern. The same holds for filaments, walls, clumps and the step-like edges of cosmic strings.

Why random phases look Gaussian. Randomising the phases makes almost any field look Gaussian. Each pixel of the randomised map is a sum over many modes, $\delta(\boldsymbol{x}) = \frac{1}{V}\sum_{\boldsymbol{k}} |\delta_{\boldsymbol{k}}| \cos(\boldsymbol{k}\cdot\boldsymbol{x} + \varphi_{\boldsymbol{k}})$ (pairing $\pm\boldsymbol{k}$), and the phases are now independent random numbers. By the central limit theorem (Chapter 2) a sum of many independent terms is close to Gaussian, whatever the amplitudes. So, as long as the power is spread over many modes, phase randomisation gives an almost Gaussian field with the same power spectrum.

Seeing it. The script `code/ch07/09_phases.py` builds a map of 70 thin rings of random radii on a 256×256 periodic grid, standardised to zero mean and unit variance, and its phase-randomised twin. The phases are taken from the FFT of fresh white noise, which has the Hermitian symmetry built in, so the twin is real automatically:

Listing 7.2: Phase randomisation: keep $|\delta_{\boldsymbol{k}}|$, draw new phases (`code/ch07/lib_fields.py`)

```
154 def randomise_phases(field, rng=None):
155     """Same |FFT| as field, phases replaced by those of fresh white noise (Hermitian
        symmetry automatic)."""
```

```
156     rng = np.random.default_rng() if rng is None else rng
157     field = np.asarray(field)
158     d = field.ndim
159     N = field.shape[0]
160     fk = np.fft.rfftn(field)
161     wk = np.fft.rfftn(rng.standard_normal(field.shape))
162     u = wk / np.abs(wk)                       # unit complex numbers, uniform phases
163     out = np.abs(fk) * u
164     out.flat[0] = fk.flat[0]                  # keep the mean
165     return np.fft.irfftn(out, s=[N] * d, axes=tuple(range(d)))
```

Lines 160–161 transform the map and a white-noise map of the same size. Line 162 reduces the white-noise modes to unit complex numbers, whose phases are uniform (Section 7.3.3) and already satisfy $u_{-\boldsymbol{k}} = u^*_{\boldsymbol{k}}$. Line 163 attaches them to the old amplitudes, line 164 keeps the $\boldsymbol{k} = 0$ mode (the mean), and line 165 transforms back.

The two maps have the same $\hat{P}(k)$ to 7.772×10^{-16}, rounding error (Figure 7.18, left). Everything else differs. *The one-point distribution.* For the rings it is strongly skewed, with skewness 1.856: most pixels sit on the flat background just below the mean, and a few on the bright rings. The twin has skewness 0.01711 and a Gaussian histogram, as the central-limit argument says. *The excursion sets* differ most of all. Count the connected regions where the field exceeds ν standard deviations (`scipy.ndimage.label` labels the connected clusters of pixels). At $\nu = 1$ the rings give 15 regions, chains of overlapping rings; their Gaussian twin gives 487 small blobs. The ring curve has a sharp spike near $\nu \approx 1.6$, the crest height of a single ring. The grid samples each crest at slightly different heights. So a threshold just below the crest height chops every ring into many short arcs, while a threshold just above it leaves only the places where rings overlap.

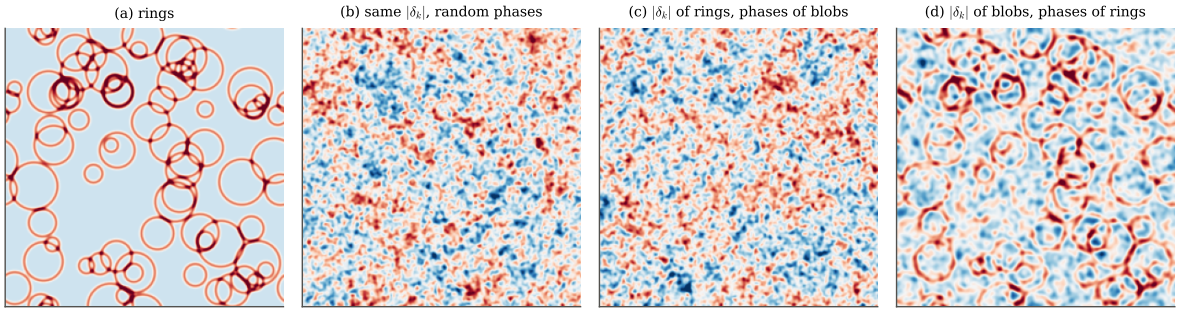


Figure 7.17: Amplitudes and phases. (a) A field of thin rings. (b) The same Fourier amplitudes with random phases: the same power spectrum, and the rings are gone. (c) The amplitudes of the rings with the phases of a Gaussian blob field: the phases are random, and so is the picture. (d) The amplitudes of the blob field with the phases of the rings: the rings come back. Where structures are, and what shape they have, is written in the phases.

The amplitudes say how much of each scale is present; the phases say where it is and how it is arranged. The phase swap in Figure 7.17(c, d) shows this directly. Give the ring map's amplitudes the phases of a Gaussian blob field, and the result is another random-looking Gaussian pattern: the phases of a Gaussian field are random, and they decide what the picture looks like. Give the blob amplitudes the ring phases, and the rings reappear, blurred because the blob amplitudes are weak at high k.

7.7.1 A taxonomy of summaries

The power spectrum is one choice of what to measure. Given a field $\phi : \mathcal{M} \to \mathbb{R}$ on a domain $\mathcal{M}$, the summaries used in physics fall into a few families, each answering a different question

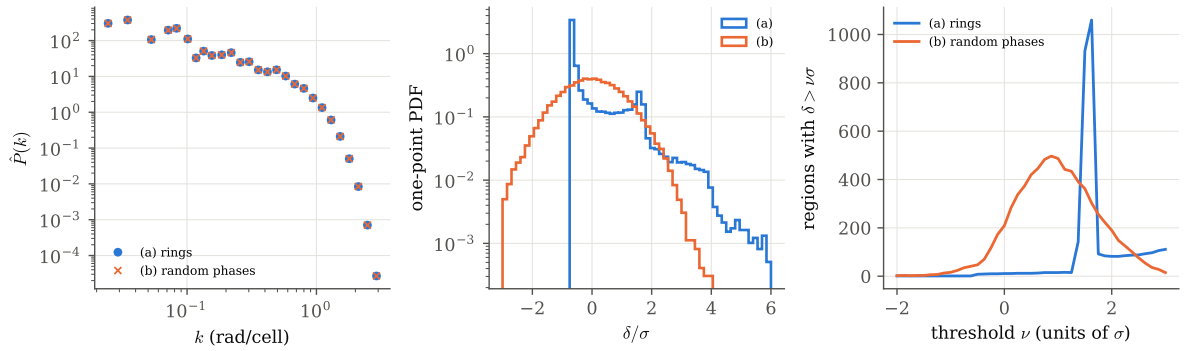


Figure 7.18: Three summaries of the ring map (a) and its phase-randomised twin (b). Left: identical power spectra. Middle: very different one-point distributions, skewed for the rings, Gaussian for the twin. Right: the number of connected regions above a threshold ν, which tells the two apart at all but the lowest and highest thresholds.

about the same data:

$$\phi \longrightarrow \begin{cases} \text{amplitude} & \langle\phi\rangle,\ \mathrm{Var}(\phi),\ P(\phi), \\ \text{correlation} & \xi(r),\ P(k),\ C_\ell, \\ \text{higher-order structure} & B,\ T, \ldots, \\ \text{geometry/morphology} & \text{sizes, shapes, curvature, excursion geometry,} \\ \text{topology} & \beta_i,\ \chi,\ \text{persistence.} \end{cases} \tag{7.37}$$

Each row answers its own question. *Amplitude:* how large is the field typically? It is read from the one-point distribution $P(\phi)$; the skewness above is an amplitude statistic. *Correlation:* how are values at different places related? This is everything for a Gaussian field, and blind to phases. *Higher-order structure:* the three- and four-point functions and their Fourier transforms, the bispectrum B and trispectrum T. The failed Wick test of the lognormal field in Figure 7.5 was a four-point statistic. *Geometry and topology:* what shapes does the field make in real space? An example is the excursion set $\{\boldsymbol{x} : \phi(\boldsymbol{x}) > \nu\}$, whose connected pieces we just counted. In topology β_0 counts connected components, β_1 counts holes, and χ, the Euler characteristic, combines them (Chapter 12).

Each of these summaries S, once chosen, is an ordinary random variable. Repeated universes give it a sampling distribution $P(S \mid \theta)$, with θ the parameters of the theory, and the machinery of the rest of the book applies to it unchanged.

The ring example raises the question behind the last row. Is there physical information in the way the regions of a field are connected that the statistic we are already using, here the power spectrum, does not expose? For the rings the answer is plainly yes, and the region count of Figure 7.18 already sees it. Chapter 12 builds that observation into a method: excursion sets at every threshold, their Betti numbers and Euler characteristic, and persistence, the record of how long each structure survives as the threshold moves. It also turns them into statistics with sampling distributions computed from exactly the Gaussian simulations of Section 7.4, because a Gaussian field with the measured $P(k)$ is the natural null hypothesis: it is the field that contains the two-point information and nothing else.

How this is used in real research: Simulated Gaussian fields: verification and research

Every simulation of a flat Gaussian field above has been a *verification*: we knew the answer analytically (the Wiener–Khinchin theorem, the recipe's normalisation, the exponential distribution of a mode, $\sqrt{2/N_{\mathrm{modes}}}$, $|W|^2P$) and the simulation confirmed it, which is how one tests both the code and one's understanding. In research the same machinery

is the instrument itself, used where no formula exists. Survey analyses estimate the covariance of $\hat{P}(k)$ or $\hat{\xi}(r)$ including the survey window, non-linear evolution and galaxy bias from hundreds to thousands of mock catalogues, typically generated from Gaussian or lognormal fields made exactly as in Section 7.4 or from N-body runs; the finite number of mocks itself biases the inverse covariance, and corrections for it are standard [6, 165]. Searches for non-Gaussianity in the CMB compare the data with large sets of Gaussian simulations that share the measured power spectrum [112]; phase-randomised "surrogate" maps play the same role for any field. On the sphere, the transfer of power by the mask and beam is calibrated on simulated skies (Chapter 8).

The two-point summaries are where the analysis of a field starts and, for a Gaussian field, where it ends. On the sky the plane waves become spherical harmonics, the power spectrum becomes C_ℓ, and the mode counting of Section 7.5.2 becomes cosmic variance; that is the CMB's version of the same story.

Recap

- A theory predicts the statistics of a field, not the field. A random field is a random variable at every point, specified by its n-point distributions; the ensemble is the stack of possible universes.
- Homogeneity and isotropy make the covariance a function of distance only, $\xi(r) = \langle \delta(\boldsymbol{x})\delta(\boldsymbol{x}+\boldsymbol{r})\rangle$, with $\xi(0) = \sigma^2$, $|\xi| \le \xi(0)$, and positive definiteness.
- A Gaussian field is fixed by ξ: the joint law of any set of points is $\mathcal{N}(\mathbf{0}, \mathbf{C})$ with $\mathbf{C}_{ij} = \xi(r_{ij})$; the mean field around a hot spot is $\nu\xi(r)/\sigma$; four-point moments are sums over pairings (Wick).
- Fourier modes of a homogeneous field are uncorrelated, $\langle \delta_{\boldsymbol{k}} \delta^*_{\boldsymbol{k}'} \rangle = (2\pi)^d \delta_D(\boldsymbol{k} - \boldsymbol{k}') P(k)$, with P the Fourier transform of ξ (Wiener–Khinchin) and $P \ge 0$. For a Gaussian field each mode has an exponential power and a uniform phase.
- Simulation: FFT white noise, multiply by $\sqrt{P(k)/\Delta^d}$, inverse FFT; a box holds only $k_f \le k \le k_{\mathrm{Ny}}$ and has zero mean.
- One universe: spatial averages replace ensemble averages only for ergodic fields; the mean over a volume has variance $P(0)/V$; a shell-averaged power spectrum has $\Delta\hat{P}/P = \sqrt{2/N_{\mathrm{modes}}}$, even without noise.
- Smoothing by a window multiplies the spectrum by $|W(k)|^2$; the spectral moments σ_j set the variance, the gradients and the feature spacing.
- $P(k)$ discards the phases. For a Gaussian field they are noise; for any other field they hold the shapes, which morphology and topology (Chapter 12) measure.

7.8 The Fourier modes of the sphere

Point a radio telescope tuned to millimetre wavelengths at any direction on the sky and it records a temperature of about 2.7255 K. Subtract that average and what remains is a pattern of hot and cold patches, typically about a hundred microkelvin above or below the average (that is the rms). This pattern is the cosmic microwave background (CMB) anisotropy. It is a random field on the sky: one real number $\Theta(\boldsymbol{e})$ for every direction $\boldsymbol{e}$. The number is the temperature

contrast $\delta T/T$. We always quote it multiplied by $T_0 = 2.7255 \times 10^6\ \mu$K, so that Θ and everything built from it are in microkelvin, the units CAMB uses (Chapter T1).

For fields in flat space (Sections 7.3 and 7.3.2) the key step was to expand the field in plane waves $e^{i\boldsymbol{k}\cdot\boldsymbol{x}}$. Each plane wave is a pattern of one definite wavelength. Homogeneity made the amplitudes of different plane waves uncorrelated, and the variance of each amplitude was the power spectrum $P(k)$. The sky, however, is a sphere, and a sphere has no plane waves: $e^{i\boldsymbol{k}\cdot\boldsymbol{x}}$ has no meaning on a curved, closed surface. So before any statistics we need a piece of geometry.

The question. Which patterns on the sphere play the role that plane waves play in flat space: patterns of a definite angular size, from which any sky can be built, and which a rotation of the sky only shuffles among themselves?

7.8.1 Spherical harmonics: what they are and why they are the right modes

The answer is the spherical harmonics. To see why, ask first what makes a plane wave special. It is an eigenfunction of the Laplacian: $\nabla^2 e^{i\boldsymbol{k}\cdot\boldsymbol{x}} = -k^2 e^{i\boldsymbol{k}\cdot\boldsymbol{x}}$. The eigenvalue $-k^2$ measures how rapidly the pattern wiggles, and $2\pi/k$ is its wavelength. The Laplacian also makes sense on the unit sphere, so the natural replacement for plane waves is the eigenfunctions of the Laplacian on the sphere,

$$\nabla^2_{S^2} Y = \frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\Big(\sin\theta\,\frac{\partial Y}{\partial\theta}\Big) + \frac{1}{\sin^2\theta}\frac{\partial^2 Y}{\partial\phi^2} = -\ell(\ell+1)\,Y, \tag{7.38}$$

where $\theta \in [0,\pi]$ is the colatitude (angle from the north pole) and $\phi \in [0, 2\pi)$ the longitude of the direction $\boldsymbol{e}$. We have met these functions before: they are the angular parts of the hydrogen wavefunctions. In quantum mechanics (with $\hbar = 1$) the orbital angular momentum $\boldsymbol{L} = -i\,\boldsymbol{r}\times\boldsymbol{\nabla}$ satisfies $\boldsymbol{L}^2 = -\nabla^2_{S^2}$, and the **spherical harmonics** $Y_{\ell m}(\theta,\phi)$ are the simultaneous eigenfunctions

$$\boldsymbol{L}^2\,Y_{\ell m} = \ell(\ell+1)\,Y_{\ell m}, \qquad L_z\,Y_{\ell m} = m\,Y_{\ell m}, \qquad \ell = 0,1,2,\ldots, \quad m = -\ell,\ldots,\ell, \tag{7.39}$$

with $L_z = -i\,\partial/\partial\phi$. So $Y_{\ell m} \propto e^{im\phi}$, and in general $Y_{\ell m}(\theta,\phi) = N_{\ell m} P_\ell^m(\cos\theta)\,e^{im\phi}$, with P_ℓ^m an associated Legendre function and $N_{\ell m}$ a normalisation constant. We use the standard (Condon–Shortley) phase convention of quantum mechanics, the one healpy and CAMB also use. The first few, written out so that we can compute with them, are

$$\begin{aligned}
&Y_{00} = \frac{1}{\sqrt{4\pi}}, && Y_{10} = \sqrt{\frac{3}{4\pi}}\cos\theta, && Y_{1,\pm1} = \mp\sqrt{\frac{3}{8\pi}}\sin\theta\,e^{\pm i\phi},\\
&Y_{20} = \sqrt{\frac{5}{16\pi}}\,(3\cos^2\theta - 1), && Y_{2,\pm1} = \mp\sqrt{\frac{15}{8\pi}}\sin\theta\cos\theta\,e^{\pm i\phi}, && Y_{2,\pm2} = \sqrt{\frac{15}{32\pi}}\sin^2\theta\,e^{\pm 2i\phi}.
\end{aligned} \tag{7.40}$$

Three properties of the $Y_{\ell m}$ are all we will use.

1. **Orthonormality.** With $\mathrm{d}\Omega = \sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi$ the solid-angle element,

$$\int \mathrm{d}\Omega\; Y_{\ell m}(\boldsymbol{e})\,Y^*_{\ell' m'}(\boldsymbol{e}) = \delta_{\ell\ell'}\delta_{mm'}. \tag{7.41}$$

   In quantum mechanics this is $\langle \ell m|\ell' m'\rangle = \delta_{\ell\ell'}\delta_{mm'}$: eigenfunctions of a Hermitian operator with different eigenvalues are orthogonal.

2. **Completeness.** Any square-integrable function on the sphere is a sum of $Y_{\ell m}$'s, just as any periodic function is a Fourier series.

3. **Complex conjugation flips m.** Since $e^{im\phi}$ becomes $e^{-im\phi}$ and the Condon–Shortley phase carries a sign $(-1)^m$,

$$Y^*_{\ell m} = (-1)^m\, Y_{\ell,-m}. \tag{7.42}$$

Check it on (7.40): $Y^*_{11} = -\sqrt{3/8\pi}\sin\theta\, e^{-i\phi} = -Y_{1,-1}$, which is $(-1)^1 Y_{1,-1}$.

What do ℓ and m mean on the sky? In short, ℓ sets the angular size of the pattern and m sets how its zero lines are oriented. Figure 7.19 shows the real parts of six harmonics in the Mollweide projection, the equal-area map of the whole sky in which the equator is the horizontal midline and the poles are the top and bottom points. The real part of $Y_{\ell m}$ has ℓ nodal lines (lines where it is zero) in total: m of them are great circles through the poles (the factor $\cos m\phi$ vanishes on $2m$ half-meridians) and $\ell - |m|$ are circles of constant latitude (the zeros of P_ℓ^m). The $m = 0$ harmonics depend on latitude only ("zonal"), the $|m| = \ell$ ones on longitude only ("sectoral"), and the rest form a chessboard ("tesseral"). In every case the distance between neighbouring nodal lines is about π/ℓ radians, so

$$\text{multipole } \ell \quad\longleftrightarrow\quad \text{angular scale } \vartheta \approx \frac{\pi}{\ell} = \frac{180^\circ}{\ell}. \tag{7.43}$$

This is the sphere's version of "wavenumber $k \leftrightarrow$ half-wavelength π/k".

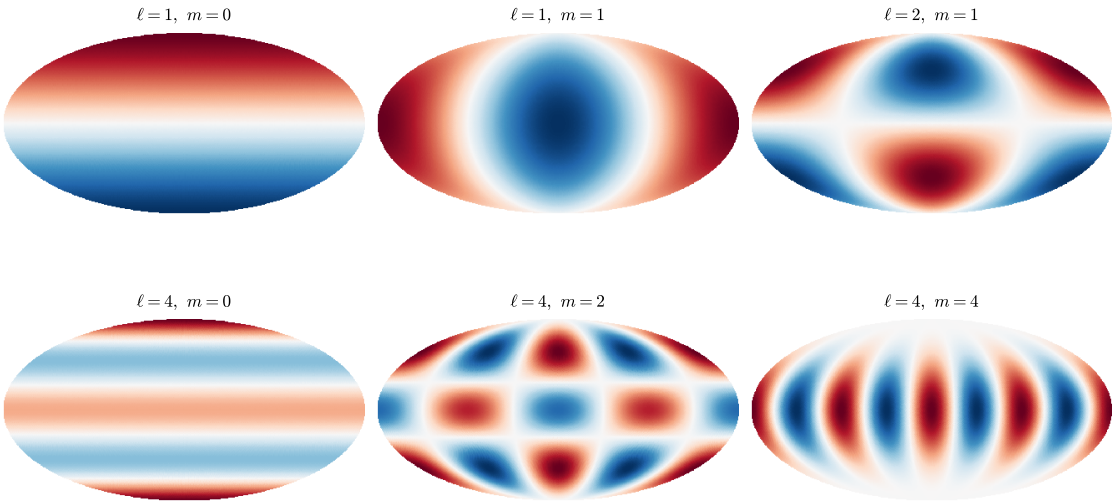


Figure 7.19: Real parts of six spherical harmonics on the whole sky (Mollweide projection, north pole at the top; red positive, blue negative). Top row: the $\ell = 1$ dipole along the pole ($m = 0$) and along the equator ($m = 1$), and a quadrupole with $m = 1$. Bottom row: three $\ell = 4$ harmonics. With $m = 0$ the pattern has four latitude rings of zeros; with $m = 4$ it has eight half-meridians of zeros; $m = 2$ mixes the two. All three bottom patterns have the same angular scale, about $180^\circ/4 = 45^\circ$.

7.8.2 Expanding a sky: the coefficients $a_{\ell m}$

Completeness (property 2 above) lets us write any sky as a sum of harmonics, and orthonormality (property 1) tells us how to find the coefficients. Every CMB analysis starts with this step [HobsonBMC (10.2)], [Hivon (4)–(5)].

Working definition 7.5 (spherical-harmonic coefficients)

A sky $\Theta(\boldsymbol{e})$ is expanded as

$$\Theta(\boldsymbol{e}) = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} a_{\ell m}\, Y_{\ell m}(\boldsymbol{e}), \qquad a_{\ell m} = \int \mathrm{d}\Omega\; \Theta(\boldsymbol{e})\, Y^*_{\ell m}(\boldsymbol{e}). \tag{7.44}$$

The coefficient $a_{\ell m}$ [one complex number per (ℓ, m), the overlap of the sky with the pattern $Y_{\ell m}$] measures how much of the pattern $Y_{\ell m}$ the sky contains. *Operationally:* multiply the map pixel by pixel by $Y^*_{\ell m}$, multiply by the pixel area, and add up (Section 7.11.2).

Why is the coefficient given by that integral? Because the harmonics are orthonormal, so projecting the sky onto one harmonic picks out its coefficient and nothing else. In steps: multiply the expansion by $Y^*_{\ell' m'}(\boldsymbol{e})$ and integrate over the sky,

$$\int \mathrm{d}\Omega\; \Theta(\boldsymbol{e})\, Y^*_{\ell' m'}(\boldsymbol{e}) \overset{(1)}{=} \sum_{\ell m} a_{\ell m} \int \mathrm{d}\Omega\; Y_{\ell m}(\boldsymbol{e})\, Y^*_{\ell' m'}(\boldsymbol{e}) \overset{(2)}{=} \sum_{\ell m} a_{\ell m}\, \delta_{\ell \ell'} \delta_{m m'} \overset{(3)}{=} a_{\ell' m'}. \tag{7.45}$$

(1) Insert the expansion and exchange the (convergent) sum with the integral.
(2) Orthonormality (7.41).
(3) The Kronecker deltas keep only the term $\ell = \ell'$, $m = m'$.

Cosmological analyses ignore the first two multipoles and start at $\ell = 2$ [HobsonBMC p. 230]. The reason is that neither carries the primordial pattern. The monopole a_{00} is $\sqrt{4\pi}$ times the sky average (because $Y_{00} = 1/\sqrt{4\pi}$), and we subtracted that average. The dipole $\ell = 1$ is dominated by our own motion through the CMB, which Doppler-shifts the temperature by a few millikelvin.

The sky is real, so half of the coefficients are redundant. The $a_{\ell m}$ are complex, but a temperature is a real number, $\Theta^* = \Theta$. What does that imply for the coefficients? Each negative-m coefficient is fixed by its positive-m partner:

$$a_{\ell,-m} \overset{(1)}{=} \int \mathrm{d}\Omega\; \Theta\, Y^*_{\ell,-m} \overset{(2)}{=} \int \mathrm{d}\Omega\; \Theta\, (-1)^m Y_{\ell m} \overset{(3)}{=} (-1)^m \left(\int \mathrm{d}\Omega\; \Theta\, Y^*_{\ell m} \right)^* \overset{(4)}{=} (-1)^m\, a^*_{\ell m}. \tag{7.46}$$

(1) Definition (7.44) with $m \to -m$.
(2) Conjugate (7.42): $Y^*_{\ell,-m} = (-1)^{-m} Y_{\ell m} = (-1)^m Y_{\ell m}$.
(3) Θ is real, so $\Theta Y_{\ell m} = (\Theta Y^*_{\ell m})^*$, and the conjugate can be taken outside the integral.
(4) Definition (7.44) again.

Two consequences follow. For $m = 0$, (7.46) reads $a_{\ell 0} = a^*_{\ell 0}$, so $a_{\ell 0}$ *is real*. For $m \neq 0$ the negative-m coefficients carry no new information, because the positive-m ones fix them. So at each multipole ℓ the sky is described by

$$\underbrace{1}_{a_{\ell 0}\ \text{real}} + \underbrace{2\ell}_{\mathrm{Re}\, a_{\ell m},\ \mathrm{Im}\, a_{\ell m},\ m=1,\ldots,\ell} = 2\ell + 1 \text{ real numbers}. \tag{7.47}$$

This count becomes the number of degrees of freedom of the power-spectrum estimator in Section 7.10. A second way to reach it: count the independent real patterns at multipole ℓ. They are the real and imaginary parts of the $Y_{\ell m}$ with $m \geq 0$, which is $2(\ell + 1)$ functions, minus $\mathrm{Im}\, Y_{\ell 0}$, which is identically zero. That leaves $2\ell + 1$, the same number. A real sky band-limited to $\ell \leq \ell_{\max}$ is therefore $\sum_{\ell=0}^{\ell_{\max}} (2\ell + 1) = (\ell_{\max} + 1)^2$ real numbers. Libraries store only $m \geq 0$: $(\ell_{\max} + 1)(\ell_{\max} + 2)/2$ complex numbers.

The total power of a sky. How much of a sky's variance sits at each multipole? The mean square of the sky over the sphere is a sum of $|a_{\ell m}|^2$. This is the sphere's version of Parseval's theorem:

$$\frac{1}{4\pi}\int \mathrm{d}\Omega\, \Theta^2(\boldsymbol{e}) \overset{(1)}{=} \frac{1}{4\pi}\sum_{\ell m}\sum_{\ell' m'} a_{\ell m}\, a^*_{\ell' m'} \int \mathrm{d}\Omega\, Y_{\ell m} Y^*_{\ell' m'} \overset{(2)}{=} \frac{1}{4\pi}\sum_{\ell}\sum_{m=-\ell}^{\ell} |a_{\ell m}|^2. \tag{7.48}$$

(1) Write $\Theta^2 = \Theta\,\Theta^*$ and expand the first factor in $Y_{\ell m}$ and the second in $Y^*_{\ell' m'}$ (the conjugate of the expansion, since Θ is real).
(2) Orthonormality collapses the double sum to $\ell = \ell'$, $m = m'$.

So $\sum_m |a_{\ell m}|^2$ (divided by 4π) is the share of the sky's mean square that lives at angular scale π/ℓ. Divided by the number of terms, $2\ell + 1$, it will be our estimator $\widehat{C}_\ell$ of the power spectrum (Section 7.10).

Example: the dipole along the pole. Take $\Theta(\boldsymbol{e}) = A\cos\theta$, hot in the north and cold in the south, with A the temperature at the north pole. The pattern has the shape of $Y_{10} = \sqrt{3/4\pi}\cos\theta$, so we expect only a_{10} to be nonzero. Its value is

$$a_{10} \overset{(1)}{=} \int \mathrm{d}\Omega\, A\cos\theta \sqrt{\frac{3}{4\pi}}\cos\theta \overset{(2)}{=} A\sqrt{\frac{3}{4\pi}}\, 2\pi \int_{-1}^{1} \mu^2\, \mathrm{d}\mu \overset{(3)}{=} A\sqrt{\frac{3}{4\pi}}\,\frac{4\pi}{3} = A\sqrt{\frac{4\pi}{3}}, \tag{7.49}$$

(1) Definition (7.44); Y_{10} is real.
(2) Substitute $\mu = \cos\theta$, $\mathrm{d}\Omega = \mathrm{d}\mu\,\mathrm{d}\phi$; the ϕ integral gives 2π.
(3) $\int_{-1}^{1} \mu^2 \mathrm{d}\mu = 2/3$.

and every other $a_{\ell m}$ vanishes, because $\cos\theta$ is proportional to Y_{10} and Y_{10} is orthogonal to all the other $Y_{\ell m}$. The script `code/ch07/21_spherical_harmonics.py` builds this map with $A = 3\,\mu$K on a HEALPix grid (Section 7.11.2) and computes the coefficients numerically: $a_{10} = 6.139960\,\mu$K against $A\sqrt{4\pi/3} = 6.139960\,\mu$K, and the largest of all the other coefficients up to $\ell = 4$ is $1.009 \times 10^{-15}\,\mu$K, zero to machine precision.

Example: the same dipole along the equator. Now turn the dipole on its side: $\Theta = B\sin\theta\cos\phi$, hottest at $\phi = 0$ on the equator. The question is how its coefficients compare with those of the polar dipole. Write $\cos\phi = (e^{i\phi} + e^{-i\phi})/2$ and read $\sin\theta\, e^{\pm i\phi}$ off (7.40):

$$B\sin\theta\cos\phi \overset{(1)}{=} \frac{B}{2}\left(\sin\theta\, e^{i\phi} + \sin\theta\, e^{-i\phi}\right) \overset{(2)}{=} \frac{B}{2}\sqrt{\frac{8\pi}{3}}\left(-Y_{11} + Y_{1,-1}\right) = B\sqrt{\frac{2\pi}{3}}\left(-Y_{11} + Y_{1,-1}\right). \tag{7.50}$$

(1) Euler's formula.
(2) From (7.40): $\sin\theta\, e^{i\phi} = -\sqrt{8\pi/3}\, Y_{11}$ and $\sin\theta\, e^{-i\phi} = +\sqrt{8\pi/3}\, Y_{1,-1}$.

So $a_{11} = -B\sqrt{2\pi/3}$ and $a_{1,-1} = +B\sqrt{2\pi/3}$, with $a_{10} = 0$. The reality condition checks out: $(-1)^1 a^*_{11} = +B\sqrt{2\pi/3} = a_{1,-1}$.

Now compare the two dipoles (take $B = A$). They are the same pattern, rotated by 90°. Their coefficients are completely different: one sits in $m = 0$, the other in $m = \pm 1$. But the sum of $|a_{\ell m}|^2$ over m is the same: $|a_{10}|^2 = 4\pi A^2/3$ for the first and $|a_{11}|^2 + |a_{1,-1}|^2 = 2 \times 2\pi B^2/3 = 4\pi B^2/3$ for the second. A rotation reshuffles the $a_{\ell m}$ within one ℓ and leaves $\sum_m |a_{\ell m}|^2$ alone. This is the seed of everything in Section 7.9.

Example: how many numbers is a sky? By the count (7.47), a sky band-limited to $\ell_{\max}$ needs about $\ell_{\max}^2$ real numbers, and $\ell_{\max}$ is set by the telescope's resolution through (7.43). The Planck satellite resolves features of about 5′, which is $\ell \approx 180^\circ/5' \approx 2000$. A sky band-limited to $\ell_{\max} = 600$, roughly the resolution of the earlier WMAP satellite, already has $(601)^2 - 4 = 361197$ real numbers in $2 \le \ell \le 600$, stored by healpy as 180901 complex numbers with $m \ge 0$ (the storage includes $\ell = 0, 1$ and the zero imaginary parts of the $m = 0$ entries). Figure 7.20 shows one random sky drawn from the fiducial cosmological model (Section 7.11.1) and built up from its harmonics: with $\ell \le 4$ we see a few continent-sized blobs; each factor of four in $\ell_{\max}$ shrinks the smallest features by a factor of four.

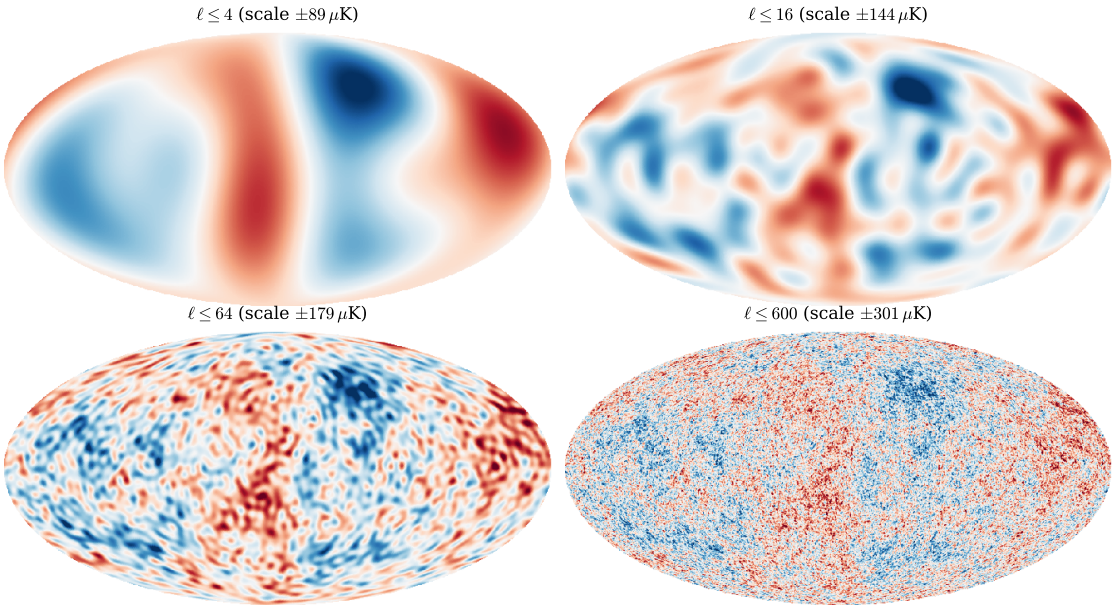


Figure 7.20: One simulated CMB sky (fiducial ΛCDM spectrum) summed up to $\ell_{\max} = 4$, 16, 64 and 600. The coefficients are the same in all four panels; each panel simply keeps more of them. The colour scale is set separately in each panel, because adding multipoles adds variance. The large-scale pattern of the first panel is still visible underneath the fine structure of the last.

7.9 No preferred direction: why the $a_{\ell m}$ are uncorrelated

Inflation predicts the statistics of the primordial fluctuations, and linear evolution carries them forward to the last scattering surface. The particular hot and cold spots we observe constitute one draw from the resulting ensemble of skies (Chapter T1). We assume statistical isotropy, meaning that rotating the ensemble leaves its probability law unchanged. An individual sky can still contain a striking feature in a particular direction. Isotropy means that, across realisations, the feature has no preferred orientation.

In flat space the matching symmetry was homogeneity (invariance under shifts), and it made the Fourier amplitudes of different wavevectors uncorrelated, $\langle \delta_{\boldsymbol{k}} \delta^*_{\boldsymbol{k}'} \rangle \propto \delta(\boldsymbol{k} - \boldsymbol{k}') P(k)$ (Section 7.3.2). The reason was simple: shifting a plane wave only changes its phase. Rotations of the sphere are harder, because rotations about different axes do not commute and a general rotation mixes harmonics with different m. The tool that handles this is one we already know, the angular-momentum algebra of quantum mechanics.

The question. If no direction on the sky is special, which pairs of coefficients $a_{\ell m}$, $a_{\ell' m'}$ can be correlated, and can the variance of $a_{\ell m}$ depend on m?

Working definition 7.6 (statistical isotropy)

A random sky $\Theta(\boldsymbol{e})$ is **statistically isotropic** if for every rotation R the rotated sky $\Theta(R^{-1}\boldsymbol{e})$ has the same joint distribution as $\Theta(\boldsymbol{e})$ [same probabilities for the temperatures at any finite set of directions]. For the mean and the two-point function this means

$$\langle \Theta(\boldsymbol{e}) \rangle = 0, \qquad \langle \Theta(\boldsymbol{e})\Theta(\boldsymbol{e}') \rangle = C(\boldsymbol{e}\cdot\boldsymbol{e}'), \tag{7.51}$$

where C is the **angular correlation function** [the average product of the temperatures at two directions, a function of the angle ϑ between them only, with $\cos\vartheta = \boldsymbol{e}\cdot\boldsymbol{e}'$] [HobsonBMC (10.1)]. Angle brackets are averages over the ensemble of skies, never over our one sky.

The mean is zero because we subtracted the average temperature; the correlation depends on the angle alone because a rotation can carry any pair of directions to any other pair with the same separation.

7.9.1 The covariance of two coefficients is a double integral

Start from the definition of the coefficients (7.44) and average:

$$\begin{aligned} \langle a_{\ell m}\, a^*_{\ell' m'} \rangle &\overset{(1)}{=} \left\langle \int \mathrm{d}\Omega\, \Theta(\boldsymbol{e}) Y^*_{\ell m}(\boldsymbol{e}) \int \mathrm{d}\Omega'\, \Theta(\boldsymbol{e}') Y_{\ell' m'}(\boldsymbol{e}') \right\rangle \\ &\overset{(2)}{=} \int \mathrm{d}\Omega \int \mathrm{d}\Omega'\, \langle \Theta(\boldsymbol{e})\Theta(\boldsymbol{e}') \rangle\, Y^*_{\ell m}(\boldsymbol{e})\, Y_{\ell' m'}(\boldsymbol{e}') \\ &\overset{(3)}{=} \int \mathrm{d}\Omega \int \mathrm{d}\Omega'\, C(\boldsymbol{e}\cdot\boldsymbol{e}')\, Y^*_{\ell m}(\boldsymbol{e})\, Y_{\ell' m'}(\boldsymbol{e}'). \end{aligned} \tag{7.52}$$

(1) Definition (7.44) for both coefficients; conjugating $a_{\ell' m'}$ conjugates the harmonic but not the real sky.
(2) The average of a sum (an integral) is the sum of the averages, and the harmonics are fixed functions, not random.
(3) Isotropy (7.51).

Since $\langle a_{\ell m} \rangle = \int \mathrm{d}\Omega\, \langle \Theta \rangle Y^*_{\ell m} = 0$, this average is the covariance of the two coefficients. The question has become: what does this double integral give, given that its kernel $C(\boldsymbol{e}\cdot\boldsymbol{e}')$ depends only on the angle between the two directions? We answer it in two steps. First we use only the rotations about the pole, which need nothing but the factor $e^{im\phi}$ in each harmonic. Then we use all rotations, which need the angular-momentum algebra.

7.9.2 Rotations about the pole: m must equal m'

The question. What does a rotation about the pole do to the coefficients? Rotate the sky by an angle α about the z-axis. The new sky has, at longitude ϕ, the temperature the old sky had at $\phi - \alpha$: $\Theta_\alpha(\theta,\phi) = \Theta(\theta, \phi - \alpha)$. Its coefficients are

$$a^{(\alpha)}_{\ell m} \overset{(1)}{=} \int \mathrm{d}\Omega\, \Theta(\theta, \phi-\alpha)\, Y^*_{\ell m}(\theta,\phi) \overset{(2)}{=} \int \mathrm{d}\Omega'\, \Theta(\theta,\phi')\, Y^*_{\ell m}(\theta, \phi'+\alpha) \overset{(3)}{=} e^{-im\alpha}\, a_{\ell m}. \tag{7.53}$$

(1) Definition (7.44) applied to the rotated sky.
(2) Substitute $\phi' = \phi - \alpha$; the solid-angle element and the range of a full turn are unchanged.
(3) $Y_{\ell m} \propto e^{im\phi}$, so $Y^*_{\ell m}(\theta, \phi'+\alpha) = e^{-im\alpha} Y^*_{\ell m}(\theta,\phi')$.

So a rotation about the pole only changes the phase of each coefficient, by an angle proportional to m.

Now use isotropy. The rotated sky is a draw from the same ensemble, so its coefficients have the same covariances as the original ones. Hence, for every angle α,

$$\langle a_{\ell m} a^*_{\ell' m'} \rangle = \langle a^{(\alpha)}_{\ell m} a^{(\alpha)*}_{\ell' m'} \rangle = e^{-i(m-m')\alpha} \langle a_{\ell m} a^*_{\ell' m'} \rangle. \tag{7.54}$$

If $m \neq m'$ we can pick α with $e^{-i(m-m')\alpha} \neq 1$ (for example $\alpha = \pi/|m-m'|$, which gives -1). Then the covariance equals -1 times itself, and the only such number is zero. So coefficients with different m are uncorrelated. This is the same logic as momentum conservation: a symmetry under shifts of ϕ forbids correlations between different eigenvalues of the generator of those shifts, $L_z = -i\partial_\phi$.

Example: turning the equatorial dipole. The dipole $B \sin\theta \cos\phi$ of Section 7.8.2 had $a_{11} = -B\sqrt{2\pi/3}$. Rotate it by $\alpha = 90°$ about the pole: the hot spot moves from $\phi = 0$ to $\phi = 90°$, and the sky becomes $B \sin\theta \sin\phi$. Equation (7.53) predicts $a_{11} \to e^{-i\pi/2} a_{11} = iB\sqrt{2\pi/3}$. Directly, with $\sin\phi = (e^{i\phi} - e^{-i\phi})/2i$ and the same table (7.40),

$$B \sin\theta \sin\phi \overset{(1)}{=} \frac{B}{2i}\left(\sin\theta e^{i\phi} - \sin\theta e^{-i\phi}\right) \overset{(2)}{=} \frac{B}{2i}\sqrt{\frac{8\pi}{3}}\left(-Y_{11} - Y_{1,-1}\right) \overset{(3)}{=} iB\sqrt{\frac{2\pi}{3}}\, Y_{11} + iB\sqrt{\frac{2\pi}{3}}\, Y_{1,-1}, \tag{7.55}$$

(1) Euler's formula for $\sin\phi$.
(2) $\sin\theta e^{i\phi} = -\sqrt{8\pi/3}\, Y_{11}$ and $\sin\theta e^{-i\phi} = \sqrt{8\pi/3}\, Y_{1,-1}$.
(3) $-1/(2i) = i/2$, and $\frac{1}{2}\sqrt{8\pi/3} = \sqrt{2\pi/3}$.

so $a_{11} = iB\sqrt{2\pi/3}$, as predicted. Only the phase moved; $|a_{11}|$ did not.

7.9.3 All rotations: one ℓ at a time, and no dependence on m

Rotations about other axes mix different m, so the phase argument no longer works on its own. The idea instead is to write the covariance as a matrix element of an operator, and then use what quantum mechanics tells us about operators that commute with rotations. Define the **covariance operator** $\hat{C}$, which acts on a function f on the sphere by smoothing it with the correlation function,

$$(\hat{C} f)(\boldsymbol{e}) = \int \mathrm{d}\Omega'\; C(\boldsymbol{e}\cdot\boldsymbol{e}')\, f(\boldsymbol{e}'), \tag{7.56}$$

and use the inner product $\langle f|g\rangle = \int \mathrm{d}\Omega\, f^* g$ of wavefunctions on the sphere, with $|\ell m\rangle$ standing for $Y_{\ell m}$. Then (7.52) reads

$$\langle a_{\ell m}\, a^*_{\ell' m'} \rangle = \int \mathrm{d}\Omega\; Y^*_{\ell m}(\boldsymbol{e})\, (\hat{C} Y_{\ell' m'})(\boldsymbol{e}) = \langle \ell m | \hat{C} | \ell' m' \rangle. \tag{7.57}$$

The covariance matrix of the coefficients is the matrix of $\hat{C}$ in the basis of spherical harmonics, exactly as the matrix of a Hamiltonian in the $|\ell m\rangle$ basis.

The covariance operator commutes with every rotation. Let U_R rotate a function, $(U_R f)(\boldsymbol{e}) = f(R^{-1}\boldsymbol{e})$. Then

$$\begin{aligned}(U_R \hat{C} f)(\boldsymbol{e}) &\overset{(1)}{=} \int \mathrm{d}\Omega'\; C\big((R^{-1}\boldsymbol{e})\cdot\boldsymbol{e}'\big)\, f(\boldsymbol{e}') \overset{(2)}{=} \int \mathrm{d}\Omega''\; C\big((R^{-1}\boldsymbol{e})\cdot(R^{-1}\boldsymbol{e}'')\big)\, f(R^{-1}\boldsymbol{e}'') \\ &\overset{(3)}{=} \int \mathrm{d}\Omega''\; C(\boldsymbol{e}\cdot\boldsymbol{e}'')\, (U_R f)(\boldsymbol{e}'') \overset{(4)}{=} (\hat{C} U_R f)(\boldsymbol{e}).\end{aligned} \tag{7.58}$$

(1) Definition of U_R applied to the function $\hat{C}f$.
(2) Change variables $\boldsymbol{e}' = R^{-1}\boldsymbol{e}''$; a rotation preserves solid angle, so $\mathrm{d}\Omega' = \mathrm{d}\Omega''$.
(3) A rotation preserves dot products, $(R^{-1}\boldsymbol{a})\cdot(R^{-1}\boldsymbol{b}) = \boldsymbol{a}\cdot\boldsymbol{b}$; and $f(R^{-1}\boldsymbol{e}'') = (U_R f)(\boldsymbol{e}'')$.
(4) Definition (7.56).

So $U_R\hat{C} = \hat{C}U_R$ for every rotation. Now take a rotation by a small angle ε about a unit axis $\boldsymbol{n}$. In quantum mechanics the angular momentum is the generator of rotations, $U_R = 1 - i\varepsilon\,\boldsymbol{n}\cdot\boldsymbol{L} + O(\varepsilon^2)$. Insert this into $U_R\hat{C} = \hat{C}U_R$: the terms $\hat{C}$ cancel, and the terms of order ε give $(\boldsymbol{n}\cdot\boldsymbol{L})\hat{C} = \hat{C}(\boldsymbol{n}\cdot\boldsymbol{L})$. Taking $\boldsymbol{n}$ along x, y and z in turn, and then noting that $\boldsymbol{L}^2$ and $L_\pm$ are built from L_x, L_y, L_z, we get

$$[\hat{C}, L_x] = [\hat{C}, L_y] = [\hat{C}, L_z] = 0 \quad\Longrightarrow\quad [\hat{C}, \boldsymbol{L}^2] = 0, \qquad [\hat{C}, L_\pm] = 0, \tag{7.59}$$

with the ladder operators $L_\pm = L_x \pm iL_y$. These commutators answer the question in three short steps.

Step 1: different ℓ are uncorrelated. Sandwich $\hat{C}\boldsymbol{L}^2 = \boldsymbol{L}^2\hat{C}$ between $\langle \ell m|$ and $|\ell' m'\rangle$:

$$\ell'(\ell'+1)\langle \ell m|\hat{C}|\ell' m'\rangle \overset{(1)}{=} \langle \ell m|\hat{C}\boldsymbol{L}^2|\ell' m'\rangle \overset{(2)}{=} \langle \ell m|\boldsymbol{L}^2\hat{C}|\ell' m'\rangle \overset{(3)}{=} \ell(\ell+1)\langle \ell m|\hat{C}|\ell' m'\rangle. \tag{7.60}$$

(1) $\boldsymbol{L}^2|\ell' m'\rangle = \ell'(\ell'+1)|\ell' m'\rangle$ from (7.39).
(2) $\hat{C}$ commutes with $\boldsymbol{L}^2$ (7.59).
(3) $\boldsymbol{L}^2$ is Hermitian, so it acts to the left on $\langle \ell m|$ with its real eigenvalue.

If $\ell \neq \ell'$ then $\ell(\ell+1) \neq \ell'(\ell'+1)$ and the matrix element vanishes.

Step 2: different m are uncorrelated. The same sandwich with L_z gives $(m - m')\langle \ell m|\hat{C}|\ell m'\rangle = 0$. This is the infinitesimal version of (7.54).

Step 3: the variance does not depend on m. After steps 1 and 2, only the diagonal elements $c_m \equiv \langle \ell m|\hat{C}|\ell m\rangle = \langle |a_{\ell m}|^2\rangle$ survive. The question is whether they can differ from one m to the next. The ladder operators move between neighbouring m with the real, positive coefficients of quantum mechanics, $L_+|\ell m\rangle = \lambda_m|\ell, m+1\rangle$ and $L_-|\ell, m+1\rangle = \lambda_m|\ell m\rangle$, where $\lambda_m = \sqrt{\ell(\ell+1) - m(m+1)}$. Compute one matrix element in two ways:

$$\lambda_m\, c_{m+1} \overset{(1)}{=} \langle \ell, m+1|\hat{C}\,L_+|\ell m\rangle \overset{(2)}{=} \langle \ell, m+1|L_+\hat{C}|\ell m\rangle \overset{(3)}{=} \big(L_-|\ell, m+1\rangle\big)^\dagger \hat{C}|\ell m\rangle \overset{(4)}{=} \lambda_m\, c_m. \tag{7.61}$$

(1) $L_+|\ell m\rangle = \lambda_m|\ell, m+1\rangle$, and the definition of c_{m+1}.
(2) $\hat{C}$ commutes with L_+ (7.59).
(3) The adjoint of L_+ is L_- (because L_x and L_y are Hermitian), so L_+ acting to the left is L_- acting on the bra's ket.
(4) $L_-|\ell, m+1\rangle = \lambda_m|\ell m\rangle$ with λ_m real, and the definition of c_m.

For $m = -\ell, \ldots, \ell - 1$ the factor λ_m is strictly positive (it vanishes only at $m = \ell$), so we may divide by it: $c_{m+1} = c_m$. Walking from $m = -\ell$ up to $m = \ell$, all $2\ell + 1$ diagonal elements are equal. Call their common value C_ℓ.

Key idea: Isotropy diagonalises the sky

For a statistically isotropic sky,

$$\langle a_{\ell m}\rangle = 0, \qquad \langle a_{\ell m}\, a^*_{\ell' m'}\rangle = C_\ell\,\delta_{\ell\ell'}\,\delta_{mm'}. \tag{7.62}$$

The **angular power spectrum** C_ℓ is the variance of each coefficient at multipole ℓ, the same

for all $2\ell+1$ values of m [HobsonBMC (10.3)], [Hivon (7)].

In group-theory language this is Schur's lemma: the harmonics with a given ℓ carry an irreducible representation of the rotation group, and an operator that commutes with an irreducible representation is a multiple of the identity on it. The ladder-operator argument above is the proof of that lemma in the one case we need. The result also shows that C_ℓ is *all* an isotropic two-point function can contain: one number per ℓ.

The result also fixes the average of a product without a conjugate. The reality condition (7.46), read with $m \to -m'$, says $a_{\ell' m'} = (-1)^{m'} a^*_{\ell', -m'}$. Insert this and use (7.62): $\langle a_{\ell m} a_{\ell' m'} \rangle = (-1)^{m'} \langle a_{\ell m} a^*_{\ell', -m'} \rangle = (-1)^m C_\ell \, \delta_{\ell\ell'} \delta_{m,-m'}$. In particular $\langle a^2_{\ell m} \rangle = 0$ whenever $m \neq 0$, a fact that fixes the distribution of the real and imaginary parts in Section 7.10.

Watch out: Isotropy is a property of the ensemble, not of our sky

Every single sky has directions that stand out: the $\ell \le 4$ panel of Figure 7.20 has a hot band and two cold blobs. Isotropy says that a rotated version of that sky was exactly as likely, not that the sky looks the same in all directions. The $a_{\ell m}$ of one sky are what they are; the statement (7.62) is about their averages over many hypothetical skies.

7.9.4 From C_ℓ back to the correlation function

The correlation function $C(\vartheta)$ of (7.51) and the power spectrum C_ℓ carry the same information: each is a Legendre series of the other. To show it we need the **addition theorem** for spherical harmonics, the sphere's version of $\sum_{\boldsymbol{k}} e^{i\boldsymbol{k}\cdot(\boldsymbol{x}-\boldsymbol{x}')}$ depending only on $\boldsymbol{x}-\boldsymbol{x}'$:

$$\sum_{m=-\ell}^{\ell} Y_{\ell m}(\boldsymbol{e})\, Y^*_{\ell m}(\boldsymbol{e}') = \frac{2\ell+1}{4\pi} P_\ell(\boldsymbol{e}\cdot\boldsymbol{e}'), \tag{7.63}$$

with P_ℓ the Legendre polynomial ($P_0=1$, $P_1(x)=x$, $P_2(x)=(3x^2-1)/2$). Check it for $\ell=1$ with both directions at the north pole: only $m=0$ survives at $\theta=0$, and $|Y_{10}|^2 = 3/4\pi = (3/4\pi)P_1(1)$. Then

$$\begin{aligned} C(\boldsymbol{e}\cdot\boldsymbol{e}') = \langle \Theta(\boldsymbol{e})\Theta(\boldsymbol{e}')\rangle &\overset{(1)}{=} \sum_{\ell m}\sum_{\ell' m'} \langle a_{\ell m} a^*_{\ell' m'}\rangle\, Y_{\ell m}(\boldsymbol{e})\, Y^*_{\ell' m'}(\boldsymbol{e}') \\ &\overset{(2)}{=} \sum_{\ell} C_\ell \sum_{m=-\ell}^{\ell} Y_{\ell m}(\boldsymbol{e})\, Y^*_{\ell m}(\boldsymbol{e}') \overset{(3)}{=} \sum_{\ell} \frac{2\ell+1}{4\pi} C_\ell\, P_\ell(\cos\vartheta). \end{aligned} \tag{7.64}$$

(1) Expand $\Theta(\boldsymbol{e})$ in $Y_{\ell m}$ and the real $\Theta(\boldsymbol{e}') = \Theta^*(\boldsymbol{e}')$ in $Y^*_{\ell' m'}$, and average term by term.
(2) Equation (7.62) keeps $\ell=\ell'$, $m=m'$.
(3) Addition theorem (7.63), with $\cos\vartheta = \boldsymbol{e}\cdot\boldsymbol{e}'$.

This is [HobsonBMC (10.4)]. To invert it, multiply by $P_{\ell'}(\cos\vartheta)$ and integrate over $x=\cos\vartheta$, using $\int_{-1}^{1} P_\ell P_{\ell'}\,\mathrm{d}x = 2\delta_{\ell\ell'}/(2\ell+1)$:

$$C_\ell = 2\pi \int_{-1}^{1} \mathrm{d}(\cos\vartheta)\; P_\ell(\cos\vartheta)\, C(\vartheta). \tag{7.65}$$

(Hobson et al. print the prefactor as 4π [HobsonBMC (10.5)]; the orthogonality integral above gives 2π.)[1]

[1] Insert (7.64) into the right-hand side: $2\pi \sum_{\ell'} \frac{2\ell'+1}{4\pi} C_{\ell'} \cdot \frac{2\delta_{\ell\ell'}}{2\ell+1} = C_\ell$. With 4π the right-hand side would be $2C_\ell$.

How much does each multipole add to the temperature variance at one point? At zero separation, $P_\ell(1) = 1$, and the correlation function is that variance,

$$\sigma_\Theta^2 = C(0) = \sum_\ell \frac{2\ell+1}{4\pi} C_\ell . \tag{7.66}$$

Each multipole contributes $(2\ell+1)C_\ell/4\pi$: the factor $2\ell+1$ counts the patterns at that ℓ and C_ℓ is the variance of each. Per logarithmic interval of ℓ the contribution is $\ell(2\ell+1)C_\ell/4\pi \approx \ell(\ell+1)C_\ell/2\pi = D_\ell$, which is why D_ℓ is the quantity that is plotted (Chapter T1).

The covariance of a map. In pixels the covariance is a full matrix; in harmonics it is diagonal. A map is the sky sampled at pixel centres $\boldsymbol{e}_p$, usually smoothed by the telescope beam. A circular beam multiplies each $a_{\ell m}$ by a number B_ℓ (its harmonic transform), so $C_\ell \to C_\ell B_\ell^2$; this is the sphere's version of smoothing in flat space (Section 7.6). The signal in two pixels p and p' then has the covariance

$$\boldsymbol{C}^S_{pp'} = \sum_\ell \frac{2\ell+1}{4\pi} C_\ell B_\ell^2 P_\ell(\cos\vartheta_{pp'}), \tag{7.67}$$

which is (7.64) with the beam included and $\vartheta_{pp'}$ the angle between the two pixel centres [HobsonBMC (10.18)]. This $N_{\rm pix} \times N_{\rm pix}$ matrix is the signal covariance of a Gaussian likelihood written in pixels. Every entry is nonzero, since any two pixels are correlated. In harmonic space the same covariance (7.62) is diagonal. That is why we work with the $a_{\ell m}$.

Example: the rms of the CMB. For the fiducial spectrum (Section 7.11.1), `code/ch07/22_isotropy.py` evaluates (7.66) up to $\ell = 2000$: $\sigma_\Theta^2 = 12635\,\mu\mathrm{K}^2$, an rms of $112.4\,\mu$K, the "hundred microkelvin" of the opening paragraph. The multipoles $\ell \le 30$ supply 19.9% of this variance, although they are only about 10^3 of the four million coefficients up to $\ell = 2000$. Figure 7.21(a) shows how fast the correlation falls: by a separation of one degree, $C(\vartheta)$ has dropped to about a quarter of its value at zero, which is the statement that most of the variance sits in features smaller than about a degree, the scale of the first acoustic peak.

Figure 7.21(b) makes a point that matters for statistics: the correlation function measured on one sky can differ from the ensemble prediction by as much as the prediction itself, at large separations. What we measure on one sky is the average of $\Theta(\boldsymbol{e})\Theta(\boldsymbol{e}')$ over all pairs of directions at separation ϑ. This equals the same Legendre series (7.64) with C_ℓ replaced by that sky's own average over m, $\widehat{C}_\ell = \sum_m |a_{\ell m}|^2/(2\ell+1)$. (Averaging the pair over all orientations does to the product of harmonics what the ensemble average did in step (2) of (7.64).) Five simulated skies give five visibly different curves at large separations. They scatter around the ensemble prediction by 80 to $180\,\mu\mathrm{K}^2$ at 30° to 90°, which is as large as the signal itself there ($|C| \approx 160$ to $230\,\mu\mathrm{K}^2$). This is cosmic variance seen in real space. It is large at large ϑ because large separations are controlled by the few low multipoles, and a low multipole has few coefficients to average over. We quantify it in Section 7.10.

7.9.5 Seeing isotropy, and its violation, in simulated skies

The argument of Section 7.9.3 makes two predictions we can test: within one ℓ, the average power $\langle |a_{\ell m}|^2 \rangle$ is the same for every m; and different ℓ are uncorrelated. A test is convincing only if it can fail, so we also build a sky that is *not* isotropic and check that both predictions break.

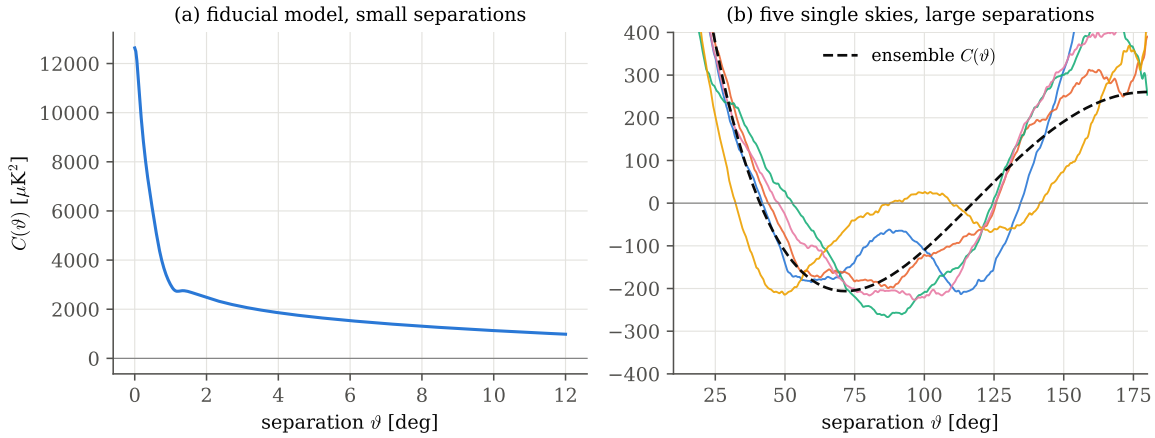


Figure 7.21: (a) The angular correlation function of the fiducial model for separations up to 12°. It falls from the point variance $C(0) \approx 1.3\times10^4\,\mu\mathrm{K}^2$ to a quarter of that within one degree, then decays slowly. (b) Large separations: the ensemble $C(\vartheta)$ (black dashed) and the full-sky pair average for five single simulated skies (coloured). The five skies are drawn from the same model; their differences are cosmic variance.

Example: a sky with a preferred axis. Multiply an isotropic sky by $1 + A\cos\theta$. Fluctuations are then amplified by a factor $1 + A$ at the north pole and damped by $1 - A$ at the south pole: the z-axis is special. (A small effect of this kind, a "hemispherical power asymmetry", is one of the anomalies that the Planck collaboration tests for [114].) What does the modulation do to the coefficients? In harmonic space, multiplying by $\cos\theta$ is simple, because $\cos\theta \propto Y_{10}$ carries one unit of angular momentum along z. It changes ℓ by ±1 and keeps m, the dipole selection rule of atomic physics:

$$\cos\theta\, Y_{\ell m} = c_{\ell+1,m}\, Y_{\ell+1,m} + c_{\ell m}\, Y_{\ell-1,m}, \qquad c_{\ell m} = \sqrt{\frac{\ell^2 - m^2}{(2\ell+1)(2\ell-1)}}. \tag{7.68}$$

Check it on $\ell = 0$: $\cos\theta\, Y_{00} = \cos\theta/\sqrt{4\pi} = \sqrt{1/3}\, Y_{10}$, and $c_{10} = \sqrt{1/3}$. The coefficients of the modulated sky $\Theta' = (1 + A\cos\theta)\Theta$ are

$$a'_{LM} \overset{(1)}{=} a_{LM} + A\sum_{\ell m} a_{\ell m} \int \mathrm{d}\Omega\, \cos\theta\, Y_{\ell m} Y^*_{LM} \overset{(2)}{=} a_{LM} + A\big(c_{LM}\, a_{L-1,M} + c_{L+1,M}\, a_{L+1,M}\big). \tag{7.69}$$

(1) Definition (7.44) for Θ', with Θ expanded in harmonics.
(2) Equation (7.68) and orthonormality: the term $Y_{\ell+1,m}$ contributes when $\ell + 1 = L$, the term $Y_{\ell-1,m}$ when $\ell - 1 = L$, and only for $m = M$.

Each new coefficient is a mix of three old ones, which are independent. Averaging products with (7.62) for the old sky gives

$$\langle a'_{\ell+1,m}\, a'^*_{\ell m}\rangle = A\, c_{\ell+1,m}\,(C_\ell + C_{\ell+1}), \tag{7.70}$$

$$\langle |a'_{\ell m}|^2\rangle = C_\ell + A^2\big(c^2_{\ell m}\, C_{\ell-1} + c^2_{\ell+1,m}\, C_{\ell+1}\big). \tag{7.71}$$

Both predictions of isotropy now fail. By (7.70), neighbouring multipoles are correlated. By (7.71), the power depends on m through $c^2_{\ell m}$ and $c^2_{\ell+1,m}$, which are both largest at $m = 0$ and shrink towards $|m| = \ell$ (where $c_{\ell m}$ vanishes). Both violations come from the preferred axis.

The script `code/ch07/22_isotropy.py` draws 6000 isotropic skies up to $\ell = 42$ from the fiducial spectrum, modulates each one with $A = 0.5$ using (7.69), and averages. The modulation is done exactly in harmonic space by the function `modulate_dipole` of the small library `code/ch07/lib_harmonic.py`:

Listing 7.3: Multiplying a sky by $1+A\cos\theta$, coefficient by coefficient (`code/ch07/lib_harmonic.py`)

```
84 def modulate_dipole(alm, lmax: int, A: float):
85     """a_lm of (1 + A cos theta) * map, computed exactly in harmonic space (result
       truncated at lmax-1).
86
87     a'_LM = a_LM + A [ c_{L,M} a_{L-1,M} + c_{L+1,M} a_{L+1,M} ].
88     """
89     ell, m = getlm(lmax)
90     out = alm.copy()
91     down = ell - 1 >= m                    # a_{L-1,M} exists
92     up = ell + 1 <= lmax                   # a_{L+1,M} exists
93     i = np.arange(ell.size)
94     out[down] += A * ccoef(ell[down], m[down]) * alm[i[down] - 1]
95     out[up] += A * ccoef(ell[up] + 1, m[up]) * alm[i[up] + 1]
96     return out
```

Lines 91–92 find which neighbours $a_{L\mp1,M}$ exist (the stored list starts at $\ell = m$ and stops at $\ell_{\max}$); lines 94–95 add the two terms of (7.69). In the storage order of healpy, which the library uses (Section 7.11.2), $a_{L-1,M}$ and $a_{L+1,M}$ sit just before and just after a_{LM}.

Figure 7.22 shows the result at $\ell = 20$. For the isotropic skies the power is the same in every m: the largest deviation of $\langle|a_{20,m}|^2\rangle/C_{20}$ from 1 among $m = 1,\dots,20$ is 0.02503, about two standard errors of 0.01291 (since $|a_{\ell m}|^2/C_\ell$ has standard deviation 1 for $m > 0$, as we will see), which is what the largest of twenty independent deviations typically is. The correlation between $\ell = 20$ and $\ell = 21$ is zero within 0.01502. For the modulated skies both predictions fail exactly as (7.70) and (7.71) say: the power rises to $1.126\,C_{20}$ at $m = 0$, and the neighbour correlation reaches $0.5005\,C_{20}$ at $m = 0$ and falls towards $|m| = \ell$.

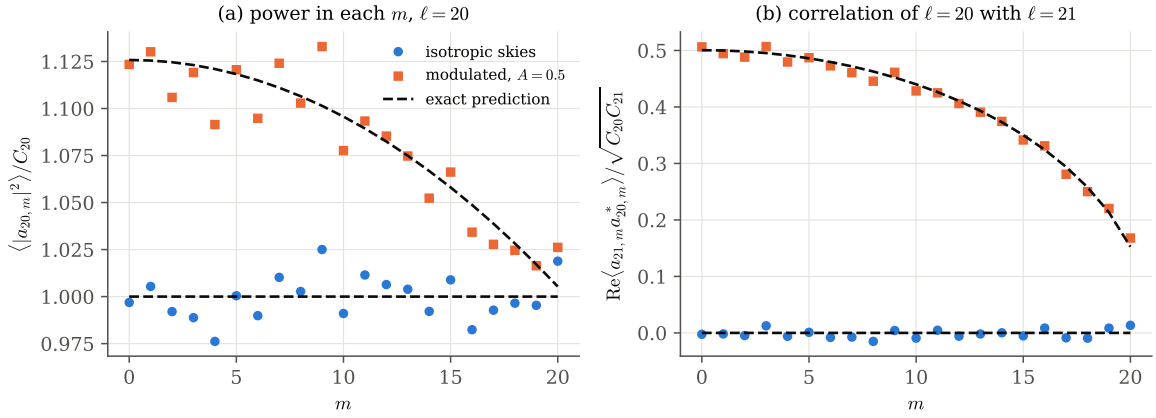


Figure 7.22: Two predictions of isotropy, tested on 6000 simulated skies at $\ell = 20$. (a) Average power in each m, in units of C_{20}: flat for isotropic skies (blue), tilted for skies modulated by $1 + 0.5\cos\theta$ (orange). (b) Average product of the coefficients at $\ell = 21$ and $\ell = 20$ with the same m: zero for isotropic skies, large for modulated ones. Black dashed lines are the exact predictions (7.70) and (7.71) (and 1, 0 for the isotropic case).

Gaussianity turns "uncorrelated" into "independent". So far isotropy has told us which coefficients are uncorrelated. For a Gaussian sky that is enough to make them independent. The coefficients are integrals of the sky with fixed weights, so they are linear combinations of the field values. If the field is Gaussian (as linear evolution of Gaussian initial conditions guarantees, [HobsonBMC p. 230]), the coefficients are jointly Gaussian, and for jointly Gaussian variables zero correlation means independence (Chapter 2). For complex coefficients we must be careful about what "zero correlation" means, since a complex number is two real numbers; Section 7.10 does

that bookkeeping and turns (7.62) into the exact distribution of the power-spectrum estimator.

7.10 One sky, $2\ell+1$ numbers: the estimator $\widehat{C}_\ell$ and cosmic variance

The model predicts $C_\ell = \langle |a_{\ell m}|^2 \rangle$, an average over an ensemble of skies we will never see. We have one sky. What saves us is the result of Section 7.9: in an isotropic sky the coefficients at one multipole are uncorrelated and all have the same variance C_ℓ (7.62). So at each multipole the sky hands us $2\ell+1$ real numbers (the count (7.47)) with a common variance. That is the situation of a laboratory physicist who has measured the same noisy quantity $2\ell+1$ times and wants its variance: average the squares (the mean is known to be zero). Replacing the ensemble average by this average over the sphere is called the ergodic replacement. Its weakness is at low ℓ, where there are very few numbers to average. That limit comes from the Universe, not from the instrument.

The question. If we estimate C_ℓ by averaging $|a_{\ell m}|^2$ over the $2\ell+1$ values of m, is the estimate right on average, how is it distributed, and how large is its scatter around the truth?

Working definition 7.7 (full-sky power-spectrum estimator)

For a full, noiseless sky,

$$\widehat{C}_\ell = \frac{1}{2\ell+1}\sum_{m=-\ell}^{\ell} |a_{\ell m}|^2 = \frac{1}{2\ell+1}\left(a_{\ell 0}^2 + 2\sum_{m=1}^{\ell} |a_{\ell m}|^2\right) \tag{7.72}$$

[the mean square of the $2\ell+1$ coefficients at one ℓ, each $m>0$ counted twice because $|a_{\ell,-m}| = |a_{\ell m}|$] [Hivon (8)], [HobsonBMC (10.25)]. healpy's `alm2cl` computes exactly this from the stored $m \geq 0$ coefficients.

The second form uses the reality condition (7.46), $|a_{\ell,-m}| = |a_{\ell m}|$. Its expectation follows in one line from (7.62):

$$\mathbb{E}\,\widehat{C}_\ell = \frac{1}{2\ell+1}\sum_{m=-\ell}^{\ell} \langle |a_{\ell m}|^2 \rangle = \frac{1}{2\ell+1}\,(2\ell+1)\,C_\ell = C_\ell. \tag{7.73}$$

So $\widehat{C}_\ell$ is unbiased. Here is the estimator written out from scratch, as every simulation script in this chapter uses it:

Listing 7.4: The full-sky estimator (7.72) from the stored coefficients (`code/ch07/lib_harmonic.py`)

```
48 def alm2cl(alm, lmax: int):
49     """Full-sky estimator C_hat_l = (1/(2l+1)) sum_{m=-l}^{l} |a_lm|^2.
50
51     The sum over negative m is folded in: |a_{l,-m}| = |a_lm|, so each m > 0 counts twice.
52     """
53     ell, m = getlm(lmax)
54     w = np.where(m == 0, 1.0, 2.0)
55     s = np.bincount(ell, weights=w * np.abs(alm) ** 2, minlength=lmax + 1)
56     return s / (2 * np.arange(lmax + 1) + 1)
```

Line 54 gives weight 1 to $m=0$ and weight 2 to $m>0$, which folds in the negative m that are not stored. Line 55 adds up the weighted $|a_{\ell m}|^2$ for each ℓ with `np.bincount`, and line 56 divides by $2\ell+1$. Applied to the same coefficients, it agrees with healpy's `alm2cl` to a relative 6.474×10^{-16}, rounding error.

7.10.1 Counting real degrees of freedom

$\widehat{C}_\ell$ is a sum of squares. To find its distribution we need the distribution of the real numbers being squared, the real and imaginary parts of the $a_{\ell m}$. Fix ℓ and write $a_{\ell m} = x_m + iy_m$ for $m = 1, \dots, \ell$, with x_m and y_m real. Section 7.9 gave us two averages: $\langle a_{\ell m} a^*_{\ell m}\rangle = C_\ell$, and $\langle a^2_{\ell m}\rangle = 0$ for $m \neq 0$. Expand both in real and imaginary parts:

$$C_\ell \overset{(1)}{=} \langle |a_{\ell m}|^2\rangle = \langle x_m^2\rangle + \langle y_m^2\rangle,$$
$$0 \overset{(2)}{=} \langle a^2_{\ell m}\rangle = \langle x_m^2\rangle - \langle y_m^2\rangle + 2i\,\langle x_m y_m\rangle \quad\Longrightarrow\quad \langle x_m^2\rangle = \langle y_m^2\rangle = \frac{C_\ell}{2}, \qquad \langle x_m y_m\rangle = 0. \tag{7.74}$$

(1) Equation (7.62) with $\ell' = \ell$, $m' = m$, and $|x+iy|^2 = x^2 + y^2$.

(2) $\langle a_{\ell m} a_{\ell' m'}\rangle = (-1)^m C_\ell \delta_{\ell\ell'}\delta_{m,-m'}$ vanishes for $m' = m \neq 0$. The real part of the zero gives $\langle x^2\rangle = \langle y^2\rangle$, the imaginary part gives $\langle xy\rangle = 0$; with the first line, each is $C_\ell/2$.

What about two different coefficients, $m \neq m'$ with both positive? Both averages vanish: $\langle a_{\ell m} a^*_{\ell m'}\rangle = 0$ because $m \neq m'$, and $\langle a_{\ell m} a_{\ell m'}\rangle = 0$ because $m' \neq -m$. Taking the real and imaginary parts of these two zeros makes all four products $\langle x_m x_{m'}\rangle$, $\langle y_m y_{m'}\rangle$, $\langle x_m y_{m'}\rangle$, $\langle y_m x_{m'}\rangle$ zero. The same argument with $m' = 0$ shows that $a_{\ell 0}$ is uncorrelated with every x_m and y_m. Finally, $a_{\ell 0}$ is real with $\langle a^2_{\ell 0}\rangle = C_\ell$. In summary, at each ℓ:

real number	how many	variance
$a_{\ell 0}$	1	C_ℓ
$x_m = \mathrm{Re}\, a_{\ell m}$, $m = 1, \dots, \ell$	ℓ	$C_\ell/2$
$y_m = \mathrm{Im}\, a_{\ell m}$, $m = 1, \dots, \ell$	ℓ	$C_\ell/2$
$a_{\ell m}$ with $m < 0$	–	fixed by $a_{\ell,-m} = (-1)^m a^*_{\ell m}$

For a Gaussian sky these $2\ell+1$ real numbers are jointly Gaussian and uncorrelated, hence independent (Chapter 2). Figure 7.23 draws the bookkeeping for $\ell = 3$.

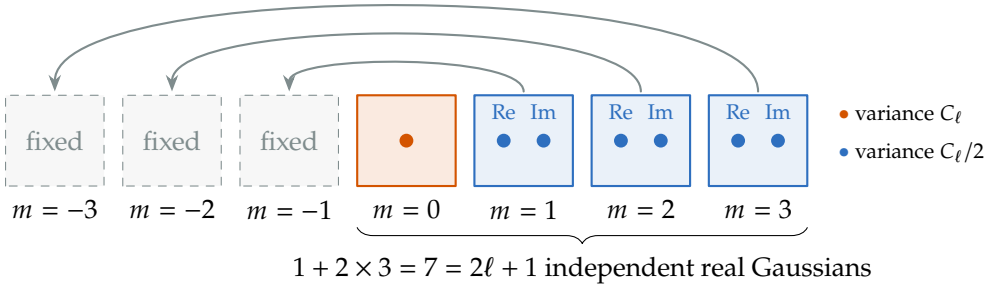


Figure 7.23: The real degrees of freedom at $\ell = 3$. The coefficient with $m = 0$ is one real number of variance C_ℓ (orange). Each $m > 0$ coefficient is two real numbers, its real and imaginary parts, each of variance $C_\ell/2$ (blue). The grey arcs send each $m > 0$ cell to its partner $-m$, which the reality condition fixes completely, so the negative-m cells add nothing new.

Now standardise each real number to unit variance, $z_0 = a_{\ell 0}/\sqrt{C_\ell}$, $z^{\mathrm{R}}_m = x_m/\sqrt{C_\ell/2}$, $z^{\mathrm{I}}_m = y_m/\sqrt{C_\ell/2}$, all independent $\mathcal{N}(0,1)$, and rewrite the estimator:

$$\frac{(2\ell+1)\widehat{C}_\ell}{C_\ell} \overset{(1)}{=} \frac{1}{C_\ell}\left(a^2_{\ell 0} + 2\sum_{m=1}^{\ell}(x_m^2 + y_m^2)\right) \overset{(2)}{=} z_0^2 + \frac{2}{C_\ell}\frac{C_\ell}{2}\sum_{m=1}^{\ell}\left[(z^{\mathrm{R}}_m)^2 + (z^{\mathrm{I}}_m)^2\right]$$
$$\overset{(3)}{=} z_0^2 + \sum_{m=1}^{\ell}\left[(z^{\mathrm{R}}_m)^2 + (z^{\mathrm{I}}_m)^2\right] \;\sim\; \chi^2_{2\ell+1}. \tag{7.75}$$

(1) Equation (7.72) multiplied by $(2\ell+1)/C_\ell$, with $|a_{\ell m}|^2 = x_m^2 + y_m^2$.
(2) Substitute $a_{\ell 0} = \sqrt{C_\ell}\, z_0$, $x_m = \sqrt{C_\ell/2}\, z_m^{\mathrm{R}}$, $y_m = \sqrt{C_\ell/2}\, z_m^{\mathrm{I}}$.
(3) The factor 2 from the folded negative m cancels the $1/2$ in the variance. What remains is a sum of $1 + 2\ell$ squares of independent standard normals, which is the definition of a χ^2 variable with $2\ell+1$ degrees of freedom (Chapter 2).

Step (3) is why the careful count matters: the factor 2 from folding and the halved variance cancel exactly, so every real degree of freedom enters with weight one.

Watch out: Two ways to get the count wrong

(i) Treating all $2\ell+1$ coefficients as independent complex Gaussians gives $2(2\ell+1)$ real degrees of freedom and a cosmic variance half the true value: it forgets that $a_{\ell,-m}$ is a copy of $a_{\ell m}$. (ii) Giving the real and imaginary parts of an $m > 0$ coefficient the variance C_ℓ each doubles $\langle |a_{\ell m}|^2 \rangle$. The one-point density $p(a_{\ell m}) = e^{-a_{\ell m}^2/2C_\ell}/\sqrt{2\pi C_\ell}$ of and [HobsonBMC (10.17)] is the density of a *real* variable of variance C_ℓ: exact for $a_{\ell 0}$; for $m > 0$ it has to be read as the joint density of x_m and y_m, each of variance $C_\ell/2$. Either way the count is $2\ell+1$, and so is the χ^2.

7.10.2 The distribution of $\widehat{C}_\ell$ and cosmic variance

We now know that $\widehat{C}_\ell$ is a rescaled χ^2 variable. Its density and its scatter follow from those of χ^2. Write $\nu = 2\ell+1$ and $V = \nu\widehat{C}_\ell/C_\ell \sim \chi^2_\nu$, whose density is $f_\nu(v) = v^{\nu/2-1}e^{-v/2}/\left(2^{\nu/2}\Gamma(\nu/2)\right)$ for $v > 0$. Since $\widehat{C}_\ell = C_\ell V/\nu$ is a rescaling, the change-of-variables rule gives the density of the estimator,

$$p\left(\widehat{C}_\ell \,|\, C_\ell\right) = \frac{\nu}{C_\ell} f_\nu\left(\frac{\nu\widehat{C}_\ell}{C_\ell}\right), \tag{7.76}$$

a *scaled* χ^2_ν. Its first two moments follow from those of χ^2_ν, which is a Gamma$(\nu/2, 2)$ variable with mean ν and variance 2ν (Chapter 2):

$$\operatorname{Var}\widehat{C}_\ell \overset{(1)}{=} \left(\frac{C_\ell}{\nu}\right)^2 \operatorname{Var} V \overset{(2)}{=} \frac{C_\ell^2}{\nu^2}\, 2\nu = \frac{2C_\ell^2}{2\ell+1}. \tag{7.77}$$

(1) $\operatorname{Var}(cV) = c^2 \operatorname{Var} V$ with $c = C_\ell/\nu$.
(2) $\operatorname{Var}\chi^2_\nu = 2\nu$.

Key idea: Cosmic variance

From one full, noiseless sky the power spectrum is measured with

$$\frac{\Delta C_\ell}{C_\ell} \equiv \frac{\sqrt{\operatorname{Var}\widehat{C}_\ell}}{C_\ell} = \sqrt{\frac{2}{2\ell+1}}, \tag{7.78}$$

the sampling error of a variance estimated from $2\ell+1$ Gaussian numbers. No instrument can reduce it; only more independent modes can. Chapter 3 showed (in the section that finds an estimator from a model) that $\widehat{C}_\ell$ is sufficient for C_ℓ, is its maximum-likelihood estimator, and attains the Cramér–Rao bound, which equals (7.77): no unbiased estimator built from one sky does better.

Hivon et al. state the result in one sentence, "C_ℓ-s are χ^2_ν-distributed with the mean equal to

$C_\ell^{\rm th}$, $\nu = 2\ell+1$ degrees of freedom, and a variance of $2C_\ell^{\rm th\,2}/\nu$" [Hivon p. 3].[2] Knox gives the same number as the noiseless limit of his error formula [166, eq. 4], with the same reading: for each ℓ there are $2\ell+1$ "samples" from a Gaussian of variance C_ℓ, and the sampling variance of a variance is twice its square divided by the number of samples. On the full sky this sampling variance is called **cosmic variance**.

The shape matters as much as the width. A χ^2_ν has skewness $\sqrt{8/\nu}$ and its mode sits at $\nu-2$, below the mean ν. So at low ℓ the estimator is more often below C_ℓ than above it, and occasionally far above. Figure 7.24(a) shows the density (7.76) in units of C_ℓ for $\ell = 2, 5, 20$ and 100. At $\ell = 2$ the density starts at zero, peaks at $0.6\,C_\ell$, and has a long tail; by $\ell = 100$ it is a narrow, nearly symmetric bell, close to the Gaussian of the same mean and variance (dotted lines, shown for $\ell = 5$ and $\ell = 100$).

Example: the quadrupole. At $\ell = 2$, $\nu = 5$, and (7.78) gives a fractional error of 0.6325: one sky determines the quadrupole only to 63%. The distribution is skewed (skewness 1.265), and the script `code/ch07/23_cl_distribution.py` evaluates $\mathrm{P}(\widehat{C}_2 < C_2) = \mathrm{P}(\chi^2_5 < 5) = 0.5841$: a low quadrupole is more likely than a high one. Suppose the measured quadrupole were a fifth of the model's. In the language of Chapter 4: if the model were true, would getting a quadrupole this low, or even lower, be something we'd reasonably expect? The probability is $\mathrm{P}(\widehat{C}_2 < 0.2\,C_2) = \mathrm{P}(\chi^2_5 < 1) = 0.03743$: small but not negligible. It is even less alarming than it looks, because the quadrupole was singled out for testing only after the sky had been seen; with many multipoles to look at, one of them is likely to look odd. A low quadrupole in the real sky is one of the large-angle features examined in the Planck isotropy analysis [114], and this χ^2_5 tail is the first thing any such test computes.

Example: $\ell = 30$ and $\ell = 1000$. At $\ell = 30$, $\nu = 61$ and the fractional error is 0.1811; still, a 20% low fluctuation, $\mathrm{P}(\widehat{C}_{30} < 0.8\,C_{30}) = 0.1299$, happens in one sky in eight. At $\ell = 1000$ the error is 0.03161, about 3%. Since different multipoles are independent on the full sky (Section 7.12.4), averaging $\Delta\ell$ neighbouring multipoles with nearly equal C_ℓ divides the variance by $\Delta\ell$:

$$\mathrm{Var}\Big(\frac{1}{\Delta\ell}\sum_{\ell'\in\mathrm{bin}}\widehat{C}_{\ell'}\Big) = \frac{1}{\Delta\ell^2}\sum_{\ell'\in\mathrm{bin}}\frac{2C_{\ell'}^2}{2\ell'+1} \approx \frac{2C_\ell^2}{(2\ell+1)\,\Delta\ell}. \tag{7.79}$$

A bin of $\Delta\ell = 50$ at $\ell = 1000$ is measured to 0.4471%. Real experiments see only part of the sky, with noise N_ℓ and a beam B_ℓ, and the rule of thumb becomes $\Delta C_\ell \approx (C_\ell + N_\ell/B_\ell^2)\sqrt{2/[(2\ell+1)\Delta\ell\, f_{\rm sky}]}$ [166, eq. 4], [Hivon (16)–(17)]: noise adds to the signal inside the bracket, and observing a fraction $f_{\rm sky}$ of the sky leaves about a fraction $f_{\rm sky}$ of the modes. Chapter 8 derives this and shows where it fails.

7.10.3 Reading the distribution backwards: which C_ℓ made our sky?

Everything so far was prediction: given C_ℓ, what values of $\widehat{C}_\ell$ come out, and how often. The physicist's problem runs the other way. We observe one sky; what does it tell us about C_ℓ?

What we observed. Our sky's estimate at multipole ℓ, $\widehat{C}_\ell = \hat{c}$, a single number in $\mu\mathrm{K}^2$. *What could have produced it?* Any value of the true C_ℓ, each one a different cosmological model at that multipole; cosmic variance lets every one of them give $\hat{c}$, some far more readily than others. *How probable was $\hat{c}$ under each candidate?* That is the density (7.76) evaluated at the observed $\hat{c}$, once for each candidate C_ℓ. *Only then the prior.* The update multiplies these numbers by prior

[2] Read literally, a χ^2_ν variable has mean ν and variance 2ν. What is meant is the scaled χ^2 of (7.76): $\nu\widehat{C}_\ell/C_\ell \sim \chi^2_\nu$.

weights and normalises (Chapter 5). *Only then the numbers.* The third step is the likelihood. Read (7.76) as a function of C_ℓ with $\hat{c}$ held fixed:

$$\mathcal{L}(C_\ell) = p(\hat{c}\,|\,C_\ell) \overset{(1)}{=} \frac{\nu}{C_\ell}\,\frac{(\nu\hat{c}/C_\ell)^{\nu/2-1}e^{-\nu\hat{c}/2C_\ell}}{2^{\nu/2}\Gamma(\nu/2)} \overset{(2)}{=} \text{const}\times C_\ell^{-\nu/2}\exp\!\Big(-\frac{\nu\hat{c}}{2C_\ell}\Big). \tag{7.80}$$

(1) Insert the χ^2_ν density into (7.76) at $\widehat{C}_\ell = \hat{c}$.

(2) Collect the powers of C_ℓ: $C_\ell^{-1}\cdot C_\ell^{-(\nu/2-1)} = C_\ell^{-\nu/2}$; the factors that depend only on $\hat{c}$ and ν do not distinguish hypotheses and go into the constant.

This is the full-sky CMB likelihood of Chapter 3; its maximum is at $C_\ell = \hat{c}$, the estimator itself. Figure 7.24(b) plots it against the candidate C_ℓ in units of the observed $\hat{c}$. At $\ell = 2$ it is very broad and lopsided: candidate spectra up to $2.03\,\hat{c}$ and down to $0.56\,\hat{c}$ lie within $\Delta(-2\ln\mathcal{L}) = 1$ of the peak. The interval extends much further upwards because a large true C_ℓ produces a low estimate fairly often (the long left part of the χ^2_5 density), while a small true C_ℓ almost never produces a high one. At $\ell = 100$ the interval $[0.91, 1.11]\,\hat{c}$ is narrow and nearly symmetric. Chapter 9 runs the same reasoning over six cosmological parameters at once, with $C_\ell(\boldsymbol{\theta})$ supplied by CAMB.

Example: adding white noise. A detector adds its own noise to the sky: $\hat{a}_{\ell m} = a_{\ell m} + n_{\ell m}$, with $n_{\ell m}$ the coefficients of a map of uniform white noise of pixel variance σ^2 and pixel area Ω_{pix}. The question is what this does to the estimator and the likelihood. First, the noise coefficients. They are uncorrelated with each other and with the signal, and each has the variance $N_\ell = \sigma^2\Omega_{\text{pix}}$. The reason: $n_{\ell m} = \sum_p n_p Y^*_{\ell m}(p)\,\Omega_{\text{pix}}$ is a sum of independent pixel noises, so its variance is $\sigma^2\Omega_{\text{pix}}^2\sum_p |Y_{\ell m}(p)|^2$, and $\sum_p |Y_{\ell m}(p)|^2\Omega_{\text{pix}} \approx \int \mathrm{d}\Omega\,|Y_{\ell m}|^2 = 1$. Second, signal plus noise is a sum of two independent Gaussians, so it has the covariance (7.62) with $C_\ell + N_\ell$ in place of C_ℓ. The whole argument of this section therefore holds with that replacement: $(2\ell+1)\hat{C}_\ell^{\text{obs}}/(C_\ell + N_\ell) \sim \chi^2_{2\ell+1}$, where $\hat{C}_\ell^{\text{obs}}$ is (7.72) computed from the $\hat{a}_{\ell m}$. The likelihood (7.80) becomes

$$\mathcal{L}(C_\ell) \propto (C_\ell + N_\ell)^{-(2\ell+1)/2}\exp\!\Big[-\frac{(2\ell+1)\,\hat{C}_\ell^{\text{obs}}}{2(C_\ell + N_\ell)}\Big], \tag{7.81}$$

which, by the same argument as before, is largest when $C_\ell + N_\ell = \hat{C}_\ell^{\text{obs}}$:

$$\hat{C}_\ell^{\text{ML}} = \hat{C}_\ell^{\text{obs}} - N_\ell \tag{7.82}$$

[HobsonBMC (10.24)–(10.25)]. The noise is subtracted on average, but the scatter is that of $C_\ell + N_\ell$, which is why N_ℓ enters the bracket of the error formula of Section 7.10.2.

7.11 Where C_ℓ and the skies come from: CAMB and HEALPix

Up to now C_ℓ has been a given function and a sky has been a given set of $a_{\ell m}$. In a real analysis both have to be produced. The spectrum comes from physics: six cosmological numbers go into a Boltzmann code, and C_ℓ comes out (Chapter T1). The skies come from geometry: we cut the sphere into pixels and use fast spherical-harmonic transforms to turn coefficients into maps and maps back into coefficients. Think of the chain physical model → field or dataset → summary → sampling distribution → inference. CAMB supplies the first arrow. HEALPix lets us make the second arrow on a computer, as many times as we like.

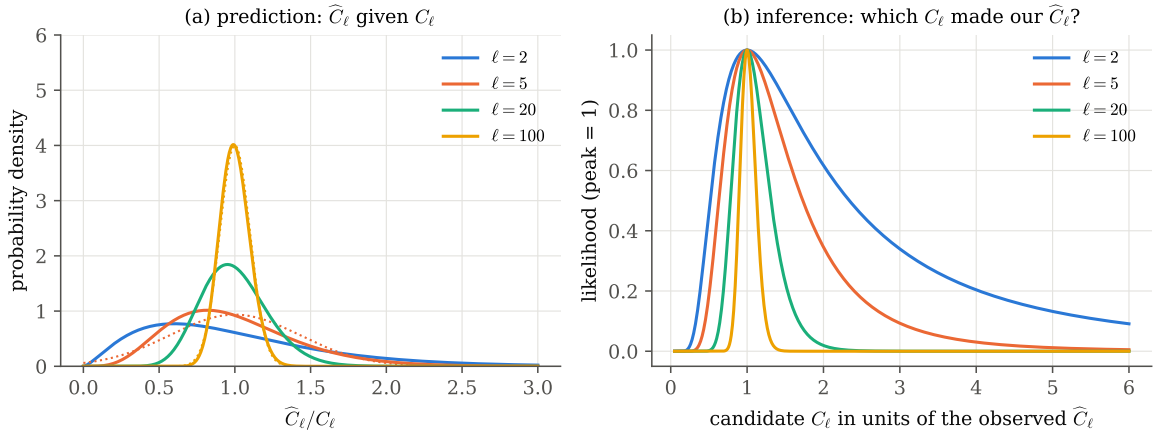


Figure 7.24: The two directions of the same formula. (a) Prediction: the density of $\widehat{C}_\ell/C_\ell$ for $\ell = 2, 5, 20, 100$, a scaled $\chi^2_{2\ell+1}$; dotted curves are Gaussians with the same mean and variance, for $\ell = 5$ and $\ell = 100$. At low ℓ the estimator is skewed, with its mode below the truth. (b) Inference: the likelihood of the true C_ℓ given an observed $\widehat{C}_\ell$, normalised to peak 1. All curves peak at the observed value; at low ℓ the likelihood has a long tail towards large C_ℓ.

7.11.1 What CAMB computes

The input. Follow one Fourier mode $\boldsymbol{k}$ of the primordial curvature perturbation ζ, the seed set up by inflation. Through gravity it perturbs every component of the Universe: photons, baryons, cold dark matter, neutrinos and the metric.

What is followed. The photon temperature perturbation in that mode depends on the direction in which the photon travels, measured relative to $\boldsymbol{k}$. It is therefore expanded in Legendre multipoles $\Theta_\ell(k, \eta)$, with η the conformal time. The equations for these multipoles form the **Boltzmann hierarchy**. For free-streaming particles it reads

$$\Theta'_\ell = \frac{k}{2\ell+1}\left[\ell\,\Theta_{\ell-1} - (\ell+1)\,\Theta_{\ell+1}\right] \qquad (\ell \geq 2), \tag{7.83}$$

where the prime is $\partial/\partial\eta$. Streaming therefore moves anisotropy from each multipole into its neighbours, and so from large angles to small. Photons obey the same equation plus two extra terms: a Thomson scattering term, which damps every $\ell \geq 2$ multipole while the plasma is ionised, and gravitational source terms at $\ell \leq 2$.

What the code does. A Boltzmann code integrates these coupled linear ODEs, together with the fluid equations for baryons and dark matter and the linearised Einstein equations, from deep in the radiation era to today, for a few thousand values of k. Integrating the hierarchy directly up to $\ell \sim 3000$ would be slow. The trick that made the codes fast is to write today's multipoles as an integral along the line of sight, $\Theta_\ell(k, \eta_0) = \int_0^{\eta_0} \mathrm{d}\eta\, S(k, \eta)\, j_\ell\big(k(\eta_0 - \eta)\big)$, where the source S needs only the first few multipoles and the spherical Bessel function j_ℓ is pure geometry, independent of the cosmology [167]. CAMB implements this method [8]; so do CMBFAST and CLASS, and all of them compute the linear CMB spectra by integrating the sources along the photon line of sight [Padilla §7.1.1].

For statistics, what matters is the structure of the answer, and it is simple. The equations are linear, so each $\boldsymbol{k}$ mode evolves on its own, and today's temperature in that mode is a fixed number times its primordial amplitude $\zeta_{\boldsymbol{k}}$. That number, for multipole ℓ, is the radiation **transfer function** $g_{T\ell}(k)$ [the temperature anisotropy at multipole ℓ produced by one unit of ζ

at wavenumber k, computed once per cosmology]. Projected on the sky [167],

$$a_{\ell m} = 4\pi\, i^{\ell} \int \frac{\mathrm{d}^3k}{(2\pi)^3}\; g_{T\ell}(k)\, Y^*_{\ell m}(\hat{\boldsymbol{k}})\, \zeta_{\boldsymbol{k}}. \tag{7.84}$$

With isotropic primordial statistics, $\langle \zeta_{\boldsymbol{k}} \zeta^*_{\boldsymbol{k}'} \rangle = (2\pi)^3 \delta^{(3)}(\boldsymbol{k} - \boldsymbol{k}') P_\zeta(k)$, the covariance of the coefficients is

$$\begin{aligned}
\langle a_{\ell m} a^*_{\ell' m'} \rangle &\overset{(1)}{=} (4\pi)^2\, i^{\ell - \ell'} \int \frac{\mathrm{d}^3k}{(2\pi)^3}\; g_{T\ell}(k)\, g_{T\ell'}(k)\, P_\zeta(k)\, Y^*_{\ell m}(\hat{\boldsymbol{k}})\, Y_{\ell' m'}(\hat{\boldsymbol{k}}) \\
&\overset{(2)}{=} \frac{16\pi^2}{8\pi^3} \int_0^\infty k^2 \mathrm{d}k\; g_{T\ell}\, g_{T\ell'}\, P_\zeta(k) \int \mathrm{d}\Omega_{\hat{\boldsymbol{k}}}\; Y^*_{\ell m} Y_{\ell' m'} \\
&\overset{(3)}{=} \underbrace{\frac{2}{\pi} \int_0^\infty k^2 \mathrm{d}k\; P_\zeta(k)\, g_{T\ell}(k)^2}_{C_\ell}\; \delta_{\ell\ell'} \delta_{mm'}.
\end{aligned} \tag{7.85}$$

(1) Multiply (7.84) by the conjugate of its primed copy, average, and do the $\boldsymbol{k}'$ integral with the delta function; $(i^{\ell'})^* = i^{-\ell'}$.

(2) Split $\mathrm{d}^3k = k^2 \mathrm{d}k\, \mathrm{d}\Omega_{\hat{\boldsymbol{k}}}$: the transfer functions and P_ζ depend only on $k = |\boldsymbol{k}|$, the harmonics only on the direction $\hat{\boldsymbol{k}}$.

(3) Orthonormality (7.41) over the directions of $\boldsymbol{k}$ sets $\ell = \ell'$, $m = m'$, which also makes $i^{\ell - \ell'} = 1$.

This is the formula for C_ℓ in Chapter T1. It is also the isotropy result (7.62), reached this time from the physics side. The Kronecker deltas came from integrating the harmonics over the directions of $\boldsymbol{k}$, which was possible only because P_ζ depends on $k = |\boldsymbol{k}|$ and not on the direction: that is where isotropy enters. The formula carries three lessons. First, $a_{\ell m}$ is a linear combination of Gaussian $\zeta_{\boldsymbol{k}}$, hence Gaussian. Second, the whole cosmology sits in P_ζ (the amplitude A_s and tilt n_s) and in $g_{T\ell}$ (the other four parameters). Third, CAMB returns C_ℓ, an expectation value. It never returns a sky; to get one we must draw the $a_{\ell m}$ ourselves.

How to call it. The fiducial spectra are computed once, in `code/camb_fiducial.py`:

Listing 7.5: The CAMB call behind every fiducial spectrum (`code/camb_fiducial.py`)

```
13 FIDUCIAL = dict(H0=67.36, ombh2=0.02237, omch2=0.1200, tau=0.0544, As=2.100e-9, ns=0.9649,
       mnu=0.06)
14
15
16 def compute(params=FIDUCIAL, lmax=LMAX):
17     import camb
18     p = camb.set_params(**params, lmax=lmax + 200, lens_potential_accuracy=1)
19     r = camb.get_results(p)
20     cls = r.get_cmb_power_spectra(p, CMB_unit="muK", raw_cl=True)
21     tot = cls["total"][: lmax + 1]
22     return np.arange(lmax + 1), tot  # columns TT, EE, BB, TE
```

Line 13 is the Planck 2018 best fit [25] in CAMB's keywords (the dictionary of Chapter T1). Line 18 builds a parameter object; computing to $\ell_{\max} + 200$ and setting `lens_potential_accuracy=1` keep the lensed spectra accurate up to $\ell_{\max} = 3000$. Line 19 solves the Boltzmann equations. Line 20 asks for the *raw* C_ℓ in $\mu\mathrm{K}^2$ (without `raw_cl=True` CAMB returns D_ℓ, a factor $\ell(\ell+1)/2\pi$ larger). Line 21 keeps the lensed total, with columns TT, EE, BB, TE. The result is cached in `data/camb_fiducial.npz`, and every script loads it with `load_fiducial()`, which returns ℓ and C_ℓ^{TT} with $C_0 = C_1 = 0$. Rerunning CAMB takes seconds, and is needed only when a script varies the parameters.

Example: the spectrum and its irreducible error bar. Figure 7.25 shows the fiducial $D_\ell = \ell(\ell+1)C_\ell/2\pi$ with the band $D_\ell[1 \pm \sqrt{2/(2\ell+1)}]$ of (7.78), and the $\widehat{D}_\ell$ of one simulated sky (`code/ch07/24_camb_healpix.py`). The first acoustic peak is at $\ell = 220$ with $D_\ell = 5732\,\mu\mathrm{K}^2$; there the band is $\pm\sqrt{2/441} = \pm 6.7\%$. At $\ell = 2$ the band is ±63% wide and the one sky wanders through it; above $\ell \sim 500$ the band is too thin to see and the simulated points trace the curve. Large scales are known poorly from one sky, and nothing can be done about it.

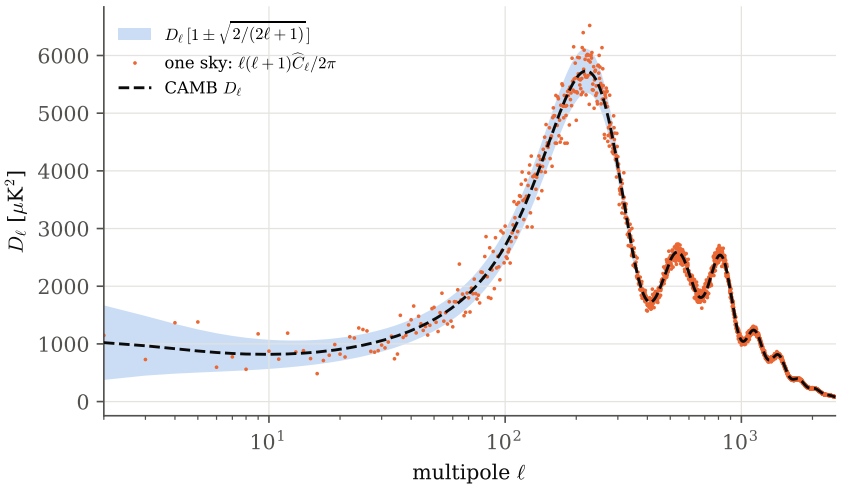


Figure 7.25: The fiducial CAMB spectrum D_ℓ (black dashed), its cosmic-variance band $\pm\sqrt{2/(2\ell+1)}$ (blue), and the estimate $\ell(\ell+1)\widehat{C}_\ell/2\pi$ from one simulated full sky up to $\ell = 2500$ (orange points). The logarithmic ℓ axis stretches the few low multipoles, where the band is wide, over a large part of the plot.

7.11.2 HEALPix: cutting the sphere into equal pixels

A computer stores a map as a list of numbers, one per pixel, so we must cut the sphere into pixels. Latitude–longitude grids are a poor choice. Their pixels shrink towards the poles. Noise averaged over a small polar pixel then has a different variance from noise averaged over a large equatorial one, so the map's noise changes with latitude. Górski et al. list three properties a good pixelisation needs [7]:

1. *Equal areas*, so that white noise in time stays white noise in the map and every pixel samples the sky with the same weight;
2. *Pixel centres on rings of constant latitude*, so that spherical-harmonic transforms are fast: $Y_{\ell m} \propto P_\ell^m(\cos\theta)e^{im\phi}$, the expensive Legendre functions need to be evaluated once per ring rather than once per pixel, and the $e^{im\phi}$ sum along each ring is an FFT;
3. *A hierarchy*: each pixel splits into four at the next resolution, so that nearby pixels are nearby in memory.

Working definition 7.8 (HEALPix grid)

The **HEALPix** grid starts from 12 equal-area base pixels and divides each into $N_{\mathrm{side}} \times N_{\mathrm{side}}$ sub-pixels [N_{side} is a power of 2 that sets the resolution]. It has

$$N_{\mathrm{pix}} = 12\,N_{\mathrm{side}}^2 \text{ pixels}, \qquad \Omega_{\mathrm{pix}} = \frac{4\pi}{N_{\mathrm{pix}}} = \frac{\pi}{3N_{\mathrm{side}}^2}, \qquad 4N_{\mathrm{side}} - 1 \text{ rings of constant latitude}, \tag{7.86}$$

and a full spherical-harmonic transform costs of order $N_{\text{pix}}^{3/2}$ operations instead of N_{pix}^2 [7], [Hivon p. 4].

Figure 7.26 shows the first three levels. The pixels are not squares (they are curvilinear quadrilaterals, and near the poles they are visibly distorted), but they all have exactly the same area.

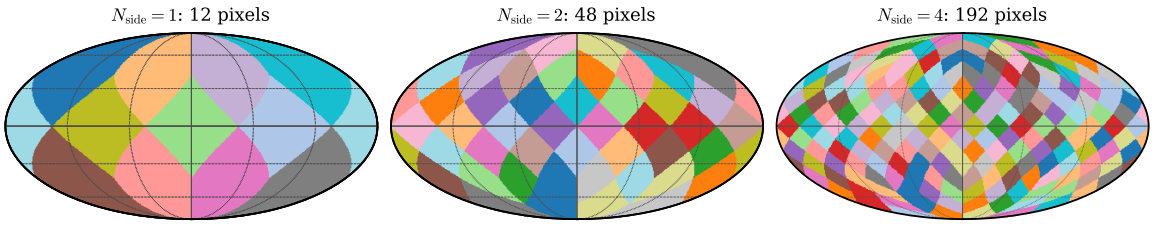


Figure 7.26: HEALPix tilings with $N_{\text{side}} = 1, 2, 4$, each pixel painted a random colour, with lines of latitude and longitude every 30° and 60°. The twelve base pixels form three rings of four; each step in N_{side} splits every pixel into four of equal area.

Where the pixel edges come from. The edges in Figure 7.26 look like arbitrary curves, but every one of them follows from the demand for equal areas and an observation of Archimedes. Wrap the unit sphere in a cylinder that touches it along the equator, and push each point of the sphere straight out, away from the axis and at constant height, onto the cylinder (Figure 7.27a). The point at polar angle θ and longitude ϕ lands at the same ϕ and at height $z = \cos\theta$. This map keeps areas. Take the small patch between θ and $\theta + \Delta\theta$ and between ϕ and $\phi + \Delta\phi$:

$$\Delta\Omega \overset{(1)}{=} (\sin\theta\,\Delta\phi)\,(\Delta\theta) \overset{(2)}{=} \Delta\phi\;|\Delta z|\,. \tag{7.87}$$

(1) The patch is nearly a rectangle. Its width runs along the circle of latitude, whose radius is $\sin\theta$, so the width is $\sin\theta\,\Delta\phi$; its height runs along the meridian and is $\Delta\theta$.

(2) On the cylinder, of radius 1, the same patch has width $\Delta\phi$ and height $|\Delta z|$, and $z = \cos\theta$ gives $|\Delta z| = \sin\theta\,\Delta\theta$.

The image on the cylinder is wider than the patch by $1/\sin\theta$ and shorter by $\sin\theta$, and the two factors cancel. Now cut the cylinder open along $\phi = 0$ and lay it flat: a rectangle $0 \le \phi < 2\pi$, $-1 \le z \le 1$, of area 4π (Figure 7.27b). A region of the sphere has the same area as its image in this rectangle, so *equal-area pixels on the sphere are the same thing as equal-area tiles in the flat* (ϕ, z) *rectangle*. We design the tiles there.

The equatorial belt. Here the pixels are laid out in rings of constant z, and every ring holds the same number of pixels, $4N_{\text{side}}$: the equator crosses four base pixels, each N_{side} pixels across. Two numbers then fix everything, the width $\Delta\phi$ of a pixel along its ring and the spacing Δz between neighbouring rings (Figure 7.27b):

$$\Delta\phi \overset{(1)}{=} \frac{2\pi}{4N_{\text{side}}} = \frac{\pi}{2N_{\text{side}}}, \qquad \Delta z \overset{(2)}{=} \frac{\Omega_{\text{pix}}}{\Delta\phi} \overset{(3)}{=} \frac{\pi/(3N_{\text{side}}^2)}{\pi/(2N_{\text{side}})} = \frac{2}{3N_{\text{side}}}. \tag{7.88}$$

(1) A ring goes once round, 2π, and holds $4N_{\text{side}}$ pixels.

(2) In the flat picture area is $\Delta\phi\,\Delta z$, and each pixel must have area Ω_{pix}.

(3) $\Omega_{\text{pix}} = \pi/(3N_{\text{side}}^2)$ from (7.86).

The rings of pixel centres in the belt are therefore at

$$z_j = j\,\Delta z = \frac{2j}{3N_{\text{side}}}, \qquad \cos\theta_j = \frac{2j}{3N_{\text{side}}}, \qquad j = -N_{\text{side}}, \ldots, N_{\text{side}}, \tag{7.89}$$

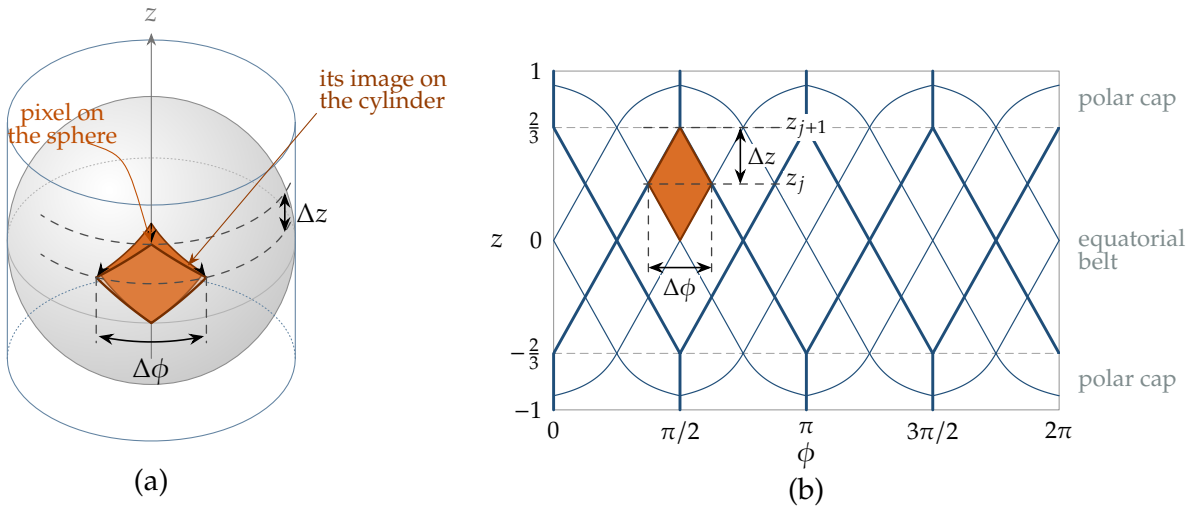


Figure 7.27: Archimedes' map and the HEALPix edges. (a) Each point of the sphere is pushed straight out, at constant height, onto the cylinder that touches the sphere along the equator. One pixel for $N_{\text{side}} = 2$ is shown on the sphere (solid) and as its image on the cylinder (see-through); the dashed rings z_j and z_{j+1} and the dashed lines below its side corners mark Δz and $\Delta\phi$ on the cylinder. (b) The cylinder cut open along $\phi = 0$ and laid flat. The thick lines are the boundaries of the twelve base pixels, the thin ones those for $N_{\text{side}} = 2$: straight (7.91) in the belt $|z| \le \frac{2}{3}$, curved (7.94) in the two polar caps. The same pixel is shaded: its width along its ring is $\Delta\phi = \pi/(2N_{\text{side}})$, and the rings z_j and z_{j+1} of pixel centres are $\Delta z = 2/(3N_{\text{side}})$ apart (7.88), so its area is $\Delta\phi\,\Delta z = \Omega_{\text{pix}}$.

and the belt ends at $j = \pm N_{\text{side}}$, $z = \pm\frac{2}{3}$, which is where the number $\frac{2}{3}$ in HEALPix comes from. If the pixels of neighbouring rings sat directly above one another, they would be rectangles $\Delta\phi \times \Delta z$. HEALPix instead shifts each ring by half a pixel, $\Delta\phi/2$, against its neighbours, so a pixel centre sits directly above the gap between two pixels of the ring below. The pixel around the centre (ϕ_c, z_j) is then the diamond with corners at $(\phi_c \pm \frac{1}{2}\Delta\phi,\ z_j)$ on its own ring and at $(\phi_c,\ z_j \pm \Delta z)$, the centres' height of the rings above and below. Its area is half the product of its diagonals, $\frac{1}{2}\,\Delta\phi \cdot 2\Delta z = \Delta\phi\,\Delta z = \Omega_{\text{pix}}$, as required, and its edges are straight lines of slope

$$\pm\frac{\Delta z}{\Delta\phi/2} = \pm\frac{2/(3N_{\text{side}})}{\pi/(4N_{\text{side}})} = \pm\frac{8}{3\pi}, \tag{7.90}$$

the same for every N_{side}. Edges of the same slope pass through corners that are $\Delta\phi$ apart along a ring, so their values at $\phi = 0$ differ by $\frac{8}{3\pi}\Delta\phi = \frac{4}{3N_{\text{side}}}$, and one of them passes through the corner at $(0, \frac{2}{3})$. The pixel edges in the belt are thus

$$z = \frac{2}{3} - \frac{4k}{3N_{\text{side}}} \pm \frac{8\phi}{3\pi}, \qquad k \text{ an integer}, \quad |z| \le \frac{2}{3}, \tag{7.91}$$

straight lines on the flat rectangle and curves with $\cos\theta$ linear in ϕ on the sphere. For $N_{\text{side}} = 1$ the same construction gives the twelve base pixels: four diamonds $\Delta\phi = \frac{\pi}{2}$ wide and $2\Delta z = \frac{4}{3}$ tall around the equator, with the four northern and four southern base pixels filling the triangles between them.

The polar caps. Above $z = \frac{2}{3}$ the flat picture has a strip of height $\frac{1}{3}$ and area $2\pi/3$. It belongs to the four northern base pixels. The lower part of each, a triangle between two equatorial diamonds, has area $\frac{1}{2} \cdot \frac{\pi}{2} \cdot \frac{2}{3} = \frac{\pi}{6}$, so the remaining $\frac{\pi}{6}$ of each is a quarter of the strip, a rectangle $\phi_t \in [0, \frac{\pi}{2}]$, $z \in [\frac{2}{3}, 1]$, where ϕ_t is ϕ measured from the start of its quarter. Diamonds cannot tile it, because its top edge $z = 1$ is not a line on the sphere but a single point, the pole: on the sphere the rectangle is a triangle with one corner at the pole. We need coordinates in which it

is drawn as a triangle and areas are still uniform. Near the pole $1 - z = 1 - \cos\theta \approx \theta^2/2$, so $\sqrt{1-z}$ measures distance from the pole. Set

$$\rho = \sqrt{3(1-z)}, \qquad u = \frac{2\phi_t}{\pi}\,\rho, \qquad v = \Big(1 - \frac{2\phi_t}{\pi}\Big)\rho. \tag{7.92}$$

The pole, $\rho = 0$, becomes the single point $u = v = 0$. The edge $\phi_t = 0$ becomes $u = 0$, the edge $\phi_t = \frac{\pi}{2}$ becomes $v = 0$, and the bottom edge $z = \frac{2}{3}$, where $\rho = 1$, becomes the line $u + v = 1$. The quarter-cap is the triangle $u, v \geq 0$, $u + v \leq 1$, half of a unit square. Its areas are uniform:

$$\Delta\Omega \overset{(1)}{=} |\Delta z|\;\Delta\phi_t \overset{(2)}{=} \frac{2\rho}{3}\,\Delta\rho\,\Delta\phi_t \overset{(3)}{=} \frac{\pi}{3}\,\Delta u\,\Delta v. \tag{7.93}$$

(1) Archimedes' rule (7.87).

(2) $z = 1 - \rho^2/3$, so $|\Delta z| = \frac{2\rho}{3}\Delta\rho$.

(3) The area of a small cell in the (u, v) plane is the Jacobian times $\Delta\rho\,\Delta\phi_t$: with $\partial u/\partial\rho = 2\phi_t/\pi$, $\partial u/\partial\phi_t = 2\rho/\pi$, $\partial v/\partial\rho = 1 - 2\phi_t/\pi$ and $\partial v/\partial\phi_t = -2\rho/\pi$, the determinant is $-\frac{2\phi_t}{\pi}\frac{2\rho}{\pi} - \frac{2\rho}{\pi}\big(1 - \frac{2\phi_t}{\pi}\big) = -\frac{2\rho}{\pi}$, so $\Delta u\,\Delta v = \frac{2\rho}{\pi}\Delta\rho\,\Delta\phi_t$; replace $\frac{2\rho}{3}\Delta\rho\,\Delta\phi_t$ by $\frac{\pi}{3}\Delta u\,\Delta v$.

The factor $\pi/3$ is the same everywhere, so equal areas in the (u, v) triangle are equal areas on the sphere (the whole triangle, of area $\frac{1}{2}$, is $\frac{\pi}{6}$, as it should be). In the triangle we can again cut with straight lines, $u = k/N_{\text{side}}$ and $v = k/N_{\text{side}}$, into squares of area $1/N_{\text{side}}^2$, each a pixel of area $\frac{\pi}{3N_{\text{side}}^2} = \Omega_{\text{pix}}$; the squares cut in half by $u + v = 1$ are completed by the half-diamonds just below $z = \frac{2}{3}$, since at $\rho = 1$ the line $u = k/N_{\text{side}}$ sits at $\phi_t = k\pi/(2N_{\text{side}})$, exactly where the belt edge (7.91) with the same k reaches $z = \frac{2}{3}$. Turning the straight lines back into (ϕ, z) with (7.92): $u = k/N_{\text{side}}$ means $\frac{2\phi_t}{\pi}\sqrt{3(1-z)} = \frac{k}{N_{\text{side}}}$; square both sides and solve for z, and do the same for v:

$$z = 1 - \frac{k^2}{3N_{\text{side}}^2}\Big(\frac{\pi}{2\phi_t}\Big)^2, \qquad z = 1 - \frac{k^2}{3N_{\text{side}}^2}\Big(\frac{\pi}{\pi - 2\phi_t}\Big)^2, \qquad z \geq \frac{2}{3}, \tag{7.94}$$

and the same with $z \to -z$ in the south. These are the curved edges near the top and bottom of Figure 7.27b, and (7.91) and (7.94) are the pixel boundaries of [7]. The corners again sit on rings: where $u = k/N_{\text{side}}$ meets $v = j/N_{\text{side}}$, $\rho = u + v = (k + j)/N_{\text{side}}$ and $z = 1 - (k + j)^2/(3N_{\text{side}}^2)$.

Example: pixel sizes and band limits. The side of a pixel is about the square root of its area,

$$\theta_{\text{pix}} \approx \sqrt{\Omega_{\text{pix}}} = \sqrt{\frac{\pi}{3}}\,\frac{1}{N_{\text{side}}}\ \text{rad} = \frac{58.6^\circ}{N_{\text{side}}}. \tag{7.95}$$

For $N_{\text{side}} = 256$ that is 13.74′ with 786432 pixels; $N_{\text{side}} = 512$ gives 6.871′ and 3145728 pixels; Planck's high-frequency maps use $N_{\text{side}} = 2048$, 1.718′ and 50,331,648 pixels. How high in ℓ can a map of N_{pix} numbers go? A sky band-limited to $\ell_{\max}$ has $(\ell_{\max} + 1)^2$ real degrees of freedom (Section 7.8.2), and N_{pix} numbers cannot determine more than N_{pix} unknowns:

$$(\ell_{\max} + 1)^2 \leq 12N_{\text{side}}^2 \quad\Longrightarrow\quad \ell_{\max} \lesssim \sqrt{12}\,N_{\text{side}} \approx 3.5\,N_{\text{side}}. \tag{7.96}$$

In practice healpy never goes beyond $\ell_{\max} = 3N_{\text{side}} - 1$, and analyses that need accurate coefficients stay below about $2N_{\text{side}}$, as the next example shows.

7.11.3 healpy: four operations and their from-scratch versions

The Python package healpy wraps the HEALPix library. Four functions do everything in this chapter, and two more are shorthands (Figure 7.28):

- `synalm(cl, lmax)`: draw one set of Gaussian $a_{\ell m}$ with spectrum C_ℓ;
- `alm2map(alm, nside)`: sum the series (7.44) at every pixel centre;
- `map2alm(map, lmax, iter)`: compute the coefficients of a map;
- `alm2cl(alm)`: the estimator $\widehat{C}_\ell$ of (7.72);
- `synfast` = `synalm` then `alm2map`, and `anafast` = `map2alm` then `alm2cl`.

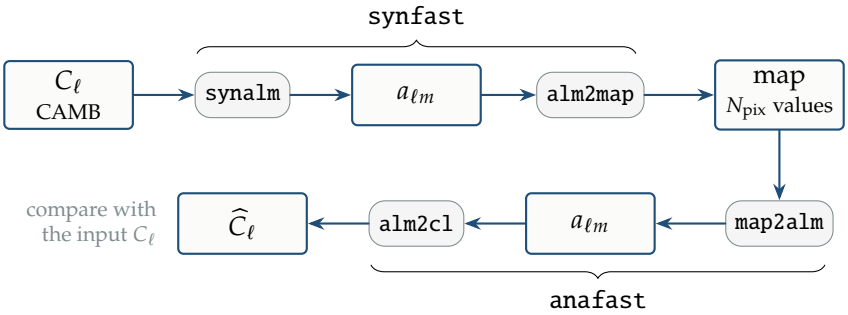


Figure 7.28: The healpy pipeline. Objects are in blue boxes, functions in grey. `synfast` draws coefficients from a spectrum and turns them into a map; `anafast` turns a map back into coefficients and squares them into $\widehat{C}_\ell$. Comparing the output $\widehat{C}_\ell$ with the input C_ℓ, over many skies, is the Monte Carlo study of Section 7.12.

Storage. healpy stores only the coefficients with $m \geq 0$, since the others follow from the reality condition (7.46). They are ordered by m first: all ℓ for $m = 0$, then all $\ell \geq 1$ for $m = 1$, and so on, so that $a_{\ell m}$ sits at position $m(2\ell_{\max} + 1 - m)/2 + \ell$. For $\ell_{\max} = 20$, the coefficient $a_{5,3}$ is at $3 \times 38/2 + 5 = 62$, which is what `healpy.Alm.getidx(20, 5, 3)` returns. Neighbouring ℓ at fixed m are neighbours in memory, which is what made the modulation code of Section 7.9.5 so short.

Drawing a sky by hand. To see the mechanism we first write `synalm` ourselves, straight from the table of real degrees of freedom in Section 7.10.1:

Listing 7.6: Isotropic Gaussian coefficients from a spectrum (`code/ch07/lib_harmonic.py`)

```
34 def synalm(cl, lmax: int, rng: np.random.Generator):
35     """One isotropic Gaussian realisation of the a_lm (m >= 0) with spectrum cl.
36
37     m = 0 : a_l0 is real,        a_l0 ~ N(0, C_l)
38     m > 0 : Re a_lm, Im a_lm are independent N(0, C_l / 2), so <|a_lm|^2> = C_l
39     """
40     cl = np.asarray(cl, dtype=float)
41     ell, m = getlm(lmax)
42     sig = np.sqrt(cl[ell])
43     re = rng.standard_normal(ell.size)
44     im = rng.standard_normal(ell.size)
45     return np.where(m == 0, sig * re, sig * (re + 1j * im) / np.sqrt(2.0)).astype(complex)
```

Line 42 gives every stored coefficient the standard deviation $\sqrt{C_\ell}$ of its multipole. Lines 43–44 draw two standard normals per coefficient. Line 45 is the table: for $m = 0$ the coefficient is real with variance C_ℓ (the imaginary draw is discarded); for $m > 0$ the real and imaginary parts are each $\sqrt{C_\ell/2}$ times a standard normal. Nothing else is needed for an isotropic Gaussian sky, because the coefficients are independent: this is the diagonal Cholesky factor of Chapter 6, one $\sqrt{C_\ell}$ per mode. The library function `healpy.synalm` does the same with numpy's global random stream; in Section 7.12 we check that the two produce statistically identical skies.

From a map back to coefficients. The integral (7.44) becomes a sum over pixels, $a_{\ell m} \approx \Omega_{\mathrm{pix}} \sum_p \Theta(p)\, Y^*_{\ell m}(p)$ [Hivon (12)], with Ω_{pix} the pixel area of (7.86). That sum is a quadrature rule, and it is not exact. healpy improves it by iteration (`iter`): synthesise a map from the current coefficients, analyse the difference between that map and the input map, and add the correction.

How accurate is the round trip $a_{\ell m} \to \text{map} \to a_{\ell m}$? We test it at $N_{\mathrm{side}} = 256$ (Figure 7.29) with the script `code/ch07/24_camb_healpix.py`. For a sky band-limited to $\ell \le 2N_{\mathrm{side}} = 512$, the recovered $\widehat{C}_\ell$ agrees with that of the input coefficients to a median relative error of 2.272×10^{-5} without iteration, and 1.116×10^{-8} with three iterations (the worst multipole is off by 3.973×10^{-7}). Now let the sky contain power up to $3N_{\mathrm{side}} - 1 = 767$. The pixels can no longer tell the highest multipoles apart from lower ones, so power leaks downwards (**aliasing**). The error below $\ell = 512$ rises to a median of 2.213×10^{-6}, and near $\ell = 3N_{\mathrm{side}}$ it reaches ten percent.

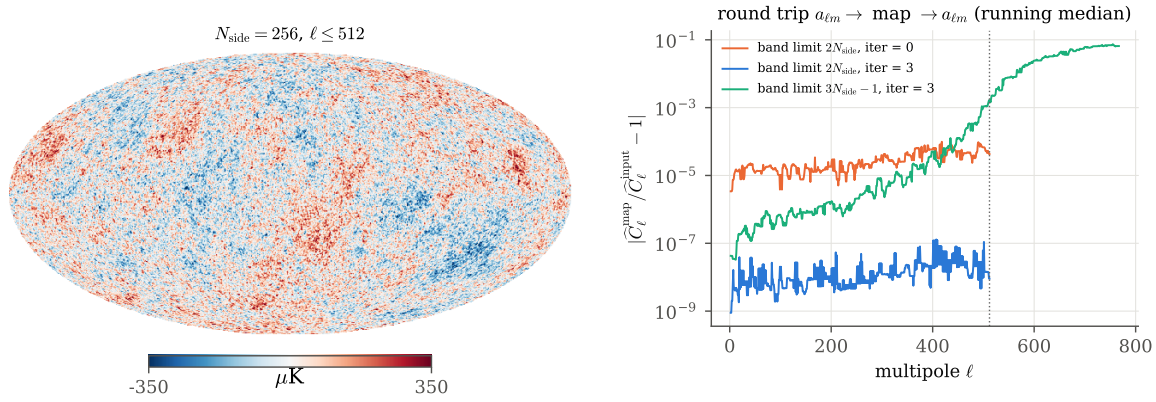


Figure 7.29: Left: a simulated sky at $N_{\mathrm{side}} = 256$ with $\ell \le 512$. Right: relative difference between $\widehat{C}_\ell$ measured on the map (`anafast`) and $\widehat{C}_\ell$ of the coefficients the map was made from, smoothed with a running median over 9 multipoles. Iterating the quadrature (blue) gains three orders of magnitude over a single pass (orange). With power up to $3N_{\mathrm{side}} - 1$ (green), aliasing spoils the high multipoles. The dotted line marks $\ell = 2N_{\mathrm{side}}$.

Watch out: A map is not a band-limited sky

The real sky has power at every ℓ, and a pixel averages the sky over its area. Both effects change the measured spectrum: power above the band limit aliases into lower multipoles, and the pixel average multiplies C_ℓ by the square of a *pixel window function* (healpy's `pixwin`), just as a beam multiplies it by B_ℓ^2 [Hivon (15)]. A simulation that draws $a_{\ell m}$ only up to ℓ_{max} and analyses at the same ℓ_{max} sees neither effect. Choose N_{side} so that the beam has already killed the signal well below $2N_{\mathrm{side}}$, and include the pixel window when comparing with data (Chapter 8).

7.12 Cosmic variance by Monte Carlo

We now have two independent routes to the same numbers. Algebra gave the exact sampling distribution of $\widehat{C}_\ell$ (Section 7.10.2): a scaled $\chi^2_{2\ell+1}$ with variance $2C_\ell^2/(2\ell+1)$. The tools of Section 7.11 let us make as many skies as we like from the CAMB spectrum. If we draw many skies, measure $\widehat{C}_\ell$ on each, and histogram the results, the histogram must reproduce the algebra. The comparison does two jobs. It checks every step of the derivation: the reality condition, the counting, the factor of two. And it builds, where the right answer is known, the same machine that Chapter 8 will use where no answer is known.

The questions. With N simulated full skies from the fiducial C_ℓ:

1. Does the histogram of $\widehat{C}_\ell$ at a given ℓ follow the scaled $\chi^2_{2\ell+1}$?
2. Does the Monte Carlo variance of $\widehat{C}_\ell$ converge to $2C_\ell^2/(2\ell+1)$, and how fast: how large is the Monte Carlo error on a Monte Carlo variance?
3. Is the covariance between different multipoles zero, and what does "zero" look like when it is estimated from N skies?

7.12.1 The simulation

The script `code/ch07/25_mc_cosmic_variance.py` simulates $N = 20{,}000$ full skies for $2 \le \ell \le 300$. It does not build maps. On the full sky a map is a lossless re-encoding of the coefficients: Figure 7.29 showed that the round trip returns them to a part in 10^8. So drawing the coefficients and applying (7.72) is the same experiment, thousands of times faster. For each multipole the script draws the $2\ell+1$ real numbers of the table in Section 7.10.1 directly:

Listing 7.7: $\widehat{C}_\ell$ for many independent skies, one multipole at a time (`code/ch07/lib_harmonic.py`)

```
59 def chat_sims(cl, ells, nsim: int, rng: np.random.Generator):
60     """C_hat_l for nsim independent full skies, multipole by multipole (memory-light).
61
62     For each l we draw the 2l+1 real numbers that make up the a_lm of one sky:
63     a_l0, and Re/Im of a_lm for m = 1..l.  Returns an array of shape (nsim, len(ells)).
64     """
65     out = np.empty((nsim, len(ells)))
66     for j, l in enumerate(ells):
67         z = rng.standard_normal((nsim, 2 * l + 1))
68         a0 = np.sqrt(cl[l]) * z[:, 0]                        # real, variance C_l
69         reim = np.sqrt(cl[l] / 2.0) * z[:, 1:]               # l real parts + l imaginary
       parts
70         out[:, j] = (a0 ** 2 + 2.0 * np.sum(reim ** 2, axis=1)) / (2 * l + 1)
71     return out
```

Line 67 draws a block of standard normals, one row per sky and $2\ell+1$ columns. Line 68 turns the first column into $a_{\ell 0}$, with variance C_ℓ; line 69 turns the other 2ℓ columns into the real and imaginary parts of $a_{\ell 1}, \ldots, a_{\ell\ell}$, each with variance $C_\ell/2$. Line 70 is the estimator (7.72) with the factor 2 for the negative m. Nowhere does the code use the χ^2 result; it only uses the covariance (7.62) and the reality condition. The driver is short:

Listing 7.8: The Monte Carlo run (`code/ch07/25_mc_cosmic_variance.py`)

```
27 N_SIM, LMAX = 20000, 300
28 ELLS = np.arange(2, LMAX + 1)
29 rng = rng_for("ch07", "25_mc_cosmic_variance", stream=0)
30
31 # ------------------------------------------------------------------ the simulation
32 chat = lh.chat_sims(cl, ELLS, N_SIM, rng)          # shape (N_SIM, 299): one row per
       simulated sky
33 y = chat / cl[ELLS]                                 # C_hat_l / C_l, mean 1, variance
       2/(2l+1)
34 nu = 2 * ELLS + 1
```

Line 32 returns an array with one row per sky and one column per multipole; line 33 divides by the input spectrum, so that every column of `y` should have mean 1 and variance $2/(2\ell+1)$. The run draws about 1.8×10^9 Gaussian numbers and takes under a minute.

7.12.2 The sampling distribution of $\widehat{C}_\ell$

The first question has a clear answer: yes. Figure 7.30 compares the histograms of $\widehat{C}_\ell/C_\ell$ at $\ell = 2$, 10, 50 and 300 with the scaled $\chi^2_{2\ell+1}$ density (7.76). The agreement is visible by eye, including the strong skew at $\ell = 2$. There the Gaussian of the same mean and variance (dotted) is plainly wrong: it puts probability on negative power and misses the peak. For a quantitative check we use the Kolmogorov–Smirnov test of Chapter 4, with the null hypothesis "$(2\ell+1)\widehat{C}_\ell/C_\ell$ is $\chi^2_{2\ell+1}$". The p-values are 0.2452, 0.8436, 0.3608 and 0.6115: with 20,000 skies per multipole, nothing distinguishes the simulation from the scaled χ^2. The mean is right too: the average of $\widehat{C}_\ell/C_\ell$ over all skies is 1 at every one of the 299 multipoles. The largest deviation, in units of its standard error $\sqrt{2/[(2\ell+1)N]}$, is 3.081, which is the typical size of the largest of 299 independent standard normal deviations. The estimator is unbiased.

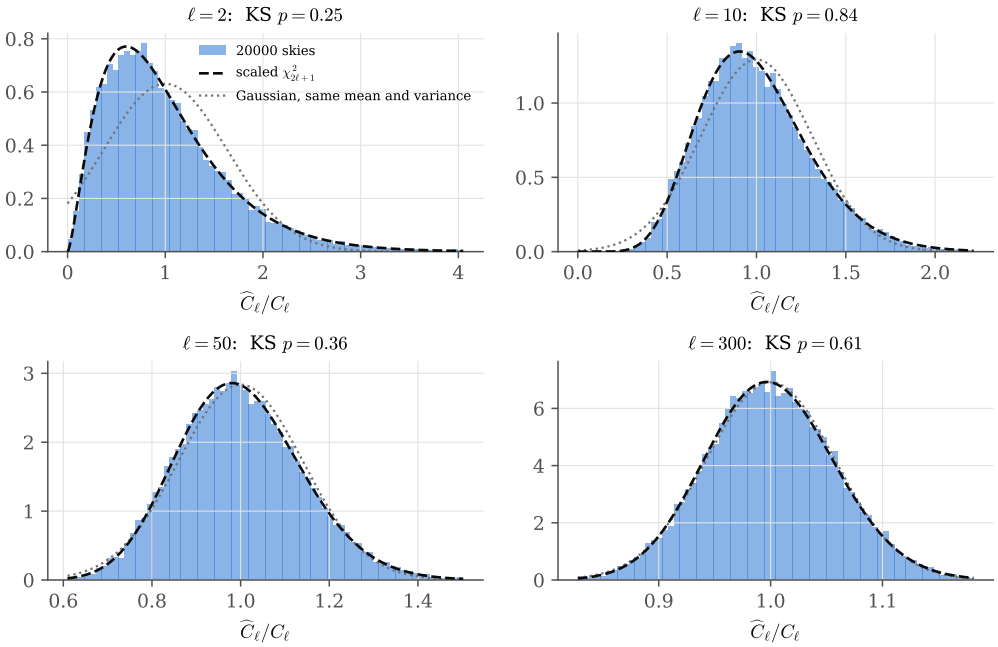


Figure 7.30: Histograms of $\widehat{C}_\ell/C_\ell$ from 20,000 simulated full skies at four multipoles (blue), the scaled $\chi^2_{2\ell+1}$ density (black dashed) and a Gaussian with the same mean and variance (grey dotted). The quadrupole is strongly skewed; by $\ell = 300$ all three agree. Each title gives the Kolmogorov–Smirnov p-value of the simulation against the χ^2.

7.12.3 How fast does the Monte Carlo variance converge?

From N simulated values $X_1, \dots, X_N$ of $\widehat{C}_\ell$ we estimate its variance with the sample variance $S^2 = \frac{1}{N-1}\sum_i (X_i - \bar{X})^2$ (Chapter 3), and compare with $2C_\ell^2/(2\ell+1)$. S^2 is itself a random quantity: run the simulation again with other seeds and it changes. To know whether a 2% discrepancy is a bug or bad luck we need the standard deviation of S^2.

For Gaussian X. If the X_i are independent $\mathcal{N}(\mu, \sigma^2)$, then $(N-1)S^2/\sigma^2 \sim \chi^2_{N-1}$ (Chapter 2), so

$$\operatorname{Var} S^2 = \left(\frac{\sigma^2}{N-1}\right)^2 2(N-1) = \frac{2\sigma^4}{N-1}, \qquad \frac{\sqrt{\operatorname{Var} S^2}}{\sigma^2} = \sqrt{\frac{2}{N-1}}. \tag{7.97}$$

This is cosmic variance again, one level up: the Monte Carlo variance is a variance estimated from N numbers, just as $\widehat{C}_\ell$ is a variance estimated from $2\ell+1$ numbers. But $\widehat{C}_\ell$ is not Gaussian,

and at low ℓ it is far from Gaussian, so we need a formula that holds for any distribution.

For any X: the fourth moment. Let the X_i be independent with mean μ, variance σ^2 and fourth central moment μ_4. Shift to $\mu = 0$ (the sample variance does not change). Expand $(N-1)S^2 = \sum_i X_i^2 - N\bar{X}^2$ into a diagonal part $A = \sum_i X_i^2$ and an off-diagonal part $B = \sum_{i\neq j} X_i X_j$:

$$\begin{aligned}(N-1)S^2 &\overset{(1)}{=} \sum_i X_i^2 - \frac{1}{N}\sum_{i,j} X_i X_j \overset{(2)}{=} \Big(1-\frac{1}{N}\Big)A - \frac{1}{N}B,\\ \operatorname{Var}\big[(N-1)S^2\big] &\overset{(3)}{=} \Big(1-\frac{1}{N}\Big)^2 N(\mu_4-\sigma^4) + \frac{1}{N^2}\,2N(N-1)\sigma^4,\\ \operatorname{Var} S^2 &\overset{(4)}{=} \frac{\mu_4-\sigma^4}{N} + \frac{2\sigma^4}{N(N-1)} \overset{(5)}{=} \frac{1}{N}\Big[\mu_4 - \frac{N-3}{N-1}\,\sigma^4\Big].\end{aligned} \tag{7.98}$$

(1) $N\bar{X}^2 = \frac{1}{N}(\sum_i X_i)^2 = \frac{1}{N}\sum_{i,j} X_i X_j$.
(2) Split the double sum into $i = j$ (which is A) and $i \neq j$ (which is B).
(3) $\operatorname{Var} A = N \operatorname{Var}(X^2) = N(\mu_4 - \sigma^4)$. $\mathbb{E}\,B = 0$, and in $\mathbb{E}\,B^2 = \sum_{i\neq j}\sum_{k\neq l}\mathbb{E}[X_iX_jX_kX_l]$ only the terms with $\{k,l\} = \{i,j\}$ survive, two for each of the $N(N-1)$ ordered pairs, each equal to σ^4. $\operatorname{Cov}(A,B) = \sum_k \sum_{i\neq j}\mathbb{E}[X_k^2 X_i X_j] = 0$, because at least one of X_i, X_j appears exactly once and has mean zero.
(4) Divide by $(N-1)^2$: $(1-1/N)^2/(N-1)^2 = 1/N^2$ times N gives $1/N$, and $2N(N-1)/[N^2(N-1)^2] = 2/[N(N-1)]$.
(5) Combine over the common factor $1/N$: $-\sigma^4 + 2\sigma^4/(N-1) = -\sigma^4(N-3)/(N-1)$.

This is the variance of the sample variance derived in Chapter 3. Divide by σ^4 and write $\mu_4/\sigma^4 = 3 + \gamma_2$, with γ_2 the **excess kurtosis** [the fourth moment in excess of a Gaussian's, zero for a Gaussian]:

$$\frac{\operatorname{Var} S^2}{\sigma^4} = \frac{1}{N}\Big[3 + \gamma_2 - \frac{N-3}{N-1}\Big] = \frac{2}{N-1} + \frac{\gamma_2}{N}. \tag{7.99}$$

For $\gamma_2 = 0$ this is (7.97). Heavy tails ($\gamma_2 > 0$) make the Monte Carlo variance converge more slowly, but only by a constant factor; the rate $N^{-1/2}$ is unchanged.

The kurtosis of $\widehat{C}_\ell$. To use (7.99) we need γ_2 for $\widehat{C}_\ell$. Since $\widehat{C}_\ell$ is a rescaled χ^2_ν, and rescaling does not change γ_2, we need the excess kurtosis of χ^2_ν. Its moment generating function is $M(t) = (1-2t)^{-\nu/2}$ (Chapter 2), so its cumulant generating function is

$$K(t) = \ln M(t) \overset{(1)}{=} -\frac{\nu}{2}\ln(1-2t) \overset{(2)}{=} \frac{\nu}{2}\sum_{n\geq 1}\frac{(2t)^n}{n} \overset{(3)}{=} \sum_{n\geq 1}\underbrace{\nu\,2^{n-1}(n-1)!}_{\kappa_n}\,\frac{t^n}{n!}, \tag{7.100}$$

(1) Logarithm of M.
(2) $-\ln(1-x) = \sum_{n\geq 1} x^n/n$ with $x = 2t$.
(3) Rewrite $\frac{\nu}{2}\,2^n/n = \nu 2^{n-1}(n-1)!/n!$; the coefficient of $t^n/n!$ is the n-th cumulant κ_n.

so $\kappa_2 = 2\nu$ (the variance) and $\kappa_4 = 48\nu$. The fourth cumulant is the fourth central moment minus its Gaussian value, $\kappa_4 = \mu_4 - 3\kappa_2^2$, so $\gamma_2 = \kappa_4/\kappa_2^2 = 48\nu/4\nu^2 = 12/\nu$. With $\nu = 2\ell + 1$, (7.99) becomes the prediction we test:

Key idea: The Monte Carlo error on a Monte Carlo variance

Estimating $\mathrm{Var}\,\widehat{C}_\ell$ from N simulated skies gives a fractional error

$$\frac{\sqrt{\mathrm{Var}\,S^2}}{2C_\ell^2/(2\ell+1)} = \sqrt{\frac{2}{N-1}+\frac{12}{(2\ell+1)\,N}} \approx \frac{1}{\sqrt{N}}\sqrt{2+\frac{12}{2\ell+1}}. \tag{7.101}$$

It falls as $N^{-1/2}$. At high ℓ the coefficient is the Gaussian $\sqrt{2}$; at the quadrupole it is $\sqrt{2+12/5} = 2.098$, half as large again, because the skewed χ^2_5 has heavy tails.

Example: how many simulations did Knox need? Knox estimated parameter errors by simulating data sets and reported that "typically we perform 100 simulations which is sufficient to determine the standard deviation to better than 10%" [166]. Is that right? A standard deviation is the square root of a variance, so a small fractional error on S^2 becomes half as large on S ($\delta S/S \approx \frac{1}{2}\,\delta S^2/S^2$, the error propagation of Chapter 3). For roughly Gaussian estimates, (7.97) with $N = 100$ gives $\frac{1}{2}\sqrt{2/99} = 7.1\%$. The claim is right, with little room to spare.

What the simulation shows. Figure 7.31(a) follows the Monte Carlo variance, in units of $2C_\ell^2/(2\ell+1)$, as the number of skies grows, for $\ell = 2$, 10 and 100, with the $\pm1\sigma$ band of (7.101) shaded. Early on the estimates wander by tens of percent (most visibly at $\ell = 2$, whose band is the widest); as N grows they narrow with their bands and settle onto 1. At $N = 20{,}000$ the ratios at $\ell = 2$, 10, 50 and 300 are 1.021, 1.000, 1.015 and 0.998, against predicted errors of 0.01483 at $\ell = 2$ and 0.01005 at $\ell = 300$: all within 1.5 standard deviations.

Panel (a) shows one run. To test the *rate* $N^{-1/2}$ we need the scatter of the Monte Carlo variance over many runs, which Figure 7.31(b) measures directly. We cut a long run into many independent batches of N skies, compute S^2 in each batch, and take the rms of $S^2/\sigma^2 - 1$ over batches. For $\ell = 2$ and $\ell = 10$ the script draws an extra 10^6 skies (cheap, since these multipoles need only 5 and 21 numbers per sky); for $100 \le \ell \le 300$ it pools the 201 independent multipoles of the main run. On logarithmic axes the points fall on straight lines of slope -0.5011 ($\ell = 2$) and -0.5133 (high ℓ), on top of (7.101). The quadrupole points lie above the Gaussian line $\sqrt{2/(N-1)}$ by the predicted factor $2.098/\sqrt{2} \approx 1.5$; the high-$\ell$ points lie on it.

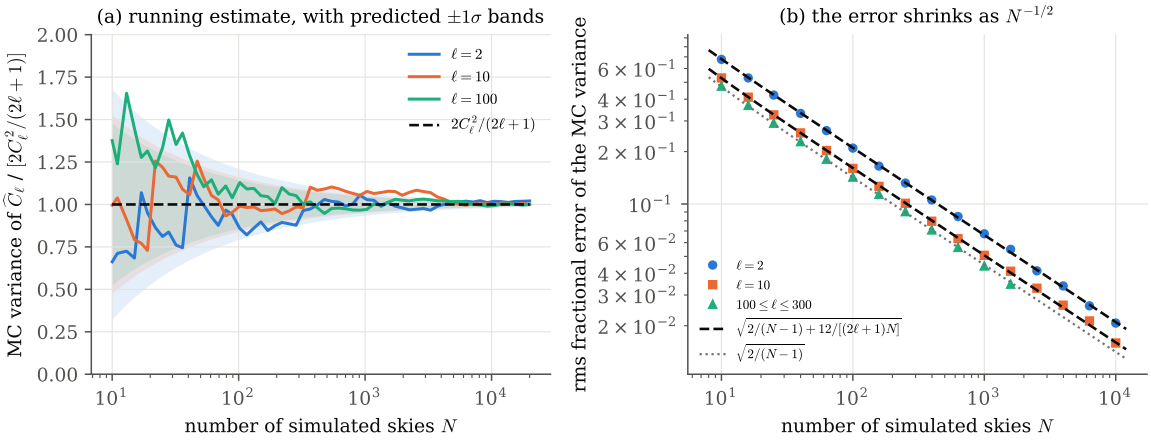


Figure 7.31: (a) Running Monte Carlo estimate of $\mathrm{Var}\,\widehat{C}_\ell$ divided by the cosmic variance $2C_\ell^2/(2\ell+1)$, for $\ell = 2$, 10, 100, with the predicted $\pm1\sigma$ bands of (7.101). (b) Measured rms fractional error of the Monte Carlo variance against the number of skies N per batch. Black dashed: (7.101) for $\ell = 2$ (upper) and $\ell = 10$ (lower); grey dotted: the Gaussian $\sqrt{2/(N-1)}$. All lines have slope $-1/2$.

7.12.4 Different multipoles are independent

The third question is about pairs of multipoles. On the full sky, $\widehat{C}_\ell$ is a function of the $2\ell+1$ real numbers at multipole ℓ only, and $\widehat{C}_{\ell'}$ of those at ℓ' only. By (7.62) and Gaussianity, the two sets are independent, so $\widehat{C}_\ell$ and $\widehat{C}_{\ell'}$ are independent for $\ell \neq \ell'$ and their covariance matrix is diagonal,

$$\operatorname{Cov}\big(\widehat{C}_\ell, \widehat{C}_{\ell'}\big) = \frac{2C_\ell^2}{2\ell+1}\,\delta_{\ell\ell'}. \tag{7.102}$$

This is what made the binning of (7.79) legitimate, and it is the property that a mask destroys (Section 7.13).

A covariance estimated from N skies is never exactly diagonal, so we need to know what "zero" looks like. The answer: the estimated correlation of two independent quantities scatters around zero with standard deviation $1/\sqrt{N}$. To see why, take two independent variables, standardise them to $u = (\widehat{C}_\ell - C_\ell)/\sigma_\ell$ and $v = (\widehat{C}_{\ell'} - C_{\ell'})/\sigma_{\ell'}$, and look at the sample correlation coefficient. For large N the corrections from estimating the means and standard deviations are of relative order $1/N$, so $r \approx \frac{1}{N}\sum_i u_i v_i$, and

$$\operatorname{Var} r \overset{(1)}{=} \frac{1}{N^2}\sum_i \mathbb{E}\big[u_i^2 v_i^2\big] \overset{(2)}{=} \frac{1}{N^2}\,N\;\mathbb{E}[u^2]\;\mathbb{E}[v^2] \overset{(3)}{=} \frac{1}{N}. \tag{7.103}$$

(1) The skies are independent and $\mathbb{E}[uv] = 0$, so the cross terms $i \neq j$ vanish.
(2) u and v are independent, so the mean of the product is the product of the means.
(3) Both are standardised.

No step used Gaussianity: the skewed χ^2_5 of the quadrupole gives the same $1/\sqrt{N}$ as anything else. So an estimated correlation matrix of truly independent quantities should look like the identity plus noise of standard deviation $1/\sqrt{N}$ off the diagonal.

Figure 7.32(a) shows the correlation matrix of $\widehat{C}_\ell$ for $2 \le \ell \le 41$ estimated from the first 500 skies. It is the identity plus a speckle of positive and negative values. Their standard deviation is 0.04741, against $1/\sqrt{500} = 0.04472$. (The 780 distinct pairs share variables, so their spread is itself uncertain by a few percent.) With all 20,000 skies the spread is 0.007064 against $1/\sqrt{N} = 0.007071$, and Figure 7.32(b) shows the $N^{-1/2}$ law over two and a half decades. Over all 44551 pairs with $2 \le \ell < \ell' \le 300$ the largest correlation is 0.02929, about $4.1/\sqrt{N}$, the size one expects for the largest of so many Gaussian deviates. The script computes the matrix with `np.corrcoef`:

Listing 7.9: Off-diagonal correlations of $\widehat{C}_\ell$ from the first N skies (`code/ch07/25_mc_cosmic_variance.py`)

```
130 def corr_offdiag(N):
131     R = np.corrcoef(y[:N, jc], rowvar=False)
132     off = R[~np.eye(len(LC), dtype=bool)]
133     return R, off
134 
135 R500, off500 = corr_offdiag(500)
136 Rall, offall = corr_offdiag(N_SIM)
137 Nc = np.unique(np.round(np.geomspace(50, N_SIM, 14)).astype(int))
138 sd_off = np.array([corr_offdiag(N)[1].std() for N in Nc])
139 # all multipoles l=2..300 with the full run: the largest off-diagonal element
140 Rbig = np.corrcoef(y, rowvar=False)
141 off_big = Rbig[~np.eye(len(ELLS), dtype=bool)]
```

Line 131 estimates the 40×40 correlation matrix from the first N rows; line 132 keeps the off-diagonal elements; lines 140–141 repeat it for all 299 multipoles with every sky.

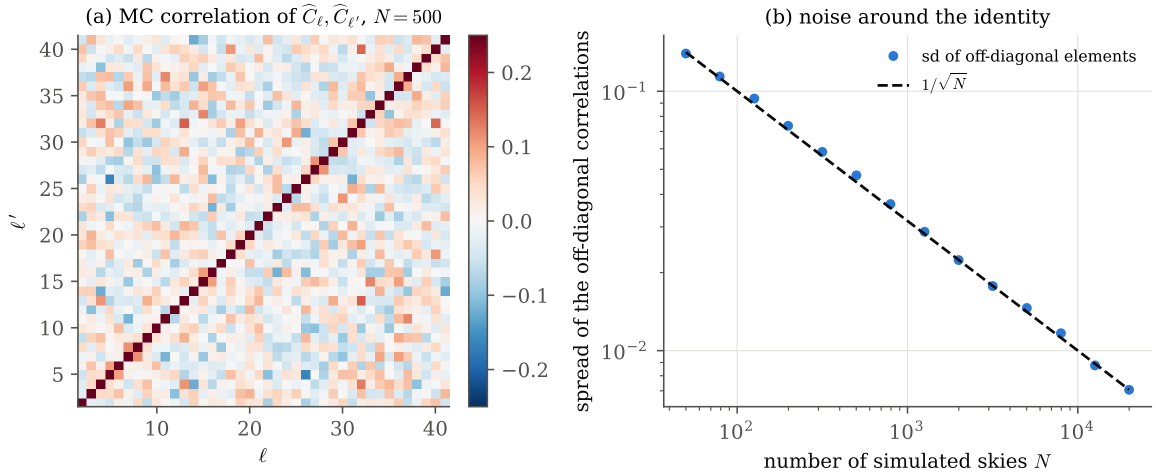


Figure 7.32: (a) Correlation matrix of $\widehat{C}_\ell$ and $\widehat{C}_{\ell'}$ for $2 \le \ell, \ell' \le 41$, estimated from 500 simulated skies (the colour scale saturates at ±0.25, so the diagonal 1 appears as the darkest red). Off the diagonal there is only noise. (b) Standard deviation of the off-diagonal elements against the number of skies, with the prediction $1/\sqrt{N}$ of (7.103) (black dashed).

Cross-check with healpy. Does the library's generator give the same skies as our from-scratch one? The script draws 3000 skies up to $\ell = 50$ with `healpy.synalm` and measures them with `healpy.alm2cl`. Two-sample Kolmogorov–Smirnov tests against our own skies give $p = 0.5166$, 0.998 and 0.9504 at $\ell = 2$, 10 and 50, and the healpy variances in units of the cosmic variance are 0.973 ($\ell = 2$) and 0.996 ($\ell = 50$), within the $\pm\sqrt{4.4/3000} \approx 4\%$ and ±2.6% that (7.101) allows for 3000 skies. The two generators produce the same skies, statistically.

How this is used in real research: Verification here, research instrument next

These simulations *verify* a result we already had in closed form: the $\chi^2_{2\ell+1}$ distribution, cosmic variance $2C_\ell^2/(2\ell+1)$, and a diagonal covariance. They test the code and our understanding where the answer is known, and they tell us how many simulations buy how much precision (7.101). In research the same loop (draw skies with `synfast`, push them through the real pipeline, measure with `anafast`) is the instrument itself. A real sky is masked, noisy, beamed and filtered; the pseudo-spectrum is then biased by mode coupling, its covariance is not diagonal, and no formula gives either exactly. MASTER calibrates the transfer function of the pipeline and the covariance of the estimate with exactly such simulations [Hivon §2–3], [168]; Planck's likelihoods are validated on thousands of simulated skies. With (7.101) we know in advance that 1000 simulations determine a variance to about 4.5% and a standard deviation to about 2.2%. Chapter 8 builds that pipeline.

7.13 What C_ℓ does not see

The estimator $\widehat{C}_\ell$ keeps one number per multipole, the mean of $|a_{\ell m}|^2$ over m. It throws away the phase of every $a_{\ell m}$ and the way the power is shared among the m. Is anything lost? For a Gaussian, isotropic sky, no. The covariance (7.62) is then the whole distribution, the phases are uniformly random, and Chapter 3 showed that $\widehat{C}_\ell$ is sufficient for C_ℓ. For any other sky, yes: the phases carry the shapes. In flat space (Section 7.7), randomising the Fourier phases of a non-Gaussian field left $P(k)$ untouched and erased every structure in it. The same experiment

on the sphere takes a few lines.

Example: hot spots and their phase-scrambled twin. Can two skies with the same $\widehat{C}_\ell$ look completely different? Paint 40 Gaussian hot spots of width 4° at random positions on a sky and compute its coefficients. Then build a second sky with the same $|a_{\ell m}|$ but new phases. For $m > 0$ each $a_{\ell m}$ is multiplied by a random unit complex number, and each real $a_{\ell 0}$ by a random sign (the only phase a real number can have). The reality condition (7.46) is respected automatically, because only $m \geq 0$ is stored and the negative m are rebuilt from it. By construction both skies have the same $\widehat{C}_\ell$ at every ℓ; numerically they agree to 5.322×10^{-16}. Figure 7.33 shows how different they look: isolated spots on a flat background, against a featureless Gaussian-looking pattern. The pixel histograms in Figure 7.34(b) put a number on the difference: the spot sky has skewness 2.903, its twin 0.01184. A test based on $\widehat{C}_\ell$ alone could never tell them apart. (The experiment is `code/ch07/26_phases_sphere.py`.)

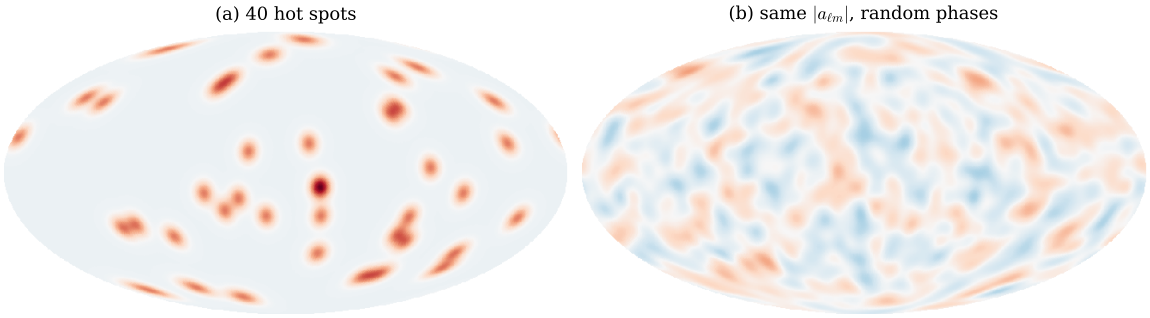


Figure 7.33: (a) A sky made of 40 hot spots, band-limited to $\ell \leq 255$. (b) The same $|a_{\ell m}|$ with random phases. Same colour scale in both panels. The two skies have identical $\widehat{C}_\ell$ at every multipole.

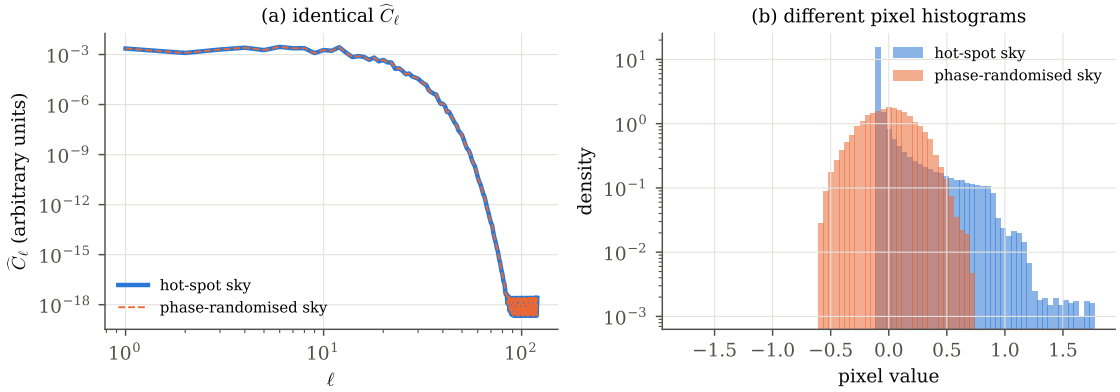


Figure 7.34: (a) $\widehat{C}_\ell$ of the two skies of Figure 7.33 up to $\ell = 120$ (the 4° spots have essentially no power beyond): the curves lie on top of each other. (b) Histograms of their pixel values (logarithmic vertical axis): the spot sky has a long positive tail, the phase-randomised sky is symmetric and close to Gaussian.

Two things break the clean full-sky picture of this chapter. The first is the sky itself. Primordial non-Gaussianity ($f_{\rm NL}$, Chapter T1), lensing and foregrounds put information into the phases, where C_ℓ cannot see it; Chapter 12 builds summaries that can, from the shapes of hot and cold regions. The second is the observation: a real map covers only part of the sky. Let $W(\boldsymbol{e})$ be the mask, 1 where we observe and 0 elsewhere. The coefficients of the masked sky are $\tilde{a}_{\ell m} = \int \mathrm{d}\Omega\, \Theta\, W\, Y^*_{\ell m}$ [Hivon (11)]. Because the mask multiplies the sky, these coefficients

mix neighbouring multipoles, just as multiplying by $1 + A\cos\theta$ did in Section 7.9.5. Their mean square is then a weighted sum of the true spectrum, $\langle \widetilde{C}_\ell \rangle = \sum_{\ell'} M_{\ell\ell'} C_{\ell'}$ [Hivon (14)]. The covariance (7.102) is no longer diagonal, the $\chi^2_{2\ell+1}$ law is gone, and the Monte Carlo loop of Section 7.12 becomes the only way to get the sampling distribution. That is where Chapter 8 begins.

Recap

- The spherical harmonics $Y_{\ell m}$ (eigenfunctions of $\boldsymbol{L}^2$ and L_z) are the Fourier modes of the sphere; multipole ℓ is an angular scale of about $180^\circ/\ell$. A sky is $\Theta = \sum a_{\ell m} Y_{\ell m}$ with $a_{\ell m} = \int \mathrm{d}\Omega\, \Theta Y^*_{\ell m}$.
- A real sky has $a_{\ell,-m} = (-1)^m a^*_{\ell m}$: $a_{\ell 0}$ is real, and there are $2\ell + 1$ real numbers per ℓ.
- Statistical isotropy makes $\hat{C}$ commute with all rotations; with $\boldsymbol{L}^2$, L_z and $L_\pm$ this gives $\langle a_{\ell m} a^*_{\ell' m'} \rangle = C_\ell \delta_{\ell\ell'} \delta_{mm'}$. $C(\vartheta) = \sum (2\ell+1) C_\ell P_\ell(\cos\vartheta)/4\pi$. A preferred axis breaks both the diagonality and the m-independence.
- $\widehat{C}_\ell = \sum_m |a_{\ell m}|^2/(2\ell+1)$ is unbiased; $(2\ell+1)\widehat{C}_\ell/C_\ell \sim \chi^2_{2\ell+1}$ because $a_{\ell 0}$ has variance C_ℓ and $\mathrm{Re}, \mathrm{Im}\, a_{\ell m}$ have $C_\ell/2$. Cosmic variance: $\mathrm{Var}\, \widehat{C}_\ell = 2C_\ell^2/(2\ell+1)$, 63% at the quadrupole. Read backwards, the same density is the likelihood $\mathcal{L}(C_\ell) \propto C_\ell^{-(2\ell+1)/2} e^{-(2\ell+1)\widehat{C}_\ell/2C_\ell}$.
- CAMB solves the linear Boltzmann hierarchy and returns $C_\ell = \frac{2}{\pi}\int k^2 \mathrm{d}k\, P_\zeta g^2_{T\ell}$, an expectation, never a sky. HEALPix has $12N^2_{\mathrm{side}}$ equal-area pixels on rings; healpy's `synfast`/`anafast` go between C_ℓ, $a_{\ell m}$, maps and $\widehat{C}_\ell$, accurately below $\ell \approx 2N_{\mathrm{side}}$.
- Monte Carlo with N skies reproduces the χ^2, converges to the cosmic variance with fractional error $\sqrt{2/(N-1) + 12/[(2\ell+1)N]}$, and gives a correlation matrix equal to the identity plus $1/\sqrt{N}$ noise. This verifies a known result; Chapter 8 uses the same loop where no formula exists.
- $\widehat{C}_\ell$ ignores the phases of the $a_{\ell m}$: skies with the same $\widehat{C}_\ell$ can look nothing alike.

Reading the literature

- **Random fields and their two-point statistics.** Hobson et al. (eds.), *Bayesian Methods in Cosmology*: the pages on parameterising a field by pixels or by Fourier modes, where homogeneity makes the Fourier covariance diagonal and a window couples the modes, and Chapter 10 (Jaffe) on the CMB as a hierarchical model, with the harmonic expansion, the power spectrum, the Wiener filter and the full-sky likelihood [HobsonBMC pp. 229–238]. Cowan, *Statistical Data Analysis*, Chapter 1, for covariance matrices and their diagonalisation by an orthogonal transformation; Wasserman, *All of Statistics*, Chapter 3, for covariance and correlation.
- **Gaussian random fields in cosmology.** Bardeen, Bond, Kaiser and Szalay on the statistics of peaks of Gaussian fields [111]; Peacock's *Cosmological Physics* [164] for power spectra, correlation functions and windows; Kaiser on clustering in redshift space [108].
- **The CMB power spectrum and cosmic variance.** Knox on the errors of C_ℓ from noise, beam, sky coverage and cosmic variance [166]; Hivon et al., MASTER, Section 2 for the harmonic expansion, the full-sky estimator and its χ^2, and the pseudo-spectrum of a masked sky [168].
- **Computing C_ℓ.** Seljak and Zaldarriaga on line-of-sight integration [167]; Lewis, Challinor and Lasenby, the CAMB paper [8]; Padilla et al. for a survey of Boltzmann and sampling

codes [40]; the Planck 2018 parameters that define the fiducial model [25], and the Planck tests of isotropy [114]. For the physics behind C_ℓ (the Boltzmann hierarchy, transfer functions, acoustic peaks), see the cosmology notes and Chapter T1.

- **Pixelising the sphere.** Górski et al. on HEALPix [7].

Chapter summary

A physical theory of fluctuations predicts the statistics of a field, never the field itself, and we observe one realisation. Homogeneity and isotropy make the natural modes of the field uncorrelated: plane waves in flat space, spherical harmonics on the sky. Their variances are $P(k)$ or C_ℓ, and for a Gaussian field those variances are everything. One realisation offers only a finite number of independent modes at each scale: $N_{\rm modes}$ in a k-shell, or $2\ell+1$ at a multipole. That number sets an error $\sqrt{2/N_{\rm modes}}$ on every measured power, and no instrument can reduce it. In the pipeline

$$\text{physical model} \longrightarrow \text{field or dataset} \longrightarrow \text{summary } S \longrightarrow \boxed{P(S \mid \theta)} \longrightarrow \text{infer physics},$$

random-field theory supplies the first summaries of a field, $\xi(r)$, $P(k)$ and C_ℓ, and their exact sampling distributions in the ideal case of a complete, noiseless field; Monte Carlo confirms them. Chapter 8 adds the mask, the beam and the noise, Chapters 9 and 10 turn the sampling distribution into parameter constraints, and Chapter 12 looks at what the two-point summaries throw away.

Random fields in flat space

concept	operational meaning	used later in
random field	a random variable at every point, with a joint distribution for any finite set of points; theory predicts its statistics, we see one draw	Chapters 8 and 12
homogeneity, isotropy	all joint distributions unchanged by translations (rotations); the two-point function depends only on the separation	Chapters 8 and 12
correlation function $\xi(r)$	$\langle\delta(\boldsymbol{x})\delta(\boldsymbol{x}+\boldsymbol{r})\rangle$: how strongly the field at two points a distance r apart is related	Chapters 10 and 12
Gaussian random field	all joint distributions Gaussian; mean and $\xi(r)$ determine everything	Chapters 8, 9 and 12
power spectrum $P(k)$	variance of each Fourier mode; Fourier modes of a homogeneous field are uncorrelated; P is the Fourier transform of ξ (Wiener–Khinchin)	Chapters 8 and 10
simulating a field	colour white noise by $\sqrt{P(k)}$ in Fourier space and transform back (FFT)	Chapters 8 and 12
ergodicity, sample variance	a spatial average replaces the ensemble average when distant regions are nearly independent; a k-shell with $N_{\rm modes}$ modes gives $\Delta P/P = \sqrt{2/N_{\rm modes}}$	Chapters 8 and 10
smoothing	a beam, pixel or counting cell multiplies $P(k)$ by $\lvert W(k)\rvert^2$	Chapters 8 and 12
spectral moments σ_j	integrals of $k^{2j}P(k)$ (with the window): the variance of the field, of its gradient and the typical spacing of its features	Chapter 12
phases	$P(k)$ keeps the amplitudes and discards the phases; phases hold the shapes of non-Gaussian fields	Chapter 12

Fields on the sphere

concept	operational meaning	used later in
spherical harmonics, $a_{\ell m}$	eigenfunctions of $\boldsymbol{L}^2$, L_z; $a_{\ell m} = \int \mathrm{d}\Omega\, \Theta Y^*_{\ell m}$; multipole $\ell \leftrightarrow$ angle $180°/\ell$	Chapters 8 and 12
reality condition	$a_{\ell,-m} = (-1)^m a^*_{\ell m}$; $a_{\ell 0}$ real; $2\ell+1$ real numbers per ℓ, $(\ell_{\max}+1)^2$ in all; store $m \geq 0$ only	Chapter 8
isotropy in harmonic space	$\langle a_{\ell m} a^*_{\ell' m'} \rangle = C_\ell \delta_{\ell\ell'} \delta_{mm'}$, from $[\hat{C}, \boldsymbol{L}] = 0$ and the ladder operators; a preferred axis couples ℓ to $\ell \pm 1$	Chapters 8 and 9
$C(\vartheta)$ and C_ℓ	$C(\vartheta) = \sum_\ell \frac{2\ell+1}{4\pi} C_\ell P_\ell(\cos\vartheta)$, $C_\ell = 2\pi \int \mathrm{d}\cos\vartheta\, P_\ell C$; point variance $\sum(2\ell+1)C_\ell/4\pi$	Chapters 8 and 12
full-sky estimator $\widehat{C}_\ell$	mean of $\lvert a_{\ell m}\rvert^2$ over m; unbiased, sufficient, maximum likelihood	Chapters 8 to 10
$\chi^2_{2\ell+1}$ law	$(2\ell+1)\widehat{C}_\ell/C_\ell \sim \chi^2_{2\ell+1}$: one real $a_{\ell 0}$ of variance C_ℓ, 2ℓ real and imaginary parts of variance $C_\ell/2$; skewed at low ℓ	Chapters 8 and 9
cosmic variance	$\mathrm{Var}\,\widehat{C}_\ell = 2C_\ell^2/(2\ell+1)$ (the Cramér–Rao bound); binned: divide by $\Delta\ell$; with noise and sky fraction, $(C_\ell + N_\ell/B_\ell^2)\sqrt{2/[(2\ell+1)\Delta\ell f_{\mathrm{sky}}]}$	Chapters 8 and 10
likelihood of C_ℓ	$\mathcal{L} \propto C_\ell^{-(2\ell+1)/2} \exp[-(2\ell+1)\widehat{C}_\ell/2C_\ell]$: broad and lopsided at low ℓ	Chapters 9 and 10
CAMB	solves the linear Boltzmann hierarchy, line of sight; $C_\ell = \frac{2}{\pi}\int k^2 \mathrm{d}k\, P_\zeta \mathcal{G}^2_{T\ell}$; call with `raw_cl=True`, $\mu\mathrm{K}^2$	Chapters 8 to 10
HEALPix and healpy	$N_{\mathrm{pix}} = 12N^2_{\mathrm{side}}$ equal areas on rings, pixel $58.6°/N_{\mathrm{side}}$; `synalm`, `alm2map`, `map2alm`, `alm2cl`; trust $\ell \lesssim 2N_{\mathrm{side}}$; pixel window	Chapters 8 and 12
Monte Carlo of $\widehat{C}_\ell$	simulate skies, histogram $\widehat{C}_\ell$; MC variance error $\sqrt{2/(N-1) + 12/[(2\ell+1)N]}$; off-diagonal correlations $\pm 1/\sqrt{N}$	Chapters 8 and 9
verification vs research	here the simulation checks a closed-form answer; with a mask, noise and a beam it is the only way to the sampling distribution	Chapter 8
what C_ℓ misses	the phases of the $a_{\ell m}$: equal $\widehat{C}_\ell$, different shapes and pixel histograms	Chapter 12

Chapter 8

Monte Carlo as a Research Instrument: Observing the CMB

Where this chapter is going

A Boltzmann code such as CAMB turns six cosmological numbers into an angular power spectrum C_ℓ (Chapter T1). That spectrum describes the *ensemble* of skies the theory could have produced. We observe a single member of that ensemble through a telescope whose beam blurs the sky. Pixelisation averages the signal further, detector noise enters the map, and masking the Galaxy removes part of the sky from the analysis. In Chapter 7, where we had a complete, noiseless sky, the spectrum estimator $\widehat{C}_\ell$ was unbiased and followed a scaled $\chi^2_{2\ell+1}$ distribution with variance $2C_\ell^2/(2\ell+1)$. To see which of these properties survive the observation, we can simulate skies with a known spectrum, pass them through the same observing procedure, and measure $\widehat{C}_\ell$ on each resulting map.

In the pipeline of the book (physical model → field → what matters → summary → sampling distribution → inference) the problem sits on one arrow: finding the sampling distribution of the summary $\widehat{C}_\ell$ under realistic observation. The beam and the noise come first, on the full sky, where every answer can still be derived by hand. There the simulations *verify* known formulas, and the same machine is tested and trusted. The mask comes next: it couples multipoles, correlates their errors and bends the distribution of $\widehat{C}_\ell$ away from any formula, and the simulations become the research instrument (Figure 8.1). The output, a mean, a covariance and a distribution shape for $\widehat{C}_\ell$, is exactly what the likelihood of Chapter 9, the Fisher forecast of Chapter 10 and the topological summaries of Chapter 12 need.

8.1 The research problem

Picture the data as they arrive. A detector sweeps across the sky and records a voltage many times per second. Every sample is the temperature of a small patch of sky, averaged over the response pattern of the telescope, plus noise from the detector and its electronics. Months of samples are binned into a map of a few million pixels. A theorist, meanwhile, has run CAMB and holds a curve C_ℓ. The question is how to put the map and the curve on the same plot, and how big the error bars on that plot really are. Answering it for the cosmic microwave background (CMB) is the research problem that organises everything below:

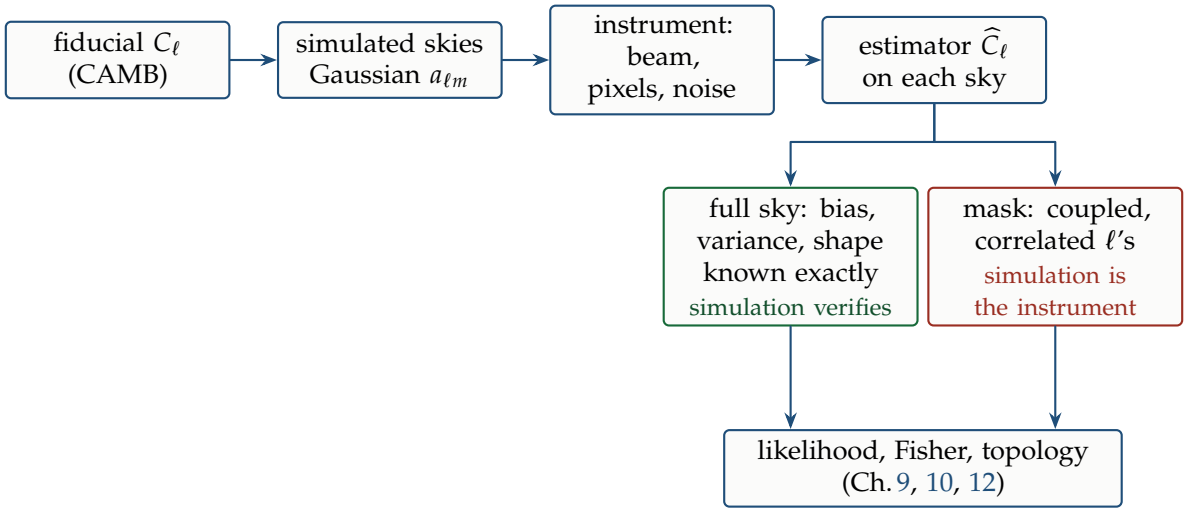


Figure 8.1: The research problem. A fiducial spectrum is turned into many simulated skies, each is observed by a model of the instrument, and the spectrum is estimated on each. On the full sky (green) the statistics of the estimator follow from a short derivation, and the simulations check it. Once the sky is masked (red) no formula is exact, and the simulations are the only way to the mean, the covariance and the shape of the distribution that the later chapters feed into a likelihood.

The central framing: The research problem

A theoretical CMB power spectrum C_ℓ, obtained from a fiducial cosmological model, describes the ensemble-averaged statistics of the CMB sky. In practice, however, CMB observations provide only a finite and potentially incomplete realization of this random field. Sky masks, instrumental noise, beam effects, and other observational limitations modify the statistical properties of the quantities extracted from the map.

The problem is to use Monte Carlo simulations of CMB skies generated from a fiducial theoretical power spectrum to investigate how these observational effects influence the statistical properties of CMB power-spectrum estimators. In particular, we seek to determine the sampling distribution, bias, variance, and covariance of the recovered spectrum and to assess the validity of commonly used analytical approximations for these quantities.

The study will begin with the full-sky Gaussian case, for which the statistical properties of $\widehat{C}_\ell$ are known analytically, and then introduce partial-sky coverage through realistic or controlled masks. The simulated realizations will be used to quantify the resulting mode coupling and correlations between multipoles, and to determine how these effects depend on the sky fraction, mask geometry, and multipole range.

The central computational question is therefore:

> **How does incomplete and imperfect observation of the CMB alter the statistical information contained in the measured angular power spectrum, and how accurately can these effects be characterized using Monte Carlo simulations?**

The statement mixes several questions, and each needs a different tool. Before any equation it helps to say them one at a time, in words, in the order in which they depend on each other. The first is physical: what exactly does the instrument do to the sky? Only when that is written down as a model can we ask what it does to the *average* of the estimator (does $\widehat{C}_\ell$ still centre on C_ℓ?), then to its *scatter* (how wide is the spread over skies, and are neighbouring multipoles correlated?), and then to the *shape* of its distribution (is it still a scaled χ^2, or close enough to a Gaussian?). The last two questions are about the method itself: do the quick formulas that

everyone uses survive the comparison with simulations, and how many simulations does it take to know?

In order, the questions are:

1. What does the instrument do to the sky? (A beam, a pixel window, noise, a mask: the observation model.)

2. Does the estimator still average to the true C_ℓ? If not, by how much is it off, and can the offset be removed? (Bias.)

3. How much does it scatter from sky to sky, and do the errors at different multipoles move together? (Variance and covariance.)

4. What is the full shape of its distribution? (The sampling distribution: the likelihood of Chapter 9 needs it, not just its width.)

5. Are the standard analytic shortcuts, such as the Knox formula with a sky fraction $f_{\rm sky}$, accurate, and where do they fail?

6. How many simulated skies are needed to answer 2–5 to a given precision?

On the full sky with a beam and noise, questions 1–4 have exact answers, which we derive below. That makes the full sky the place to build the simulation machine and to check, number by number, that it reproduces what we already know. A machine that has passed that test can then be trusted where the answers are not known.

8.1.1 Why we do not simply maximise the likelihood

The procedure of Chapter 3 for finding an estimator starts from the likelihood: write the probability of the data given the parameter, maximise it, read the error bar off its curvature. For a CMB map the data model is

$$\boldsymbol{d} = \mathsf{A}\,\boldsymbol{s} + \boldsymbol{n}, \tag{8.1}$$

where $\boldsymbol{s}$ is the sky in pixels, the matrix A applies the beam and the pointing, and $\boldsymbol{n}$ is Gaussian noise with covariance matrix N [HobsonBMC (10.8)–(10.9)]. Here the observed outcome is the whole map and the hypotheses are the possible spectra C_ℓ; each spectrum is judged by how plausible the random sky-plus-noise fluctuation it would need to produce this particular map is, the reading set out for a single multipole in Section 8.3.1. The sky is a Gaussian random field whose pixel covariance $\mathsf{S}(C_\ell)$ is fixed by the spectrum, and the noise is independent of it, so the map is Gaussian with covariance $\mathsf{S}+\mathsf{N}$, the sum of the two covariances. Its density is the multivariate Gaussian of Chapter 2, $(2\pi)^{-N_{\rm pix}/2}\det(\mathsf{S}+\mathsf{N})^{-1/2}\exp[-\frac{1}{2}\boldsymbol{d}^{\mathsf{T}}(\mathsf{S}+\mathsf{N})^{-1}\boldsymbol{d}]$, and minus twice its logarithm, read as a function of C_ℓ, is

$$-2\ln\mathcal{L}(C_\ell) = \boldsymbol{d}^{\mathsf{T}}(\mathsf{S}+\mathsf{N})^{-1}\boldsymbol{d} + \ln\det(\mathsf{S}+\mathsf{N}) + \text{const} \tag{8.2}$$

[HobsonBMC (10.23)], [134].

What does maximising this over C_ℓ give? Write $\mathsf{C} \equiv \mathsf{S}+\mathsf{N}$ for the total covariance. The covariance of the sky between two pixels is a sum over ℓ of C_ℓ times a fixed angular pattern (Chapter 7), so S is linear in the spectrum, $\mathsf{S} = \sum_\ell C_\ell\,\mathsf{S}_\ell$, where $\mathsf{S}_\ell \equiv \partial\mathsf{S}/\partial C_\ell$ is the pixel covariance that one unit of C_ℓ produces. Two matrix rules do the differentiation: the derivative of an inverse, $\partial\mathsf{C}^{-1} = -\mathsf{C}^{-1}(\partial\mathsf{C})\,\mathsf{C}^{-1}$, and the derivative of a log-determinant,

$\partial \ln\det \mathsf{C} = \mathrm{tr}\big(\mathsf{C}^{-1}\partial\mathsf{C}\big)$. With $\partial\mathsf{C}/\partial C_\ell = \mathsf{S}_\ell$ and the shorthand $\mathsf{E}_\ell \equiv \mathsf{C}^{-1}\mathsf{S}_\ell\,\mathsf{C}^{-1}$,

$$\frac{\partial \ln \mathcal{L}}{\partial C_\ell} \overset{(1)}{=} \frac{1}{2}\Big[\boldsymbol{d}^{\mathsf{T}}\mathsf{E}_\ell\,\boldsymbol{d} - \mathrm{tr}\big(\mathsf{C}^{-1}\mathsf{S}_\ell\big)\Big] \overset{(2)}{=} \frac{1}{2}\Big[\boldsymbol{d}^{\mathsf{T}}\mathsf{E}_\ell\,\boldsymbol{d} - \mathrm{tr}\big(\mathsf{E}_\ell\mathsf{N}\big)\Big] - \sum_{\ell'} F_{\ell\ell'}C_{\ell'},$$

$$\widehat{C}_\ell^{\,\mathrm{QML}} \overset{(3)}{=} \frac{1}{2}\sum_{\ell'}\big(F^{-1}\big)_{\ell\ell'}\Big[\boldsymbol{d}^{\mathsf{T}}\mathsf{E}_{\ell'}\,\boldsymbol{d} - \mathrm{tr}\big(\mathsf{E}_{\ell'}\mathsf{N}\big)\Big]. \tag{8.3}$$

(1) Differentiate (8.2) with the two rules and multiply by $-\tfrac{1}{2}$.
(2) Insert $\mathsf{C}\,\mathsf{C}^{-1}$ into the trace, $\mathrm{tr}(\mathsf{C}^{-1}\mathsf{S}_\ell) = \mathrm{tr}(\mathsf{E}_\ell\mathsf{C})$, and split $\mathsf{C} = \sum_{\ell'} C_{\ell'}\mathsf{S}_{\ell'} + \mathsf{N}$. The sky part gives $\sum_{\ell'} C_{\ell'}\,\mathrm{tr}(\mathsf{E}_\ell\mathsf{S}_{\ell'}) = 2\sum_{\ell'} F_{\ell\ell'}C_{\ell'}$, with $F_{\ell\ell'} = \tfrac{1}{2}\,\mathrm{tr}(\mathsf{C}^{-1}\mathsf{S}_\ell\mathsf{C}^{-1}\mathsf{S}_{\ell'})$ the Fisher matrix of the spectrum (Chapter 3).
(3) Set the slope to zero and solve for the $C_{\ell'}$. The matrices E_ℓ and F still contain C_ℓ, so they are evaluated at a fiducial spectrum, and the solution is one Newton step from it; repeating the step with the result as the new fiducial converges to the maximum.

The estimator (8.3) is *quadratic in the data*, because the data enter only through $\boldsymbol{d}^{\mathsf{T}}\mathsf{E}_\ell\boldsymbol{d}$. Read from the inside out, it weights the map by the inverse covariance, correlates it with the pattern S_ℓ that a change in one C_ℓ would produce, subtracts the part of that correlation the noise alone would give, $\mathrm{tr}(\mathsf{E}_\ell\mathsf{N})$, and normalises by the Fisher matrix. This is the **quadratic maximum-likelihood** (QML) estimator [134]. On a complete sky with uniform noise every matrix is diagonal in the harmonic coordinates (the real numbers u_k of (8.21)): C has $V_\ell = T_\ell^2 C_\ell + N_\ell$ on the $\nu = 2\ell+1$ modes of each ℓ, and S_ℓ has T_ℓ^2 on them. Then $\boldsymbol{d}^{\mathsf{T}}\mathsf{E}_\ell\boldsymbol{d} = T_\ell^2\nu\,\widehat{C}_\ell^{\,\mathrm{obs}}/V_\ell^2$, $\mathrm{tr}(\mathsf{E}_\ell\mathsf{N}) = T_\ell^2\nu N_\ell/V_\ell^2$ and $F_{\ell\ell} = \nu T_\ell^4/2V_\ell^2$, and (8.3) collapses to $(\widehat{C}_\ell^{\,\mathrm{obs}} - N_\ell)/T_\ell^2$, with the fiducial V_ℓ cancelling: the average of $|a_{\ell m}|^2$ over m with the noise subtracted, which Section 8.3.1 obtains directly.

The trouble is the size of the matrices. Every evaluation of (8.2) needs the inverse and the determinant of an $N_{\mathrm{pix}} \times N_{\mathrm{pix}}$ matrix, which costs of order N_{pix}^3 operations and $8N_{\mathrm{pix}}^2$ bytes of memory [HobsonBMC p. 238], [Hivon §1].

Example: how long is one likelihood evaluation? On an ordinary desktop computer, the Cholesky factor of a 6000×6000 covariance takes 0.80 s (Figure 8.2). The cost grows as N^3, so for the $N_{\mathrm{pix}} = 786432$ pixels of a HEALPix map at $N_{\mathrm{side}} = 256$ the same operation takes about 21 days, and the matrix alone occupies 4.9 TB. At $N_{\mathrm{side}} = 512$ it is about 4 years. At $N_{\mathrm{side}} = 2048$, the resolution of the Planck maps, it is 1.5×10^4 years for a single evaluation, and a parameter fit needs tens of thousands of them. A spherical-harmonic transform of the same Planck-size map, which is all the simple estimator needs, takes about 9 s, because on the iso-latitude HEALPix grid its cost grows only as $N_{\mathrm{pix}}^{3/2}$ [Hivon §2.2], [7].

Example: when is the cheap estimator as good as the expensive one? Two limits are known exactly [134]. When the signal dominates the noise, the cheap estimator with equal weight on every pixel is statistically equivalent to the maximum-likelihood one. When the noise dominates and is uncorrelated between pixels, the cheap estimator with each pixel weighted by its inverse noise variance is mathematically identical to it. Between the two limits, and with a mask, it loses some information, and the loss has to be measured.

So the field took a different road, and the recipe of Chapter 3 already contains it: its step 5, for problems whose likelihood cannot be evaluated. Choose a statistic that is cheap and sensitive to C_ℓ: the average of $|a_{\ell m}|^2$ over m, computed directly on the observed map, which on a cut sky is called the **pseudo-C_ℓ** [Hivon (13)]. Then simulate many skies, pass each through the same instrument and the same analysis, and read off the mean, the scatter and the shape of the statistic from the simulations. The statistic is suboptimal, but its bias is calibrated out and

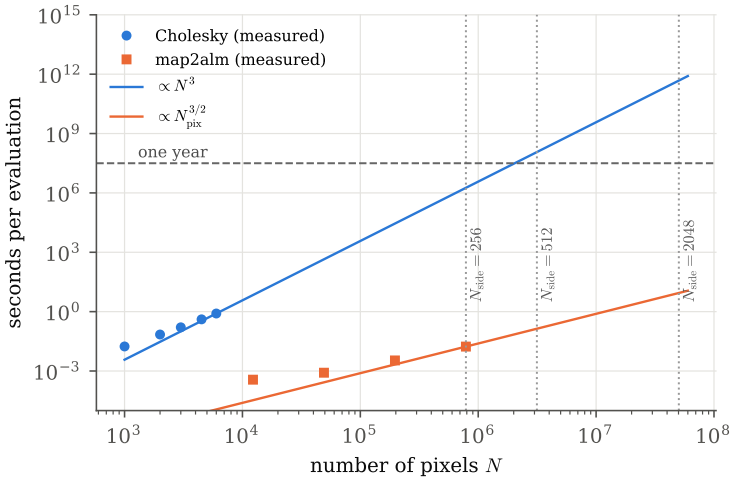


Figure 8.2: Time for one evaluation of the exact pixel likelihood (blue: a Cholesky factorisation, which grows as N^3) and for one spherical-harmonic transform (orange, growing as $N_{\mathrm{pix}}^{3/2}$). Points are measured, lines are extrapolations from the largest measured size. The dotted verticals mark three HEALPix resolutions, the dashed horizontal one year.

its error bar is honest. Hivon and collaborators named the method after this calibration step: MASTER, the Monte Carlo Apodised Spherical Transform EstimatoR [Hivon §1].

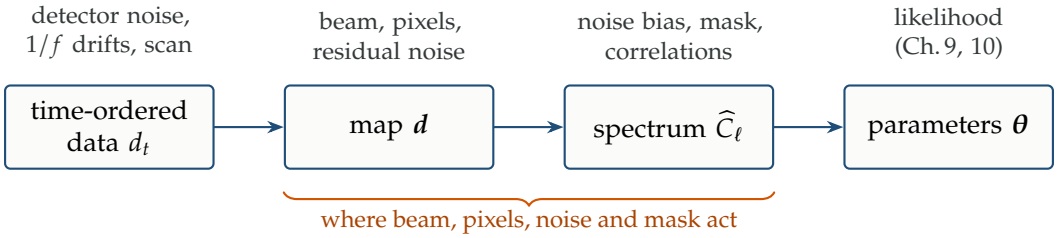


Figure 8.3: The CMB analysis as a chain of compressions [HobsonBMC §10.2]. Each step keeps the information the next one needs, and each adds its own complications (grey). The bracket marks the middle link: how the beam, the pixels and the noise of the map, and the mask, change the statistics of the spectrum computed from it.

The chain of Figure 8.3 is called a hierarchical model [HobsonBMC §10.2]: the timestream is compressed into a map, the map into a spectrum, the spectrum into parameters, and each step only needs the output of the one before. Making the map from the timestream is a least-squares problem of its own; for white noise it reduces to averaging the samples that fall in each pixel, which divides the noise variance by the number of hits [HobsonBMC (10.14)–(10.15)]. Real detector noise has extra power at low frequencies ($1/f$ noise), which leaves stripes along the scan and has to be filtered out [Hivon §2.1], [HobsonBMC p. 234]. We start from the map and assume the noise left in it is Gaussian and uncorrelated between pixels, with a variance that may change from pixel to pixel. Everything that this simplification leaves out (filtering, correlated noise, asymmetric beams) is exactly the kind of effect that real pipelines handle by putting it into the simulations [Hivon §3.2–3.3].

8.2 What the instrument does to the sky

A radio telescope pointed in direction $\hat{\boldsymbol{n}}$ collects radiation from a small cone around that direction. Its response to a point source defines the **beam**, also called the point-spread function.

This response sets the weight of each direction in the average temperature recorded by the telescope. The recorded averages are then binned into pixels, which averages them a second time over the area of each pixel. Finally every pixel carries noise from the detector. The three steps act at different places, and the order matters: the beam and the pixels act on the sky, while the noise is added afterwards and is blurred by neither.

Working definition 8.1 (observation model of a temperature map)

The **observed map** is

$$d_p = \int_{S^2} b\big(\hat{\boldsymbol{n}}_p \cdot \hat{\boldsymbol{n}}'\big)\, s(\hat{\boldsymbol{n}}')\, \mathrm{d}\Omega' + n_p \equiv (B \circledast s)(\hat{\boldsymbol{n}}_p) + n_p, \tag{8.4}$$

where $s(\hat{\boldsymbol{n}})$ is the sky temperature fluctuation in μK, $\hat{\boldsymbol{n}}_p$ is the centre of pixel p, b is the beam [the weight given to the sky at angle γ from the pointing, a function of $\cos\gamma = \hat{\boldsymbol{n}}_p \cdot \hat{\boldsymbol{n}}'$ only for a circular beam], normalised as $\int b\,\mathrm{d}\Omega = 1$, and n_p is the noise of pixel p [Gaussian, mean zero, variance σ_p^2, independent of the sky and of the noise in every other pixel]. The pixel average is included as one more smoothing of s. *Operationally:* to simulate an observation, smooth the simulated sky with the beam and the pixel, then add one independent Gaussian number to each pixel.

This is the model (8.1) written out for a single detector with a circular beam, after the timestream has been turned into a map [HobsonBMC (10.6)–(10.7)]. The same structure, data equal smoothed model plus noise, appears in the cosmology notes as $\boldsymbol{d} = \boldsymbol{s}(\boldsymbol{\theta}) + \boldsymbol{n}$ with the beam inside $\boldsymbol{s}$ (see the cosmology notes).

8.2.1 A beam multiplies each multipole by a number

Smoothing is a convolution, and on a line a convolution becomes a multiplication in Fourier space. The sphere has the same property, with spherical harmonics in place of Fourier modes and with one condition: the smoothing kernel must depend only on the angle between the two directions. To see it, expand the beam, a function of $x = \cos\gamma$ on $[-1, 1]$, in Legendre polynomials,

$$b(x) = \sum_{\ell=0}^{\infty} \frac{2\ell+1}{4\pi}\, B_\ell\, P_\ell(x), \qquad B_\ell = 2\pi \int_{-1}^{1} b(x)\, P_\ell(x)\, \mathrm{d}x. \tag{8.5}$$

The formula for B_ℓ follows by multiplying the expansion by $P_{\ell'}$ and integrating, using $\int_{-1}^{1} P_\ell P_{\ell'}\,\mathrm{d}x = 2\delta_{\ell\ell'}/(2\ell+1)$. The coefficient $B_0 = 2\pi\int b\,\mathrm{d}x = \int b\,\mathrm{d}\Omega = 1$ is the normalisation: the beam does not change the mean of the sky. Now smooth a sky $s = \sum_{\ell m} a_{\ell m} Y_{\ell m}$ (Chapter 7):

$$\begin{aligned}
(B \circledast s)(\hat{\boldsymbol{n}}) &\overset{(1)}{=} \int \sum_{\ell} \frac{2\ell+1}{4\pi} B_\ell\, P_\ell(\hat{\boldsymbol{n}} \cdot \hat{\boldsymbol{n}}')\, s(\hat{\boldsymbol{n}}')\, \mathrm{d}\Omega' \\
&\overset{(2)}{=} \sum_{\ell} B_\ell \sum_{m=-\ell}^{\ell} Y_{\ell m}(\hat{\boldsymbol{n}}) \int Y^*_{\ell m}(\hat{\boldsymbol{n}}')\, s(\hat{\boldsymbol{n}}')\, \mathrm{d}\Omega' \\
&\overset{(3)}{=} \sum_{\ell m} \big(B_\ell\, a_{\ell m}\big)\, Y_{\ell m}(\hat{\boldsymbol{n}}).
\end{aligned} \tag{8.6}$$

(1) Insert the Legendre expansion (8.5) of the beam.
(2) The addition theorem $P_\ell(\hat{\boldsymbol{n}}\cdot\hat{\boldsymbol{n}}') = \frac{4\pi}{2\ell+1}\sum_m Y_{\ell m}(\hat{\boldsymbol{n}}) Y^*_{\ell m}(\hat{\boldsymbol{n}}')$ (Chapter 7); the factors $(2\ell+1)/4\pi$ cancel.

(3) The remaining integral is the definition of the harmonic coefficient $a_{\ell m}$ of s.

Key idea: The convolution theorem on the sphere

A circular beam multiplies every harmonic coefficient of the sky by a number that depends on ℓ only: $a_{\ell m} \to B_\ell\, a_{\ell m}$. The beam does not mix different (ℓ, m), so the smoothed sky is still statistically isotropic, with power spectrum $B_\ell^2\, C_\ell$.

The last statement follows from $\langle |B_\ell a_{\ell m}|^2 \rangle = B_\ell^2 \langle |a_{\ell m}|^2 \rangle = B_\ell^2 C_\ell$. In the cosmology notes the same factor is called the window function, $W_\ell = B_\ell^2$ (see the cosmology notes); we keep B_ℓ because it is the factor that multiplies $a_{\ell m}$. A beam that is not circular has a kernel that depends on more than the angle between the two directions, and then it does mix the m's [HobsonBMC p. 232], [Hivon §3.6].

The Gaussian beam. Most telescope main lobes are close to a two-dimensional Gaussian of width σ_b,

$$b(\theta) = \frac{1}{2\pi\sigma_b^2}\, e^{-\theta^2/2\sigma_b^2}, \qquad \sigma_b = \frac{\mathrm{FWHM}}{\sqrt{8\ln 2}} \approx \frac{\mathrm{FWHM}}{2.355}, \tag{8.7}$$

where θ is the angle from the pointing and FWHM is the full width at half maximum [the angular diameter of the circle on which the beam has fallen to half its peak]. The relation between FWHM and σ_b comes from setting $e^{-\theta^2/2\sigma_b^2} = 1/2$ at $\theta = \mathrm{FWHM}/2$, which gives $\mathrm{FWHM}^2/(8\sigma_b^2) = \ln 2$. Since the beam is small, $\sin\theta \approx \theta$ and the normalisation is that of a flat two-dimensional Gaussian. Its B_ℓ follows from (8.5) with $x = \cos\theta$. For the low multipoles, where $\ell\sigma_b \ll 1$, expand P_ℓ around $x = 1$:

$$\begin{aligned} B_\ell &\overset{(1)}{=} 2\pi \int_0^\pi b(\theta)\, P_\ell(\cos\theta)\, \sin\theta\, \mathrm{d}\theta \\ &\overset{(2)}{=} 2\pi \int_0^\infty b(\theta) \Big[1 - \frac{\ell(\ell+1)}{2}\,\frac{\theta^2}{2}\Big] \theta\, \mathrm{d}\theta \\ &\overset{(3)}{=} 1 - \frac{\ell(\ell+1)}{4}\, \langle\theta^2\rangle_b \\ &\overset{(4)}{=} 1 - \frac{\ell(\ell+1)\,\sigma_b^2}{2}. \end{aligned} \tag{8.8}$$

(1) Change variable $x = \cos\theta$, $\mathrm{d}x = -\sin\theta\,\mathrm{d}\theta$.

(2) Taylor expansion $P_\ell(\cos\theta) \approx P_\ell(1) + P_\ell'(1)(\cos\theta - 1)$ with $P_\ell(1) = 1$, $\cos\theta - 1 \approx -\theta^2/2$, and $P_\ell'(1) = \ell(\ell+1)/2$. The last value follows from Legendre's equation $(1-x^2)P_\ell'' - 2xP_\ell' + \ell(\ell+1)P_\ell = 0$ evaluated at $x = 1$, where the first term vanishes. Also $\sin\theta \approx \theta$, and the upper limit can be moved to ∞ because the beam is negligible far from its centre.

(3) The first term is the normalisation $2\pi \int b\,\theta\,\mathrm{d}\theta = 1$. The second is the mean of θ^2 over the beam, $\langle\theta^2\rangle_b \equiv 2\pi \int \theta^2 b\,\theta\,\mathrm{d}\theta$.

(4) For a two-dimensional Gaussian, $\theta^2 = x^2 + y^2$ with x and y independent, each of variance σ_b^2, so $\langle\theta^2\rangle_b = 2\sigma_b^2$.

Equation (8.8) is the first-order expansion of $e^{-\ell(\ell+1)\sigma_b^2/2}$, and the cosmology notes reach the same expansion for the window function, $W_\ell \approx 1 - \sigma_b^2\ell(\ell+1)$. At high ℓ the flat-sky limit fixes the full shape: on a small patch, ℓ becomes the length of a two-dimensional wave vector, and the Fourier transform of a two-dimensional Gaussian is a Gaussian,

$$\tilde b(\boldsymbol{\ell}) = \int \frac{e^{-(x^2+y^2)/2\sigma_b^2}}{2\pi\sigma_b^2}\, e^{-i(\ell_x x + \ell_y y)}\, \mathrm{d}x\, \mathrm{d}y = e^{-\ell_x^2\sigma_b^2/2}\, e^{-\ell_y^2\sigma_b^2/2} = e^{-\ell^2\sigma_b^2/2}, \tag{8.9}$$

each factor being the familiar one-dimensional result obtained by completing the square. The formula that joins both limits is the one used everywhere,

$$B_\ell = \exp\!\left[-\frac{\ell(\ell+1)\,\sigma_b^2}{2}\right] \tag{8.10}$$

[166], and it is what `healpy.gauss_beam` returns. Is it exact? The function `gaussian_beam_exact` in `code/ch08/lib_cmbsim.py` evaluates the Legendre integral (8.5) numerically, with P_ℓ from Bonnet's recursion $(\ell+1)P_{\ell+1} = (2\ell+1)xP_\ell - \ell P_{\ell-1}$. For a 30′ beam the two agree to 8×10^{-7} at every $\ell \le 767$; even for beams of 5° and 10° they differ by at most 0.2% and 0.6% wherever $B_\ell > 10^{-3}$ (Figure 8.4, script `code/ch08/02_beam_window.py`).

Example: where the beam halves the power. Take the beam of our toy experiment, FWHM = 30′, so $\sigma_b = 12.7' = 3.71\times10^{-3}$ rad. The observed power $B_\ell^2 C_\ell$ is half the true power where $\ell(\ell+1)\sigma_b^2 = \ln 2$, that is at $\ell \approx \sqrt{\ln 2}/\sigma_b \approx 225$ (the numerical B_ℓ gives 224). For a 5′ beam, typical of the best Planck channels, $\sigma_b = 2.12' = 6.18\times10^{-4}$ rad and the power is halved only at $\ell \approx 1350$. The beam sets the multipole range an experiment can reach: angular resolution in degrees and multipole reach are two descriptions of the same number.

Example: the beam at the edge of the analysed range. At $\ell = 512$ our beam gives $B_\ell = 0.165$, so the sky power there is multiplied by $B_\ell^2 \approx 0.027$, a suppression by a factor of about 37. The sky is still there; it has been turned down, and it can be turned back up by dividing by B_ℓ^2. What that costs is the subject of Section 8.4.

The pixel window. Storing the smoothed sky in pixels averages it once more, over the area of each pixel. If every pixel had the same shape, this would be one more convolution, with a kernel equal to $1/\Omega_{\rm pix}$ inside the pixel and zero outside. HEALPix pixels have equal areas $\Omega_{\rm pix} = 4\pi/N_{\rm pix}$ but slightly different shapes [7], so the averaging is only approximately a convolution. Averaged over the sky, its effect on the spectrum is a multiplication by p_ℓ^2, where p_ℓ, the **pixel window function**, is tabulated by `healpy.pixwin` [Hivon (15)]. Its low-ℓ form follows from the same expansion as (8.8), with the beam replaced by a square of side $a = \sqrt{\Omega_{\rm pix}}$: the mean of x^2+y^2 over a square is $2\cdot a^2/12$, so

$$p_\ell \approx 1 - \frac{\ell(\ell+1)}{4}\cdot\frac{a^2}{6} = 1 - \frac{\ell(\ell+1)\,\Omega_{\rm pix}}{24}. \tag{8.11}$$

Example: the pixel window of $N_{\rm side} = 256$. The pixels have $\Omega_{\rm pix} = 1.598\times10^{-5}$ sr, sides of 13.7′. At $\ell = 256$, (8.11) gives $1 - 65792\times1.598\times10^{-5}/24 = 0.956$, against the tabulated 0.954. At $\ell = 512$ it gives 0.825, against 0.827. A 17% suppression of the amplitude at $\ell = 512$ is a 32% suppression of the power. Forgetting it is not a small error (Section 8.6).

The beam and the pixel act on the same sky one after the other, so their factors multiply. We call the product the **transfer function** of the instrument,

$$T_\ell \equiv B_\ell\, p_\ell, \qquad a_{\ell m} \;\to\; T_\ell\, a_{\ell m}, \qquad C_\ell \;\to\; T_\ell^2\, C_\ell. \tag{8.12}$$

Real pipelines add further factors to T_ℓ, for example a filter applied to the timestream, and measure them with simulations [Hivon (15), (18)].

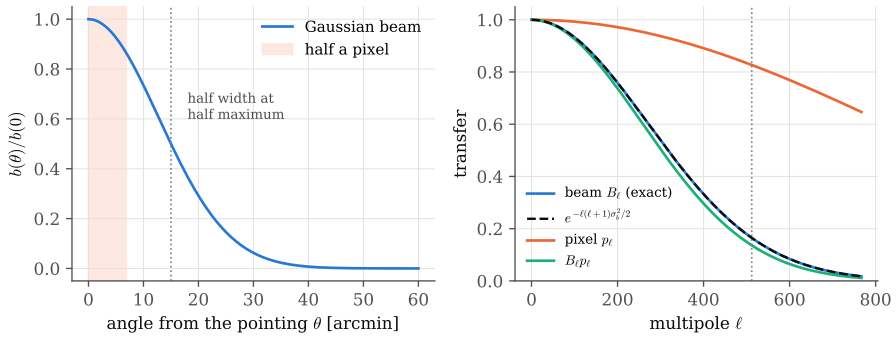


Figure 8.4: Left: the Gaussian beam of the toy experiment as a function of the angle from the pointing; the shaded band is half a pixel. Right: in harmonic space the beam becomes the transfer function B_ℓ. The exact Legendre transform (blue) and the formula $e^{-\ell(\ell+1)\sigma_b^2/2}$ (dashed) are indistinguishable. The pixel window p_ℓ (orange) is a milder suppression; the product $B_\ell p_\ell$ (green) is what multiplies each $a_{\ell m}$. The dotted line marks the analysed range $\ell \le 512$.

8.2.2 Pixel noise in harmonic space

The noise n_p enters (8.4) after the smoothing. To see what it does to the spectrum we need its harmonic coefficients, which a map transform computes as a sum over pixels, the quadrature version of $\int \mathrm{d}\Omega$:

$$n_{\ell m} = \Omega_{\mathrm{pix}} \sum_p n_p \, Y^*_{\ell m}(\hat{\boldsymbol{n}}_p). \tag{8.13}$$

Each $n_{\ell m}$ is a sum of independent Gaussian numbers, so it is Gaussian with mean zero. Its covariance is

$$\begin{aligned}
\langle n_{\ell m}\, n^*_{\ell' m'} \rangle &\overset{(1)}{=} \Omega^2_{\mathrm{pix}} \sum_{p,q} \langle n_p n_q \rangle \, Y^*_{\ell m}(\hat{\boldsymbol{n}}_p)\, Y_{\ell' m'}(\hat{\boldsymbol{n}}_q) \\
&\overset{(2)}{=} \Omega^2_{\mathrm{pix}} \sum_p \sigma_p^2 \, Y^*_{\ell m}(\hat{\boldsymbol{n}}_p)\, Y_{\ell' m'}(\hat{\boldsymbol{n}}_p) \qquad (8.14) \\
&\overset{(3)}{=} \sigma^2 \, \Omega_{\mathrm{pix}} \sum_p \Omega_{\mathrm{pix}} \, Y^*_{\ell m}(\hat{\boldsymbol{n}}_p)\, Y_{\ell' m'}(\hat{\boldsymbol{n}}_p) \\
&\overset{(4)}{=} \sigma^2 \, \Omega_{\mathrm{pix}} \, \delta_{\ell\ell'} \delta_{mm'}. \qquad (8.15)
\end{aligned}$$

(1) Insert (8.13) twice; the noise is real, $n^*_p = n_p$, and the average acts only on the noise.

(2) Different pixels are independent, $\langle n_p n_q \rangle = \sigma_p^2 \delta_{pq}$.

(3) For uniform noise, $\sigma_p = \sigma$ in every pixel; split Ω^2_{pix} into two factors.

(4) The remaining sum is the pixel approximation of $\int Y^*_{\ell m} Y_{\ell' m'} \mathrm{d}\Omega = \delta_{\ell\ell'}\delta_{mm'}$, accurate for $\ell \lesssim 2N_{\mathrm{side}}$ (Chapter 7).

Uniform, uncorrelated pixel noise is therefore **white** in harmonic space: every (ℓ, m) gets an independent Gaussian of the same variance,

$$N_\ell = \sigma^2 \, \Omega_{\mathrm{pix}} \equiv \Delta_T^2, \tag{8.16}$$

the **noise power spectrum** [Hivon §3.3], [166]. The cosmology notes write it as w^{-1}, where $w \equiv (\Omega_{\mathrm{pix}}\sigma^2)^{-1}$ is the weight per solid angle (see the cosmology notes).[1] The square root

[1] In [HobsonBMC (10.24)–(10.25)] the symbol σ^2 is the variance of the harmonic noise coefficients, $\sigma^2_{\mathrm{pix}}\Omega_{\mathrm{pix}}$ in the notation here, not the variance of a pixel; the two differ by a factor of order 10^{-5} for our pixels.

$\Delta_T = \sigma\sqrt{\Omega_{\rm pix}}$ is called the **map depth**, quoted in μK arcmin [the noise of a pixel one arcminute on a side]. It is the natural label of an experiment because it does not depend on how the map was pixelised: at a fixed total observing time, a pixel half as large in area gets half the samples, so σ^2 doubles and $\sigma^2\Omega_{\rm pix}$ stays the same.

Example: the toy experiment. Our experiment has depth $\Delta_T = 200\,\mu$K arcmin on $N_{\rm side} = 256$ pixels of side 13.7′. The noise per pixel is $\sigma = \Delta_T/13.7' = 14.6\,\mu$K, and $N_\ell = (200 \times 2.909 \times 10^{-4})^2\,\mu{\rm K}^2 = 3.38 \times 10^{-3}\,\mu{\rm K}^2$, using $1' = 2.909 \times 10^{-4}$ rad. On the map (Figure 8.5) the sky has an rms of 108 μK, the beam brings it down to 86 μK, and the noise adds 14.6 μK per pixel. By eye the noise is a faint grain on the large spots. In harmonic space it dominates the beamed sky above $\ell \approx 436$ (Section 8.3).

Example: Knox's experiment. Knox [166] considered a year of observation with $\sigma = 22\,\mu$K in $20' \times 20'$ pixels. Its depth is $\Delta_T = 22 \times 20 = 440\,\mu$K arcmin, or 7.3 μK deg. He quotes the weight per solid angle $w = (7.5\,\mu{\rm K})^{-2}\,{\rm deg}^{-2}$, which is the same number to the rounding in his σ. The same experiment pixelised into $10' \times 10'$ pixels would show $\sigma = 44\,\mu$K per pixel and exactly the same N_ℓ.

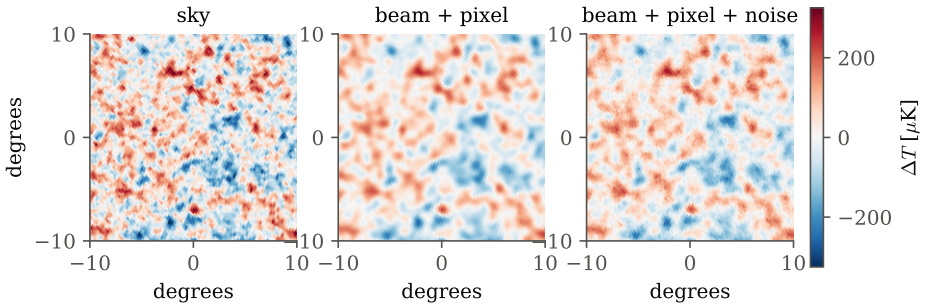


Figure 8.5: One simulated sky, 20° on a side, before and after the instrument. Left: the sky drawn from the fiducial C_ℓ. Middle: after the 30′ beam and the 13.7′ pixels; the smallest spots are gone. Right: with 200 μK arcmin of white noise added, which shows up as a fine grain. The colour scale is the same in all three panels.

Watch out: The beam acts on the sky, not on the noise

In (8.4) the noise is added after the smoothing, so the observed coefficients are $d_{\ell m} = T_\ell a_{\ell m} + n_{\ell m}$ and not $T_\ell(a_{\ell m} + n_{\ell m})$. The beam turns the sky down at high ℓ but leaves the noise at its full, flat level [166]. Dividing the observed coefficients by T_ℓ to undo the beam therefore turns the noise up by $1/T_\ell$: the effective noise on the sky is N_ℓ/T_ℓ^2, which grows like $e^{\ell(\ell+1)\sigma_b^2}$. A simulation that smooths the noise along with the sky gets every error bar at high ℓ wrong.

When the noise is not uniform. A scanning telescope does not visit every pixel equally often, and the noise variance of a pixel falls as one over its number of hits [HobsonBMC (10.15)]. Then σ_p changes across the sky and step (3) of (8.15) is no longer allowed. Line (8.14) is still exact, and it says that the harmonic noise is no longer diagonal: a pattern in σ_p^2 couples different (ℓ, m). One combination survives untouched. Average the diagonal over the $2\ell + 1$ values of m:

$$\bar{N}_\ell \equiv \frac{1}{2\ell+1}\sum_{m=-\ell}^{\ell}\langle|n_{\ell m}|^2\rangle \overset{(1)}{=} \frac{\Omega_{\rm pix}^2}{2\ell+1}\sum_p \sigma_p^2 \sum_m |Y_{\ell m}(\hat{\boldsymbol{n}}_p)|^2$$

$$\overset{(2)}{=} \frac{\Omega_{\text{pix}}^2}{4\pi} \sum_p \sigma_p^2$$

$$\overset{(3)}{=} \Omega_{\text{pix}} \langle \sigma^2 \rangle. \tag{8.17}$$

(1) Line (8.14) with $\ell' = \ell$, $m' = m$, summed over m.
(2) The addition theorem at coincident points, $\sum_m |Y_{\ell m}(\hat{\boldsymbol{n}})|^2 = (2\ell+1)/4\pi$, the same for every direction.
(3) $N_{\text{pix}}\Omega_{\text{pix}} = 4\pi$, and $\langle \sigma^2 \rangle \equiv \sum_p \sigma_p^2 / N_{\text{pix}}$ is the sky-averaged pixel variance.

However the noise is distributed over the sky, the average noise power at every ℓ is the pixel area times the mean pixel variance [134, eq. 48]. This number fixes the noise *bias* of the next section. The noise *scatter* depends on the full pattern, and Section 8.6 measures it.

8.3 The noise bias and how to remove it

Take the observed map of the toy experiment (Figure 8.5, right), transform it, and average $|d_{\ell m}|^2$ over m at $\ell = 500$, exactly as Chapter 7 did for a perfect sky. The answer comes out about four times larger than the beamed sky power $T_\ell^2 C_\ell$ at that multipole. Nothing went wrong in the transform. The noise in each $d_{\ell m}$ averages to zero, but its square does not: squaring a fluctuation turns it into power, and power from the detector adds to power from the sky. Every spectrum computed from a noisy map is biased upward, and the first task is to say by exactly how much.

Write the observed harmonic coefficients, using the transfer function $T_\ell = B_\ell p_\ell$ of (8.12) and the harmonic noise $n_{\ell m}$ of (8.13), as $d_{\ell m} = T_\ell a_{\ell m} + n_{\ell m}$, and define the **observed spectrum** as the same average that gave $\widehat{C}_\ell$ on a perfect sky:

$$\widehat{C}_\ell^{\text{obs}} \equiv \frac{1}{2\ell+1} \sum_{m=-\ell}^{\ell} |d_{\ell m}|^2. \tag{8.18}$$

Its average over skies and noise realisations is

$$\begin{aligned} \big\langle \widehat{C}_\ell^{\text{obs}} \big\rangle &\overset{(1)}{=} \frac{1}{2\ell+1} \sum_m \Big[T_\ell^2 \langle |a_{\ell m}|^2 \rangle + T_\ell \langle a_{\ell m} n_{\ell m}^* \rangle + T_\ell \langle a_{\ell m}^* n_{\ell m} \rangle + \langle |n_{\ell m}|^2 \rangle \Big] \\ &\overset{(2)}{=} \frac{1}{2\ell+1} \sum_m \Big[T_\ell^2 C_\ell + \langle |n_{\ell m}|^2 \rangle \Big] \\ &\overset{(3)}{=} T_\ell^2 \, C_\ell + \bar{N}_\ell. \end{aligned} \tag{8.19}$$

(1) Expand $|T_\ell a_{\ell m} + n_{\ell m}|^2$; T_ℓ is real.
(2) The sky and the noise are independent and both have mean zero, so $\langle a_{\ell m} n_{\ell m}^* \rangle = \langle a_{\ell m} \rangle \langle n_{\ell m}^* \rangle = 0$; statistical isotropy gives $\langle |a_{\ell m}|^2 \rangle = C_\ell$ for every m (Chapter 7).
(3) The definition (8.17) of the mean noise power; for uniform noise $\bar{N}_\ell = N_\ell = \sigma^2 \Omega_{\text{pix}}$.

The observed spectrum measures the beamed sky plus the noise. The excess $\bar{N}_\ell$ is the **noise bias** [Hivon (15)], [166]. Because it is known, it can be subtracted, and because T_ℓ is known, the beam can be divided out.

Working definition 8.2 (debiased, deconvolved estimator)

For a full-sky map with transfer function T_ℓ and noise power N_ℓ,

$$\widehat{C}_\ell = \frac{\widehat{C}_\ell^{\,\mathrm{obs}} - N_\ell}{T_\ell^2} \tag{8.20}$$

[subtract the power the noise alone would give, then undo the suppression by the beam and the pixel]. By (8.19) it is unbiased, $\langle \widehat{C}_\ell \rangle = C_\ell$, for every ℓ. *Operationally:* compute `alm2cl(map2alm(d))`, subtract N_ℓ (from (8.16), or from noise-only simulations when no formula exists), divide by $B_\ell^2 p_\ell^2$.

How large is the correction? The left panel of Figure 8.6 plots the three spectra of (8.19) for the toy experiment. The beamed sky $T_\ell^2 C_\ell$ falls steeply past $\ell \approx 300$, while the noise N_ℓ is flat, and the two cross at $\ell_{\mathrm{eq}} = 436$. Below ℓ_{eq} the measurement is limited by the sky itself; above it, by the detector.

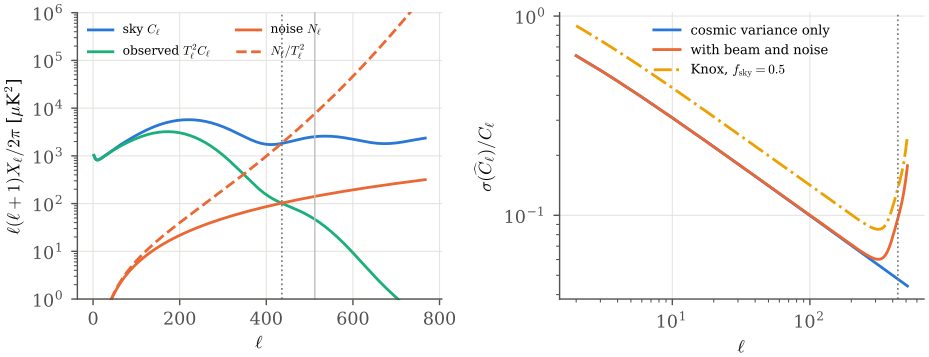


Figure 8.6: The signal and noise budget of the toy experiment, plotted as $\ell(\ell+1)X_\ell/2\pi$ for each spectrum X_ℓ. Left: the sky spectrum (blue) is turned down by the beam and pixel (green), while the white noise (orange, solid), flat in N_ℓ, rises only as ℓ^2 in these units; the two cross at ℓ_{eq} (dotted line). After dividing by T_ℓ^2 the noise (orange, dashed) climbs exponentially. The grey line marks the analysed range $\ell \le 512$. Right: the fractional error on C_ℓ at a single multipole, from cosmic variance alone (blue), with the beam and the noise (orange), and with Knox's sky-fraction factor for half the sky (yellow); the dotted line is again ℓ_{eq}.

Example: forgetting the subtraction. Without the subtraction the estimator $\widehat{C}_\ell^{\,\mathrm{obs}}/T_\ell^2$ is off by $N_\ell/(T_\ell^2 C_\ell)$. For the toy experiment this is 0.05 at $\ell = 300$, a 5% excess, close to the cosmic-variance error of a single multipole there (5.8%). At $\ell = 512$ it is 3.0: the uncorrected spectrum is four times too high.

8.3.1 Is this the best we can do?

The estimator (8.20) was built by matching an average. The procedure of Chapter 3 asks a different question: which value of C_ℓ makes the data we observed most probable? Say it in words first, in four moves.

The observed outcome is what the map gives us at one ℓ: the coefficients $d_{\ell m}$. They are complex, but only $2\ell + 1$ real numbers are free, because the map is real and $d_{\ell,-m} = (-1)^m d_{\ell m}^*$ (Chapter 7). To make these numbers all alike, rescale the ones that come in pairs. The real $d_{\ell 0}$ has variance V_ℓ; for $m > 0$ the variance $V_\ell = \langle |d_{\ell m}|^2 \rangle$ is shared equally between the real and

imaginary parts, so each of them has $V_\ell/2$. Define

$$u_1 \equiv d_{\ell 0}, \qquad u_{2m} \equiv \sqrt{2}\,\mathrm{Re}\, d_{\ell m}, \qquad u_{2m+1} \equiv \sqrt{2}\,\mathrm{Im}\, d_{\ell m} \qquad (m = 1, \dots, \ell), \tag{8.21}$$

$$\sum_{k=1}^{\nu} u_k^2 \overset{(1)}{=} d_{\ell 0}^2 + 2\sum_{m=1}^{\ell} |d_{\ell m}|^2 \overset{(2)}{=} \sum_{m=-\ell}^{\ell} |d_{\ell m}|^2 \overset{(3)}{=} \nu\, \widehat{C}_\ell^{\,\mathrm{obs}}, \qquad \nu = 2\ell + 1.$$

(1) $(\sqrt{2}\,\mathrm{Re}\, d)^2 + (\sqrt{2}\,\mathrm{Im}\, d)^2 = 2|d|^2$.
(2) $|d_{\ell,-m}| = |d_{\ell m}|$, so the terms with $m < 0$ repeat those with $m > 0$.
(3) The definition (8.18) of the observed spectrum.

Each u_k now has the same variance V_ℓ and mean zero, and the ν of them are independent. Every later count of "$2\ell + 1$ independent numbers" at one ℓ refers to these u_k.

The hypotheses are the possible values of C_ℓ. Under one of them, the model predicts mean zero for every u_k; what turns that prediction into the numbers actually observed is a random fluctuation, made of an independent sky part and an independent noise part. Variances of independent Gaussians add, so under the hypothesis C_ℓ the fluctuation is Gaussian with variance $V_\ell = T_\ell^2 C_\ell + N_\ell$. The likelihood of C_ℓ asks how plausible it is that a fluctuation of that size produced exactly these $u_1, \dots, u_\nu$: too small a C_ℓ makes the observed values improbably large, too large a C_ℓ makes them improbably small. Read as a function of C_ℓ with the data held fixed, that probability ranks the hypotheses; maximising it, or combining it with a prior by Bayes' theorem, picks among them. Then

$$-2\ln \mathcal{L}(C_\ell) \overset{(1)}{=} \sum_{k=1}^{\nu}\left[\frac{u_k^2}{V_\ell} + \ln(2\pi V_\ell)\right]$$

$$\overset{(2)}{=} \nu\left[\frac{\widehat{C}_\ell^{\,\mathrm{obs}}}{T_\ell^2 C_\ell + N_\ell} + \ln\!\left(T_\ell^2 C_\ell + N_\ell\right)\right] + \text{const}, \tag{8.22}$$

$$\frac{\partial(-2\ln\mathcal{L})}{\partial C_\ell} \overset{(3)}{=} \nu\, T_\ell^2\left[-\frac{\widehat{C}_\ell^{\,\mathrm{obs}}}{(T_\ell^2 C_\ell + N_\ell)^2} + \frac{1}{T_\ell^2 C_\ell + N_\ell}\right] = 0 \implies \hat{C}_\ell^{\mathrm{ML}} = \frac{\widehat{C}_\ell^{\,\mathrm{obs}} - N_\ell}{T_\ell^2}. \tag{8.23}$$

(1) Each u_k is $\mathcal{N}(0, V_\ell)$, with density $(2\pi V_\ell)^{-1/2} e^{-u_k^2/2V_\ell}$; independent densities multiply, so minus twice their logarithms add.
(2) $\sum_k u_k^2 = \nu \widehat{C}_\ell^{\,\mathrm{obs}}$, and $V_\ell = T_\ell^2 C_\ell + N_\ell$; the $\ln 2\pi$ terms do not depend on C_ℓ.
(3) Differentiate with the chain rule, $\partial V_\ell / \partial C_\ell = T_\ell^2$; the bracket vanishes when $V_\ell = \widehat{C}_\ell^{\,\mathrm{obs}}$.

The maximum-likelihood estimator is the debiased, deconvolved estimator [HobsonBMC (10.24)–(10.25)]. On the full sky with uniform white noise, the cheap estimator is not an approximation to the best one; it *is* the best one, the quadratic maximum-likelihood estimator of Section 8.1.1 with every matrix diagonal. Section 8.4 confirms that its variance reaches the Cramér–Rao bound.

One detail of (8.23) matters in practice. If the noise happens to fluctuate low, $\widehat{C}_\ell^{\,\mathrm{obs}} < N_\ell$ and the estimate is negative, although C_ℓ cannot be. Restricting the maximum to $C_\ell \geq 0$ would replace those estimates by zero and bias the average upward. We keep the unrestricted estimator, which is unbiased and can be negative, and leave the constraint $C_\ell \geq 0$ to the likelihood in which $\widehat{C}_\ell$ is later used (Chapter 9).

8.3.2 When the noise level is not known exactly

The subtraction in (8.20) needs N_ℓ. For white, uniform noise (8.16) gives it. Real noise is coloured, varies across the sky, and is filtered during map-making; then N_ℓ has no formula and

is estimated by simulating noise-only timestreams, making maps from them in the same way, and averaging their spectra, $\langle \tilde{N}_\ell \rangle_{\rm MC}$ [Hivon §3.3]. Two questions follow: how much bias does an error in N_ℓ cause, and how many noise simulations keep that error small?

Example: a 2% error in the noise level. Subtract $(1+\varepsilon)N_\ell$ instead of N_ℓ. By (8.19) the average of the estimator becomes $C_\ell - \varepsilon N_\ell / T_\ell^2$, a fractional bias

$$\frac{\langle \widehat{C}_\ell \rangle - C_\ell}{C_\ell} = -\varepsilon \, \frac{N_\ell}{T_\ell^2 C_\ell}. \tag{8.24}$$

At $\ell = 512$, $N_\ell/(T_\ell^2 C_\ell) = 3.0$, so $\varepsilon = 0.02$ gives -6%. The statistical error of one multipole there is 18% (Section 8.4), and of a bin of 25 multipoles about $18\%/\sqrt{25} \approx 3.6\%$. A 2% error in the noise model produces a bias of almost two standard deviations per bin.

Watch out: A small noise error is a large spectrum error at high ℓ

The bias (8.24) is the noise error ε times the noise-to-signal ratio $N_\ell/(T_\ell^2 C_\ell)$, which grows like $e^{\ell(\ell+1)\sigma_b^2}$. However well the noise is known, there is a multipole beyond which the auto-spectrum is dominated by the uncertainty in the noise model rather than by the sky.

Example: how many noise simulations? If N_ℓ is estimated as the average of the spectra of N_n independent noise-only maps, each of which scatters by $\sqrt{2/(2\ell+1)}\, N_\ell$ (the noise coefficients are Gaussian, so their spectrum is a scaled χ^2 like the sky's, Chapter 7), the estimate scatters by $\sqrt{2/(2\ell+1)}\, N_\ell/\sqrt{N_n}$. Compare it with the statistical error of $\widehat{C}_\ell^{\rm obs}$, which is $\sqrt{2/(2\ell+1)}\,(T_\ell^2 C_\ell + N_\ell)$ (Section 8.4):

$$\frac{\text{error from the noise estimate}}{\text{statistical error}} = \frac{N_\ell}{T_\ell^2 C_\ell + N_\ell} \, \frac{1}{\sqrt{N_n}} \;\le\; \frac{1}{\sqrt{N_n}}. \tag{8.25}$$

With 100 noise simulations the extra error is at most 10% of the statistical one. Errors from independent sources add in quadrature, so the total error bar is at most $\sqrt{1+0.1^2} = 1.005$ times the statistical one: it grows by at most 0.5%. This is the number Hivon et al. found for the Boomerang analysis [Hivon (42)].

8.3.3 Cross-spectra: no noise bias at all

There is a way to avoid the noise model entirely. Split the data in two: the first and second halves of the mission, or two sets of detectors. Each half sees the same sky through the same beam, $d_{\ell m}^{(i)} = T_\ell a_{\ell m} + n_{\ell m}^{(i)}$ for $i = 1, 2$, but the two noise realisations are independent. Instead of squaring one map, multiply the two:

$$\widehat{C}_\ell^{12} \equiv \frac{1}{2\ell+1} \sum_m \mathrm{Re}\big(d_{\ell m}^{(1)} d_{\ell m}^{(2)*}\big), \qquad \big\langle \widehat{C}_\ell^{12} \big\rangle = T_\ell^2 C_\ell + \big\langle \mathrm{Re}\, n_{\ell m}^{(1)} n_{\ell m}^{(2)*} \big\rangle = T_\ell^2 C_\ell. \tag{8.26}$$

The cross terms between sky and noise vanish as in (8.19), and the noise–noise term vanishes because the two noises are independent. The **cross-spectrum** has no noise bias, and it needs no noise model [16]. The Planck high-ℓ likelihood is built from cross-spectra between half-mission maps for exactly this reason [75] (Figure 8.7).

The price is paid in scatter. Each half has twice the noise variance of the full map, $N_\ell^{(1)} = N_\ell^{(2)} = 2N_\ell$, because it has half the observing time. To find the variance of (8.26),

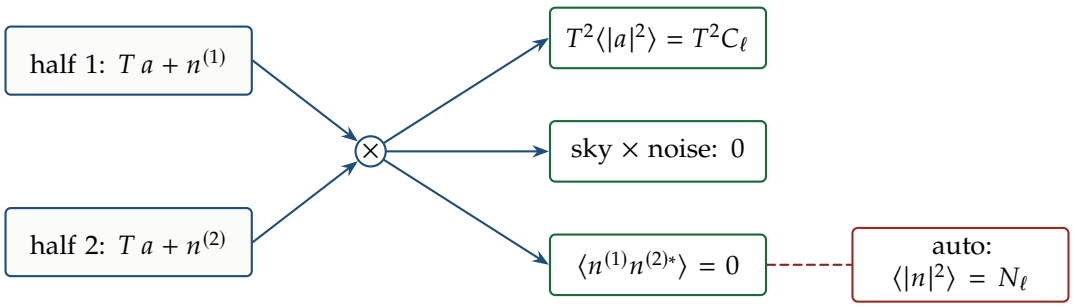


Figure 8.7: Why the cross-spectrum has no noise bias. Multiplying two maps of the same sky with independent noise leaves three kinds of terms on average: the sky power survives, sky times noise averages to zero, and noise times the *other* noise averages to zero. In the auto-spectrum (red) the last term is a noise times itself, whose average is the noise power.

write it, as for the likelihood above, as a sum over the $\nu = 2\ell + 1$ real degrees of freedom, $\nu \widehat{C}_\ell^{12} = \sum_k u_k v_k$, where u_k and v_k are the corresponding real components of the two maps. The pairs are independent for different k; within a pair u and v are zero-mean Gaussians with variances $S + 2N$ each (writing $S = T_\ell^2 C_\ell$, $N = N_\ell$) and covariance $\langle uv\rangle = S$, since only the sky is shared. Then

$$\begin{aligned}\mathrm{Var}(uv) &\overset{(1)}{=} \langle u^2 v^2\rangle - \langle uv\rangle^2 \\ &\overset{(2)}{=} \langle u^2\rangle\langle v^2\rangle + 2\langle uv\rangle^2 - \langle uv\rangle^2 \\ &\overset{(3)}{=} (S+2N)^2 + S^2, \end{aligned} \tag{8.27}$$

$$\mathrm{Var}\big(\widehat{C}_\ell^{12}\big) \overset{(4)}{=} \frac{(S+2N)^2 + S^2}{2\ell+1} \quad \text{versus} \quad \mathrm{Var}\big(\widehat{C}_\ell^{\mathrm{obs}}\big) = \frac{2(S+N)^2}{2\ell+1}. \tag{8.28}$$

(1) Definition of the variance.
(2) Isserlis' theorem for zero-mean Gaussians (Chapter 7): $\langle uuvv\rangle$ is the sum over the three ways of pairing the four factors.
(3) Insert the variances and the covariance.
(4) The ν terms are independent, so their variances add, and dividing the sum by ν divides the variance by ν^2. The auto-spectrum line is the case $u = v$ with variance $S + N$, derived in Section 8.4.

When the sky dominates ($N \ll S$) both variances are $2S^2/\nu$: nothing is lost. When the noise dominates ($S \ll N$) the cross-spectrum variance is $4N^2/\nu$ against $2N^2/\nu$, so its error bar is $\sqrt{2}$ times larger. At $\ell = 512$ the predicted ratio of error bars is 1.25; the simulations of Section 8.6 find 1.24. The cross-spectrum trades a little precision for immunity to the noise model, a trade that real analyses usually accept.

8.4 How much the estimate scatters

The debiased, deconvolved estimator $\widehat{C}_\ell = (\widehat{C}_\ell^{\mathrm{obs}} - N_\ell)/T_\ell^2$ of (8.20) centres on the true spectrum. Here $\widehat{C}_\ell^{\mathrm{obs}}$ is the average of $|d_{\ell m}|^2$ over m on the observed map, N_ℓ the noise power and $T_\ell = B_\ell p_\ell$ the beam times the pixel window (Section 8.2). Centring on the truth is a statement about the average over many skies. We have one sky, and on it a single estimate at $\ell = 512$ typically lands about a fifth away from C_ℓ. On a perfect sky the same estimate would land only about 4% away. Something in the observation has made the scatter four times larger, and we want to know what, by how much at every ℓ, and whether any other estimator could do better.

Said in words, there are four questions. How wide is the spread of $\widehat{C}_\ell$ over skies and noise realisations? How much of it comes from the sky and how much from the detector? Is it the smallest spread any estimator could achieve with these data? And what is the full shape of the distribution, not just its width?

The width. The answer comes from the same bookkeeping that gave the likelihood in Section 8.3.1. At one ℓ the observed map supplies $\nu = 2\ell + 1$ real numbers $u_1, \dots, u_\nu$, the coefficients $d_{\ell m}$ rescaled so that all have the same variance, (8.21). Each is a sky part plus an independent noise part, so each is a zero-mean Gaussian with variance $V_\ell = T_\ell^2 C_\ell + N_\ell$, they are independent of one another, and $\sum_k u_k^2 = \nu\, \widehat{C}_\ell^{\,\mathrm{obs}}$. A sum of ν squared independent standard Gaussians is a χ^2_ν variable, whose variance is 2ν (Chapter 7). Then

$$\begin{aligned}\mathrm{Var}\big(\widehat{C}_\ell\big) &\overset{(1)}{=} \frac{1}{T_\ell^4}\,\mathrm{Var}\big(\widehat{C}_\ell^{\,\mathrm{obs}}\big) \overset{(2)}{=} \frac{1}{T_\ell^4}\,\frac{V_\ell^2}{\nu^2}\,\mathrm{Var}\Big(\sum_{k=1}^{\nu}\frac{u_k^2}{V_\ell}\Big)\\ &\overset{(3)}{=} \frac{1}{T_\ell^4}\,\frac{V_\ell^2}{\nu^2}\,2\nu \overset{(4)}{=} \frac{2}{2\ell+1}\Big(C_\ell + \frac{N_\ell}{T_\ell^2}\Big)^2 .\end{aligned} \tag{8.29}$$

(1) N_ℓ and T_ℓ are fixed numbers: subtracting a constant does not change a variance, and dividing by T_ℓ^2 divides it by T_ℓ^4.
(2) $\widehat{C}_\ell^{\,\mathrm{obs}} = \frac{1}{\nu}\sum_k u_k^2 = \frac{V_\ell}{\nu}\sum_k (u_k^2/V_\ell)$, and a constant factor comes out of a variance squared.
(3) Each $u_k/\sqrt{V_\ell}$ is a standard Gaussian, so the sum is χ^2_ν, with variance 2ν.
(4) $V_\ell/T_\ell^2 = C_\ell + N_\ell/T_\ell^2$ and $\nu = 2\ell + 1$.

Key idea: The error bar of a full-sky, beamed, noisy spectrum

$$\sigma\big(\widehat{C}_\ell\big) = \sqrt{\frac{2}{2\ell+1}}\,\Big(C_\ell + \frac{N_\ell}{T_\ell^2}\Big).$$

The sky and the noise enter added, inside the bracket, and the noise enters as it looks *after* the beam has been divided out, N_ℓ/T_ℓ^2.

This is the formula of Knox [166, eq. 4], written there with the weight per solid angle $w^{-1} = N_\ell$ and a Gaussian beam, $T_\ell^2 \to e^{-\ell^2\sigma_b^2}$. The cosmology notes reach it as the inverse square root of the Fisher information of C_ℓ, with the window $W_\ell = B_\ell^2$ (see the cosmology notes). With no noise it is the cosmic variance of Chapter 7. Knox also explained the factor $2/(2\ell+1)$ in words: the error on a variance measured from n Gaussian samples is $2\sigma^4/n$, however precisely each sample is measured, and the sky offers $2\ell + 1$ samples at each ℓ [Knox1995 p. 8].

Where the variance comes from. Squaring the bracket in (8.29) splits the variance into three terms,

$$\mathrm{Var}\big(\widehat{C}_\ell\big) = \frac{1}{2\ell+1}\Big[\underbrace{2C_\ell^2}_{\text{sky}} + \underbrace{4\,C_\ell\,\frac{N_\ell}{T_\ell^2}}_{\text{sky}\times\text{noise}} + \underbrace{2\Big(\frac{N_\ell}{T_\ell^2}\Big)^2}_{\text{noise}}\Big]. \tag{8.30}$$

The first is the cosmic variance, which no instrument can remove. The last is the scatter of the noise power itself, which would be there with no sky at all. The middle one is easy to overlook. Each observed coefficient is a sum of a sky part and a noise part, so $|d_{\ell m}|^2$ contains the product of the two. That product averages to zero, which is why it did not appear in the bias (8.19), but it fluctuates, and its fluctuation is the cross term.

Example: when does the noise supply half the variance? Write $r = N_\ell/(T_\ell^2 C_\ell)$ for the noise-to-signal ratio on the sky. The two terms of (8.30) that involve the noise make up half of the total when

$$4r + 2r^2 = \tfrac{1}{2}\big(2 + 4r + 2r^2\big) \overset{(1)}{=} r^2 + 2r - 1 = 0 \overset{(2)}{=} r = \sqrt{2} - 1 \approx 0.41.$$

(1) Divide (8.30) by $C_\ell^2/(2\ell+1)$, multiply out and collect terms.
(2) The positive root of the quadratic.

The noise already supplies half the variance when its power on the sky is 41% of the signal, well before the two powers are equal. At equality, $r = 1$, the noise terms are $6/8 = 75\%$ of the variance. In the toy experiment of Section 8.2.2 (a 30′ beam and 200 μK arcmin of white noise on $N_{\mathrm{side}} = 256$ pixels) the two powers cross at $\ell_{\mathrm{eq}} = 436$, but the noise supplies half the variance from $\ell = 386$ onward.

Example: the error bar at three multipoles. At $\ell = 300$ the cosmic variance alone gives a fractional error $\sqrt{2/601} = 5.8\%$, and the noise, with $r = 0.05$, raises it to 6.0%. At $\ell = 512$ the cosmic variance gives $\sqrt{2/1025} = 4.4\%$, but $r = 3.0$ multiplies it by $1 + r$, to 18%: the factor of four of the opening paragraph. At $\ell = 700$, beyond the analysed band, $r = 246$ and the fractional error is 9.33, nearly a thousand per cent. The right panel of Figure 8.6 plots the whole curve.

The smallest possible width. Is there a better estimator? The Cramér–Rao bound of Chapter 3 says no unbiased estimator can have a variance below one over the Fisher information, $I(C_\ell) = \big\langle \partial^2(-\ln \mathcal{L})/\partial C_\ell^2 \big\rangle$. With $-\ln\mathcal{L} = \frac{\nu}{2}\big[\widehat{C}_\ell^{\,\mathrm{obs}}/V_\ell + \ln V_\ell\big]$ + const from (8.22), and $\partial V_\ell/\partial C_\ell = T_\ell^2$,

$$\frac{\partial(-\ln\mathcal{L})}{\partial C_\ell} \overset{(1)}{=} \frac{\nu}{2}\,T_\ell^2\Big[-\frac{\widehat{C}_\ell^{\,\mathrm{obs}}}{V_\ell^2} + \frac{1}{V_\ell}\Big],$$

$$I(C_\ell) \overset{(2)}{=} \frac{\nu}{2}\,T_\ell^4\Big[\frac{2\langle\widehat{C}_\ell^{\,\mathrm{obs}}\rangle}{V_\ell^3} - \frac{1}{V_\ell^2}\Big] \overset{(3)}{=} \frac{\nu\,T_\ell^4}{2V_\ell^2} \overset{(4)}{=} \frac{2\ell+1}{2\,(C_\ell + N_\ell/T_\ell^2)^2}. \tag{8.31}$$

(1) Chain rule on both terms.
(2) Differentiate once more (again $\partial V_\ell/\partial C_\ell = T_\ell^2$) and average over the data.
(3) $\langle\widehat{C}_\ell^{\,\mathrm{obs}}\rangle = V_\ell$ by (8.19), so the bracket is $1/V_\ell^2$.
(4) $V_\ell/T_\ell^2 = C_\ell + N_\ell/T_\ell^2$.

One over (8.31) is exactly (8.29). The debiased estimator reaches the bound: on the full sky with uniform white noise it extracts every bit of information about C_ℓ that the map contains. Whatever an estimator on a masked sky loses is measured against this number.

8.4.1 The shape of the distribution

The width is not the whole story; the likelihood of Chapter 9 needs the full distribution. It follows from the same bookkeeping that gave the width (8.29) and the bound (8.31). There, the $\nu = 2\ell + 1$ real numbers u_k at one ℓ were independent Gaussians of variance $V_\ell = T_\ell^2 C_\ell + N_\ell$, so $X \equiv \sum_k u_k^2/V_\ell = \nu\,\widehat{C}_\ell^{\,\mathrm{obs}}/V_\ell$ is a χ^2_ν variable. Solve for $\widehat{C}_\ell^{\,\mathrm{obs}} = V_\ell X/\nu$ and insert it in the estimator

(8.20), with the noise-to-signal ratio $r_\ell \equiv N_\ell/(T_\ell^2 C_\ell)$ of Section 8.4:

$$\frac{\widehat{C}_\ell}{C_\ell} \overset{(1)}{=} \frac{\widehat{C}_\ell^{\,\mathrm{obs}} - N_\ell}{T_\ell^2 C_\ell} \overset{(2)}{=} \frac{V_\ell}{\nu\, T_\ell^2 C_\ell}\, X - \frac{N_\ell}{T_\ell^2 C_\ell}$$
$$\overset{(3)}{=} A_\ell\, X - r_\ell, \qquad X \sim \chi^2_{2\ell+1}, \qquad A_\ell = \frac{1+r_\ell}{2\ell+1}. \tag{8.32}$$

(1) The definition (8.20), divided by C_ℓ.
(2) $\widehat{C}_\ell^{\,\mathrm{obs}} = V_\ell X/\nu$.
(3) $V_\ell/(T_\ell^2 C_\ell) = 1 + N_\ell/(T_\ell^2 C_\ell) = 1 + r_\ell$, and $\nu = 2\ell + 1$.

The debiased estimator is a scaled and shifted χ^2. Its variance closes the loop with the previous subsection: $\mathrm{Var}(\widehat{C}_\ell/C_\ell) = A_\ell^2\,\mathrm{Var}(X) = A_\ell^2\, 2\nu = 2(1+r_\ell)^2/(2\ell+1)$, so $\mathrm{Var}(\widehat{C}_\ell) = 2(C_\ell + N_\ell/T_\ell^2)^2/(2\ell+1)$, which is (8.29) and one over the Fisher information (8.31). The distribution (8.32) is the full shape whose width sits exactly on the Cramér–Rao bound. The appendix of Knox [166] derives the same density (his eqs. 31–33). Two features follow at once. At low ℓ, ν is small and the χ^2 is visibly skewed. How skewed follows from its cumulants. A χ^2_ν is a sum of ν independent squared standard Gaussians, cumulants of independent terms add, and each square contributes $\kappa_k = 2^{k-1}(k-1)!$; so $\kappa_k = 2^{k-1}(k-1)!\,\nu$ for the sum, as derived in (7.100). The skewness is the third cumulant over the variance to the power $3/2$,

$$\frac{\kappa_3}{\kappa_2^{3/2}} = \frac{8\nu}{(2\nu)^{3/2}} = \sqrt{\frac{8}{\nu}}, \qquad \text{and likewise the excess kurtosis is} \qquad \frac{\kappa_4}{\kappa_2^2} = \frac{48\nu}{(2\nu)^2} = \frac{12}{\nu}. \tag{8.33}$$

Scaling and shifting the variable, as in (8.32), changes neither ratio. The skewness is 1.07 at $\ell = 3$. And the shift $-r_\ell$ moves part of the distribution below zero. The estimate is negative whenever the observed power happens to fall below the noise power, $X < \nu\, r_\ell/(1+r_\ell)$, which becomes likely once r_ℓ is large.

Example: how often is the estimate negative? At $\ell_{\mathrm{eq}} = 436$, where $r_\ell = 1$, a negative estimate needs $X < \nu/2$, a χ^2_{873} variable more than ten standard deviations below its mean: the probability is 4×10^{-38}. At $\ell = 700$, $r_\ell = 246$ and the threshold $\nu r_\ell/(1+r_\ell)$ sits within a fraction of a standard deviation of the mean ν. The probability of a negative estimate is then 46% (Figure 8.8, right). A negative $\widehat{C}_\ell$ there is not a sign of a bug. The measurement simply contains almost no signal, and the unbiased estimator has to straddle zero.

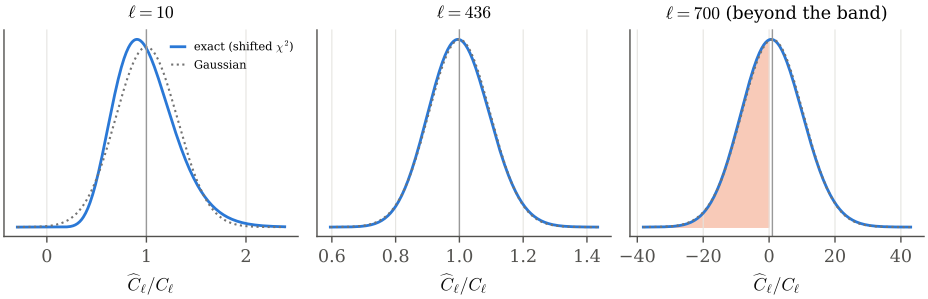


Figure 8.8: The exact distribution of $\widehat{C}_\ell/C_\ell$ for the toy experiment at three multipoles (blue), with the Gaussian of the same mean and width (dotted). At $\ell = 10$ the χ^2 is skewed, with its peak below the mean. At $\ell = 436$ it is close to Gaussian and safely positive. At $\ell = 700$, outside the analysed band, it is a broad Gaussian centred on 1, almost half of which (orange) lies below zero.

8.4.2 Knox's formula on part of the sky

Knox's formula was derived for a complete sky. Every real temperature analysis cuts away the Galaxy, and every ground or balloon experiment sees a patch. The approximation used everywhere for that case divides the variance by the observed sky fraction f_{sky} and, when $\Delta\ell$ neighbouring multipoles are averaged into one bin, by $\Delta\ell$ as well:

$$\sigma\big(\widehat{C}_b\big) \approx \sqrt{\frac{2}{(2\ell+1)\,\Delta\ell\, f_{\text{sky}}}}\left(C_\ell + \frac{N_\ell}{T_\ell^2}\right) \tag{8.34}$$

[Hivon (16)–(17)], [134, eq. 33] (see also the cosmology notes). Each of its two new factors comes from one move applied to the full-sky error (8.29).

- *The bin.* The binned estimate $\widehat{C}_b$ is the plain average of the $\Delta\ell$ values $\widehat{C}_\ell$ in the bin. On the full sky they are independent, so their variances add, and the $1/\Delta\ell$ of the average comes out squared: $\mathrm{Var}(\widehat{C}_b) = \Delta\ell^{-2}\sum_{\ell\in b}\mathrm{Var}(\widehat{C}_\ell)$. If the spectrum changes little across the bin, the $\Delta\ell$ terms are all about equal, the sum is $\Delta\ell$ times one of them, and $2/(2\ell+1)$ becomes $2/[(2\ell+1)\Delta\ell]$.
- *The sky fraction.* The factor $2/(2\ell+1)$ is $2/\nu$, with ν the number of independent modes. A patch covering a fraction f_{sky} of the sky is taken to carry that fraction of the modes, $\nu \to \nu\, f_{\text{sky}}$, which divides the variance by f_{sky}.

The same two moves are written out as labelled chains where we reproduce the papers that use the formula: (8.41) for the bin, (8.40) for the sky fraction. Hivon et al. [168] add a factor w_2^2/w_4 to f_{sky} for a weighted mask, where w_i is the sky average of the i-th power of the mask over the observed region [Hivon (17)].

The plain claim is: *observing half the sky doubles the variance*. The everyday question behind it is, how many independent numbers does a patch of sky provide at multipole ℓ? On the full sky the answer is $2\ell+1$, one per m, and they are independent, which is why the error is $\sqrt{2/(2\ell+1)}$. The formula assumes that a fraction f_{sky} of the sky provides the same fraction of independent numbers, $(2\ell+1)f_{\text{sky}}$, and that each still belongs to a single ℓ. That is the hidden assumption: the modes on the patch are taken to be as clean as on the full sky, only fewer.

It cannot be literally true, and three short moves show why. First, the observed map is the sky times the mask, $s(\hat{\boldsymbol{n}})M(\hat{\boldsymbol{n}})$, with $M = 1$ inside the patch and 0 outside. Second, expand both factors in spherical harmonics. A product $Y_{\ell'm'}Y_{LM}$ is again a sum of spherical harmonics, but of every degree from $|\ell' - L|$ to $\ell' + L$ (the rule for adding angular momenta), so a coefficient of degree ℓ of the cut map collects sky coefficients $a_{\ell'm'}$ from every ℓ' that differs from ℓ by a degree L present in the mask. A patch of angular size θ_{p} has its harmonic content concentrated at $L \lesssim \pi/\theta_{\text{p}}$, so each ℓ of the cut map mixes the sky over a width $\Delta\ell_c \sim \pi/\theta_{\text{p}}$ around it. Third, two neighbouring multipoles of the cut map then share sky modes and are correlated; the power of one ℓ is spread over its neighbours. This is called **mode coupling** [Hivon (14)]. The modes on a patch are shared between multipoles, so the counting is right only on average over a bin wider than $\Delta\ell_c$, which is why the formula comes with a bin width $\Delta\ell$. The coupling is computed exactly for the masked sky in Section 8.10.1.

Example: the counting fails where there is less than one mode. Hivon et al.'s test region covers $f_{\text{sky}} = 1.82\%$ of the sky [Hivon §4.1]. For the quadrupole, $(2\ell+1)f_{\text{sky}} = 0.09$: the formula claims a tenth of one mode and an error bar $\sqrt{2/0.09} \approx 4.7$ times the signal. A patch $40^\circ \times 24^\circ$ cannot contain a single wavelength of the quadrupole, so it cannot tell a quadrupole from a

dipole or from an $\ell = 3$ pattern. The question "how well is C_2 measured?" has no separate answer there, and no counting rule can supply one. Efstathiou [134] shows by comparing with the exact maximum-likelihood Fisher matrix that the $f_{\rm sky}$ formula fails at low ℓ even for large sky fractions.

Example: the counting works where there are many. In the same region, take a bin of $\Delta\ell = 50$ around $\ell = 200$. The number of modes is $(2 \cdot 200 + 1) \times 50 \times 0.0182 = 365$, and (8.34) gives a fractional error of $\sqrt{2/365} = 7.4\%$ on the signal-dominated bin. With the cosine-apodised weighting of their Table 1, $w_2^2/w_4 = 0.514$, the count drops to 188 and the error rises to 10.3%. Hivon et al. found that their simulated error bars match this rule of thumb almost exactly [Hivon (36), §4.2].

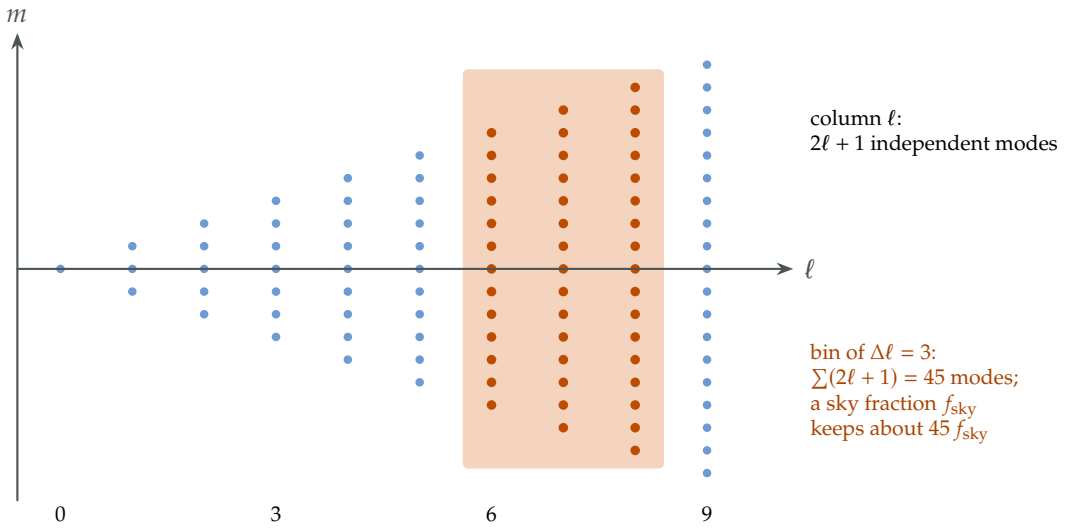


Figure 8.9: Counting modes. Each dot is one independent number on the full sky, one per (ℓ, m); column ℓ has $2\ell + 1$ of them. A bin of $\Delta\ell$ multipoles (orange) contains about $(2\ell + 1)\Delta\ell$. The $f_{\rm sky}$ rule assumes that a patch of sky keeps the fraction $f_{\rm sky}$ of these dots, each still in its own column. A mask actually smears every dot over neighbouring columns, so the rule can hold only for bins wider than the smearing.

Watch out: The $f_{\rm sky}$ formula is a count, not a theorem

Equation (8.34) counts modes. It is reliable only for bins that contain many modes and are wider than the coupling a mask induces between multipoles. At low ℓ, for small patches, and for the correlations between neighbouring bins, it says nothing; there the covariance must come from simulations or from the exact mode-coupling matrix of the masked sky [Hivon §3.1].

8.4.3 Signal-to-noise, per multipole and in total

How well does each multipole measure the spectrum? The natural number is the **signal-to-noise per multipole**, the spectrum divided by its own error bar, from (8.29):

$$\mathrm{SNR}_\ell \equiv \frac{C_\ell}{\sigma(\widehat{C}_\ell)} = \sqrt{\frac{2\ell+1}{2}}\,\frac{C_\ell}{C_\ell + N_\ell/T_\ell^2} = \sqrt{\frac{2\ell+1}{2}}\,\frac{1}{1+r_\ell}. \tag{8.35}$$

With no noise it grows as $\sqrt{\ell}$, because higher multipoles have more modes. The noise cuts it by the factor $1/(1 + r_\ell)$, which falls exponentially once the beam has turned the sky down.

Between these two behaviours there is a best multipole. For the toy experiment it is $\ell = 314$, with $\mathrm{SNR}_\ell = 16.6$; at $\ell_{\rm eq}$ the value has fallen to 10.4 (Figure 8.10, left).

A second question is how well all multipoles together measure the spectrum. Since one number per ℓ is too much to summarise, ask about the simplest possible change of the spectrum: an overall amplitude, $C_\ell = A\, C_\ell^{\rm fid}$ with $A = 1$ for the true sky. Each multipole gives an independent unbiased estimate of A, namely $\widehat{C}_\ell / C_\ell^{\rm fid}$, with variance $1/\mathrm{SNR}_\ell^2$. Independent estimates of one number are best combined with weights proportional to their inverse variances (Chapter 3):

$$\hat{A} = \frac{\sum_\ell \mathrm{SNR}_\ell^2 \, \widehat{C}_\ell / C_\ell^{\rm fid}}{\sum_\ell \mathrm{SNR}_\ell^2},$$
$$\mathrm{Var}(\hat{A}) \overset{(1)}{=} \frac{\sum_\ell \mathrm{SNR}_\ell^4 \, \mathrm{Var}(\widehat{C}_\ell / C_\ell^{\rm fid})}{\left(\sum_\ell \mathrm{SNR}_\ell^2\right)^2} \overset{(2)}{=} \frac{\sum_\ell \mathrm{SNR}_\ell^2}{\left(\sum_\ell \mathrm{SNR}_\ell^2\right)^2} = \frac{1}{\sum_\ell \mathrm{SNR}_\ell^2}. \tag{8.36}$$

(1) The terms are independent (different ℓ on the full sky), so variances add, each multiplied by the square of its weight.

(2) $\mathrm{Var}(\widehat{C}_\ell / C_\ell^{\rm fid}) = 1/\mathrm{SNR}_\ell^2$ by the definition (8.35).

Working definition 8.3 (cumulative signal-to-noise)

$$\mathrm{SNR}(\leq \ell_{\rm max}) \equiv \left[\sum_{\ell=2}^{\ell_{\rm max}} \mathrm{SNR}_\ell^2\right]^{1/2} = \left[\sum_{\ell=2}^{\ell_{\rm max}} \frac{2\ell+1}{2} \frac{1}{(1+r_\ell)^2}\right]^{1/2}$$

[the number of standard deviations by which the spectrum up to $\ell_{\rm max}$ is detected, if only its overall amplitude is fitted]. *Operationally:* one over the error of the amplitude $\hat{A}$ of (8.36).

The same sum appears in Hamimeche and Lewis [16, eq. 21] as the measure of how much information a noisy spectrum carries. It is a sum of squares, so each multipole adds its SNR_ℓ^2, and a multipole whose SNR_ℓ has dropped below one adds almost nothing.

Example: where adding multipoles stops helping. With no noise, $\mathrm{SNR}(\leq \ell_{\rm max})^2 = \sum_{\ell=2}^{\ell_{\rm max}} (2\ell+1)/2 = [(\ell_{\rm max}+1)^2 - 4]/2$, so that $\mathrm{SNR}(\leq 512) = 363$, and the square keeps growing like $\ell_{\rm max}^2$. With the beam and the noise of the toy experiment, the sum reaches 280 at $\ell = 512$, and 281 at $\ell = 767$, the largest multipole the $N_{\rm side} = 256$ maps can hold, where the cosmic-variance limit would be 543. The last 255 multipoles add one unit. Ninety per cent of the final value is already reached at $\ell = 378$ (Figure 8.10, right). This is the sense in which the beam sets the reach of an experiment: beyond a few $\ell_{\rm eq}$ there is nothing left to measure, whatever the number of pixels.

8.5 The Monte Carlo machine

Three formulas now describe the full-sky estimator: its mean (8.19), its variance (8.29) and its distribution (8.32). Each was derived under assumptions: a Gaussian sky, a circular beam, white noise, transforms that are exact on the pixel grid. A simulation makes none of these assumptions on our behalf. It does to fake skies exactly what the instrument and the analysis do to the real one, thousands of times, and reads the statistics off the results. On the full sky the outcome must reproduce the three formulas, and a disagreement means a mistake in a formula or in the code. Once the two agree, the same machine can be pointed at the masked sky, where

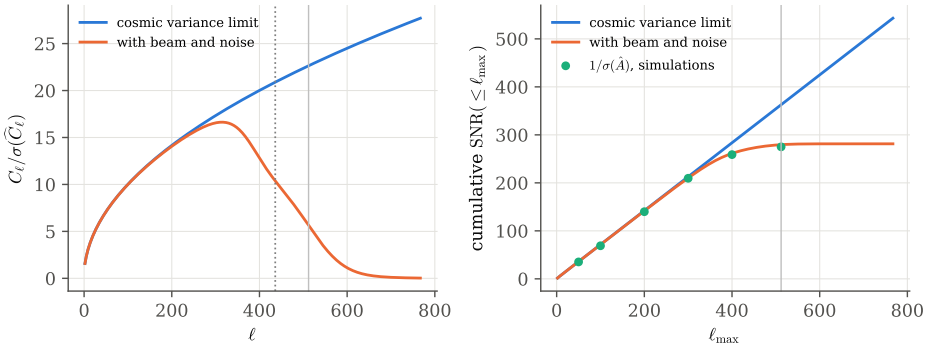


Figure 8.10: Signal-to-noise of the toy experiment. Left: per multipole, against the cosmic-variance limit $\sqrt{(2\ell+1)/2}$ (blue). The noise turns the curve over at $\ell = 314$, before the power equality $\ell_{\rm eq}$ (dotted); the grey line, in both panels, marks the analysed limit $\ell = 512$. Right: cumulative, up to $\ell_{\max}$. Without noise it would keep rising; with the beam and the noise it saturates. The green dots are $1/\sigma(\hat{A})$ measured on the simulated skies of Section 8.6.

no formula is exact. This is step 5 of the estimator recipe of Chapter 3, carried out with the Monte Carlo tools of Chapter 6.

The question to answer before writing any code is: what must be simulated for the simulated $\widehat{C}_\ell$ to have the same distribution as the $\widehat{C}_\ell$ we would compute from the real map? Two kinds of ingredient enter the real $\widehat{C}_\ell$. The random ones, the sky and the noise, must be drawn afresh from their distributions for every simulated sky. The deterministic ones, the beam, the pixels, the transforms, the noise subtraction and the deconvolution, must be applied exactly as they would be to the data, with the same code. Anything left out of the simulation is assumed to have no effect on $\widehat{C}_\ell$.

Working definition 8.4 (Monte Carlo of an observed spectrum)

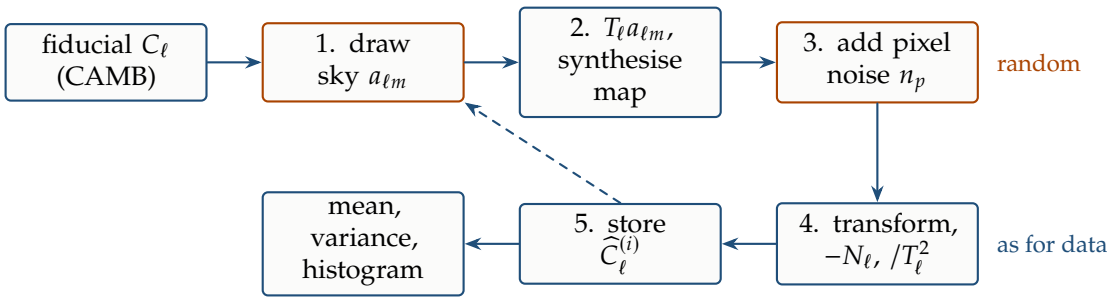
For $i = 1, \ldots, N_{\rm sim}$, with independent random numbers each time:

1. draw a Gaussian sky $a_{\ell m}$ from the fiducial C_ℓ, up to $\ell = 3N_{\rm side} - 1$;
2. multiply by the transfer function, $a_{\ell m} \to T_\ell a_{\ell m}$, and synthesise the map;
3. add an independent Gaussian number of variance σ_p^2 to every pixel;
4. analyse the map exactly as the data: transform, average $|d_{\ell m}|^2$ over m, subtract N_ℓ, divide by T_ℓ^2;
5. store the spectrum $\widehat{C}_\ell^{(i)}$, not the map.

Operationally: the sample mean, sample variance and histogram of $\{\widehat{C}_\ell^{(i)}\}$ estimate the mean, variance and distribution of $\widehat{C}_\ell$ [to a precision set by $N_{\rm sim}$, which is itself computed, Section 8.6].

Figure 8.11 draws the loop. Each step deserves the question: why is it legitimate?

Step 1: the sky. A statistically isotropic Gaussian sky is completely specified by its spectrum (Chapter 7): the $a_{\ell m}$ are independent, $a_{\ell 0} \sim \mathcal{N}(0, C_\ell)$ is real, and for $m > 0$ the real and imaginary parts are independent $\mathcal{N}(0, C_\ell/2)$, with $a_{\ell,-m} = (-1)^m a_{\ell m}^*$ fixing the negative m; after the rescaling of (8.21) these are $2\ell+1$ independent $\mathcal{N}(0, C_\ell)$ numbers. Drawing those numbers



compare with (8.19), (8.29), (8.32)

Figure 8.11: The Monte Carlo loop. The two orange steps draw random numbers, the sky and the noise, afresh for every simulated sky. The blue steps are deterministic and identical to what is done to the real map. The dashed line closes the loop; only the spectra leave it, and their statistics are compared with the formulas of the previous sections.

is therefore a correct draw of a whole sky. We draw them with our own generator rather than `healpy.synalm`, so that every random number goes through the reproducible stream `rng_for("ch08", ...)` (Chapter 6):

Listing 8.1: Drawing one Gaussian sky (`code/ch08/lib_cmbsim.py`)

```
159 def synalm(cl, lmax: int, rng: np.random.Generator) -> np.ndarray:
160     """Gaussian a_lm (healpy m >= 0 layout) with <|a_lm|^2> = C_l, from our own generator.
161
162     a_l0 ~ N(0, C_l) is real; for m > 0, Re and Im ~ N(0, C_l/2) independently.
163     """
164     ell, m = hp.Alm.getlm(lmax)
165     sig = np.sqrt(np.asarray(cl, dtype=float)[ell])
166     re = rng.standard_normal(ell.size)
167     im = rng.standard_normal(ell.size)
168     return np.where(m == 0, sig * re, sig * (re + 1j * im) / np.sqrt(2.0)).astype(complex)
```

Line 165 picks the standard deviation $\sqrt{C_\ell}$ for every stored coefficient; line 168 uses the real part alone for $m = 0$ and divides the complex draw by $\sqrt{2}$ for $m > 0$, so that $\langle |a_{\ell m}|^2 \rangle = C_\ell$ in both cases. With $\ell \le 767$ there are $768 \times 769/2 = 295\,296$ stored coefficients per sky.

Step 2: the instrument. The convolution theorem (Section 8.2.1) makes the beam a multiplication of each $a_{\ell m}$ by B_ℓ; the pixel window is applied the same way, as a factor p_ℓ. The function `observe_signal` is two lines: `almxfl` multiplies by $T_\ell = B_\ell p_\ell$, and `alm2map` synthesises the $N_{\rm side} = 256$ map. One simplification hides here. A real map averages the beamed sky over each pixel, and HEALPix pixels have slightly different shapes, so p_ℓ is only the sky-averaged effect [7]. Applying it as an exact factor assumes every pixel has the average shape. The error this makes is far below the precision of anything we measure, but a real pipeline that cared about it would simulate the pointing and the pixelisation instead [Hivon §3.2].

Step 3: the noise. Independent Gaussian pixel noise is drawn with the depth of the experiment, $\sigma = 14.6\,\mu$K per pixel. To test the cross-spectrum of Section 8.3.3 on the same skies, each simulation draws two half-mission noise maps $n^{(1)}, n^{(2)}$, each with variance $2\sigma^2$ because each half has half the observing time. The full-mission noise is their average, $(n^{(1)} + n^{(2)})/2$, of variance σ^2, which is exactly how co-adding two halves of a real mission works. A further noise map with the pole-heavy variance pattern of Section 8.2.2 tests inhomogeneous noise.

Step 4: the analysis. The map is transformed back with `healpy.map2alm` up to $\ell = 512 = 2N_{\mathrm{side}}$, the range in which the pixel quadrature reproduces the orthogonality of the $Y_{\ell m}$ (Chapter 7). The transform of a pixel map is a quadrature, not an exact integral, and healpy can refine it iteratively. Three iterations make the signal transform accurate to about 10^{-5}; for white noise, whose power fills every multipole up to the pixel scale, one iteration already reproduces its spectrum to the Monte Carlo precision, so the noise maps use one. Then `alm2cl` averages over m and `debias` applies (8.20).

One sky, seven observations. The loop of `simulate` does all of this, and it observes each simulated sky in seven ways at once:

Listing 8.2: The Monte Carlo loop (`code/ch08/lib_cmbsim.py`)

```
229     for i in range(nsim):
230         a = synalm(cl[: exp.lmax_sim + 1], exp.lmax_sim, rng)
231         s = observe_signal(a, exp)
232         n1 = noise_map(exp, rng, scale=np.sqrt(2.0))          # half-mission noise,
    variance 2 sigma^2
233         n2 = noise_map(exp, rng, scale=np.sqrt(2.0))
234         a_s = map2alm(w * s, L, iter_signal)
235         a_1 = map2alm(w * n1, L, iter_noise)
236         a_2 = map2alm(w * n2, L, iter_noise)
237         a_n = 0.5 * (a_1 + a_2)                               # map2alm is linear
238         out["sky"][i] = cross_cl(hp.resize_alm(a, exp.lmax_sim, exp.lmax_sim, L, L))
239         out["sig"][i] = cross_cl(a_s)
240         out["auto"][i] = cross_cl(a_s + a_n)
241         out["cross"][i] = cross_cl(a_s + a_1, a_s + a_2)
242         out["noise"][i] = cross_cl(a_n)
243         if varmap is not None:
244             a_i = map2alm(w * noise_map(exp, rng, varmap=varmap), L, iter_noise)
245             out["inh"][i] = cross_cl(a_s + a_i)
246             out["ninh"][i] = cross_cl(a_i)
```

Lines 234–236 make one signal transform and two half-noise transforms. Because a transform is linear, the full-mission noise needs no transform of its own (line 237), and every spectrum in lines 238–242 is formed from these three sets of coefficients: the sky itself (`sky`, cosmic variance only), the beamed sky without noise (`sig`), the full-mission map (`auto`), the cross-spectrum of the two halves (`cross`) and the noise alone (`noise`); lines 243–246 add the inhomogeneous noise with and without the sky. All seven see the *same* sky. When two estimators are compared on the same simulated skies, the sky's own fluctuation is common to both and cancels in their difference, so the difference is measured far more precisely than either one. This is the method of **common random numbers**, and it is why the bias of the cross-spectrum relative to the auto-spectrum can be measured to a fraction of a per cent with a few thousand skies (Figure 8.12).

The driver script. The run itself is a short script:

Listing 8.3: Running and caching the full-sky Monte Carlo (`code/ch08/05_mc_run.py`)

```
11 import sys, pathlib, time
12 sys.path.insert(0, str(pathlib.Path(__file__).resolve().parents[1]))
13 sys.path.insert(0, str(pathlib.Path(__file__).resolve().parent))
14 import numpy as np
15 from common import save_numbers, rng_for
16 from camb_fiducial import load_fiducial
17 import lib_cmbsim as cs
18
19 N_SIM = 2000
```

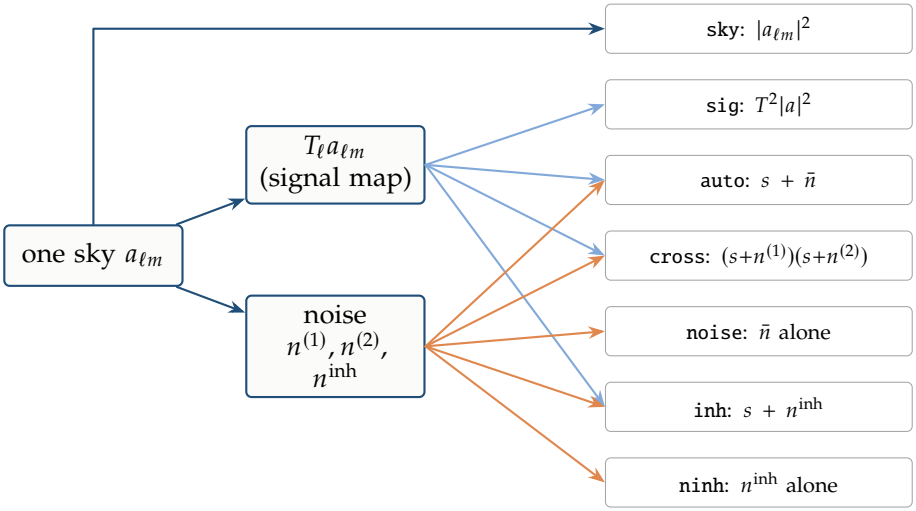


Figure 8.12: Each simulated sky is observed seven ways. Blue lines carry the beamed signal, orange lines the noise; a spectrum fed by both lines contains both. Because the seven spectra share one sky, comparisons between them are free of the sky's own fluctuation.

```
ell, cl = load_fiducial()
exp = cs.Experiment()
varmap = cs.variance_pattern(exp.nside, contrast=4.0)


def make():
    rng = rng_for("ch08", "05_mc_run")
    t0 = time.time()
    d = cs.simulate(exp, cl, N_SIM, rng, varmap=varmap, progress=200)
    d["seconds"] = np.array(time.time() - t0)
    return d


sims = cs.cached("mc_fullsky", make)
secs = float(sims["seconds"])
save_numbers("ch08", "05_mc_run", {
    "EightANsim": N_SIM,
    "EightAMinutes": f"{secs / 60:.0f}",
    "EightASecPerSky": f"{secs / N_SIM:.2f}",
})
print(f"{N_SIM} skies in {secs / 60:.1f} min")
```

The function `cs.cached` runs `make()` once and writes the seven arrays of spectra to the file `mc_fullsky.npz` in `data/ch08/`; every later call reads them back, so the analysis scripts `06`–`10` and `12` all work on the same skies without recomputing them. The run took 6 minutes, 0.19 s per sky, for $N_{\mathrm{sim}} = 2000$.

Example: what the run stores. Seven spectra of 513 multipoles in single precision, for 2000 skies, occupy $7 \times 513 \times 4 \times 2000 \approx 29$ MB. The maps themselves would be 786 432 pixels in double precision, 6.3 MB each, or 88 GB for seven maps of 2000 skies. Storing spectra and not maps is what makes the run fit on a laptop disk; any statistic of $\widehat{C}_\ell$ can still be computed later.

Example: how many skies to plan for. The number of simulations is chosen before the run, from what is to be measured. An error bar estimated from N simulations has a relative Monte Carlo error of $1/\sqrt{2(N-1)}$ (the sampling error of a standard deviation, Chapter 3), so 10%

needs $N = 1 + 1/(2 \times 0.1^2) = 51$ skies. A bias is harder. In a bin of 25 multipoles around $\ell = 400$, one simulated estimate scatters by 1.68%, and the mean of N of them by $1.68\%/\sqrt{N}$; to see a bias of 0.1% at one standard deviation needs $N = (1.68/0.1)^2 \approx 281$ skies. The 2000 skies of the run resolve a bias of about 0.04% per bin.

Example: why $N_{\rm side} = 256$. The resolution must be fine enough to show both regimes of the problem, the cosmic-variance-limited one and the noise-limited one. With a 30′ beam and 200 μK arcmin, the noise takes over at $\ell_{\rm eq} = 436$, inside the trusted range $\ell \le 512$ of $N_{\rm side} = 256$, so this resolution is enough. One pair of transforms there takes 0.04 s on this laptop. The cost grows as $N_{\rm pix}^{3/2} \propto N_{\rm side}^3$: 0.24 s at $N_{\rm side} = 512$, 1.2 s at 1024, and an extrapolated 10 s at 2048, the Planck resolution.

What the laptop version gives up. Each choice above is a reduction of what a research analysis does, and each has a price.

- *Resolution.* Planck analyses its temperature maps at $N_{\rm side} = 2048$ up to $\ell \approx 2500$ [75]; we use $N_{\rm side} = 256$ and $\ell \le 512$. At the same 2000 skies that would take about 19 s per sky, roughly ten hours of the laptop's full processor instead of six minutes, and 0.40 GB per map. The cut keeps 4.2% of the harmonic modes, so a cosmic-variance-limited amplitude would be measured 4.9 times less precisely; the beam was widened to match, so the two regimes of the problem, and every statement about them, are unchanged.
- *Number of simulations.* The Planck 2018 likelihood analysis validates its spectra with, and builds its low-ℓ polarisation likelihood from, 300 end-to-end simulations of the full mission (FFP10) [75]; we use 2000 Gaussian skies. An error bar from 300 simulations carries a Monte Carlo error of 4.1%, ours 1.6%. We can afford more because each of ours is cheap, and we pay for it in realism.
- *Realism of a simulated sky.* An end-to-end simulation generates the timestream of every detector along the real scan, with correlated $1/f$ noise, and runs the real map-maker on it [Hivon §3.2–3.3]; we draw the noise directly in pixels, white and uncorrelated. Our simulations can therefore test only what this model contains: the stripes, the filtering and the noise correlations of a real map are absent.
- *Beam.* Real effective beams are measured on planets, are not circular, and carry their own uncertainty; ours is an exact circular Gaussian. Suppose the beam used in the analysis is wrong by a small fraction $\epsilon = \delta B_\ell / B_\ell$. The estimator $\widehat{C}_\ell = (\widehat{C}_\ell^{\rm obs} - N_\ell)/T_\ell^2$ then divides by $B_\ell^2(1+\epsilon)^2$ instead of B_ℓ^2, which multiplies the estimate by $(1+\epsilon)^{-2} \approx 1 - 2\epsilon$: a relative bias of $-2\,\delta B_\ell / B_\ell$, an effect this pipeline cannot show.

Section 8.8 sets these choices beside the full analysis step by step.

8.6 What the simulations show

The run of Section 8.5 produced 2000 simulated skies of the toy experiment: a 30′ Gaussian beam, $N_{\rm side} = 256$ pixels, 200 μK arcmin of white noise, the Planck 2018 fiducial spectrum. Each sky was observed seven ways and its spectra were stored up to $\ell = 512$. Every formula of the previous sections makes a definite prediction for these numbers. We now take the predictions one at a time, in the order of the sub-questions of Section 8.1: the mean, the width, the shape, and the number of simulations it took to know.

8.6.1 The mean: is the bias gone?

The left panel of Figure 8.13 plots the average over the simulations of the raw observed spectrum $\widehat{C}_\ell^{\,\mathrm{obs}}$, the average of $|d_{\ell m}|^2$ over m on the noisy map. It lies on the prediction $T_\ell^2 C_\ell + N_\ell$ of (8.19) at every multipole: on the beamed sky $T_\ell^2 C_\ell$ at low ℓ, on the flat noise N_ℓ at high ℓ. The noise-only maps give the noise power directly, and their average over all multipoles and all simulations is 1.0001 ± 0.0001 times $\sigma^2 \Omega_{\mathrm{pix}}$, confirming (8.16).

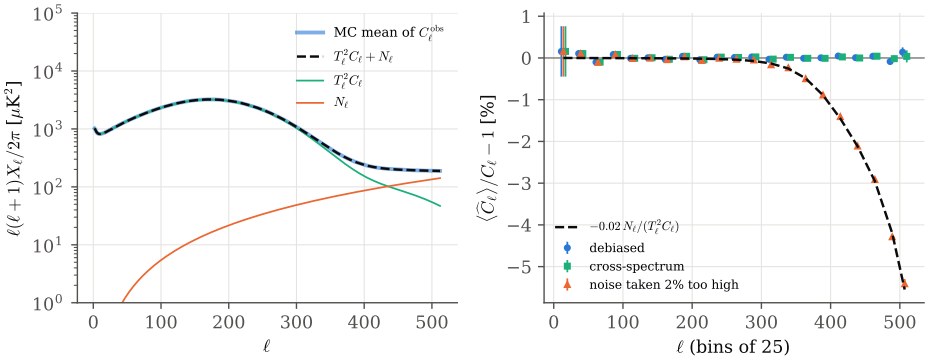


Figure 8.13: The mean of the estimators over 2000 simulated skies. Left: the raw observed spectrum (blue) follows $T_\ell^2 C_\ell + N_\ell$ (dashed), the sum of the beamed sky (green) and the noise (orange). Right: the fractional bias in bins of 25 multipoles. The debiased estimator (blue) and the cross-spectrum of the two half maps (green) scatter around zero within their Monte Carlo error bars. Subtracting a noise level that is 2% too high (orange) produces the negative bias predicted by (8.24) (dashed).

The real test is the debiased estimator (8.20), plotted as a fractional bias in 21 bins of 25 multipoles in the right panel. Its points scatter around zero, the largest by 0.16%, with typical Monte Carlo error bars of 0.04%. Is that zero? Say what was observed: 21 small deviations from zero, each measured with a known Monte Carlo error. What could have produced them? Either the estimator is unbiased and the deviations are the fluctuation of a finite set of simulations, or there is a real bias. Only the first hypothesis predicts how large the deviations should be, so it is the null hypothesis of Chapter 4: if it is true, each deviation divided by its error is a standard Gaussian, and the sum of their squares is a χ^2 variable with 21 degrees of freedom. The observed sum is 20.5. The question is then whether a value this large, or larger, would be unremarkable if there were no bias, and the answer is yes: the probability is $p = 0.49$. The simulations see no bias at the level of 0.04% per bin. The cross-spectrum, with no noise subtracted at all, passes the same test with $\chi^2 = 15.5$ ($p = 0.79$), as (8.26) predicted.

Example: three ways to get it wrong. The same simulations show what the corrections are worth, because the wrong estimators can be computed from the same stored spectra.

- *No noise subtraction.* The bias is +0.16% in the first bin, where the noise is negligible, and +278% in the last, close to the $N_\ell/(T_\ell^2 C_\ell) \approx 3.0$ of Section 8.3.

- *Pixel window forgotten.* Dividing by B_ℓ^2 instead of $B_\ell^2 p_\ell^2$ leaves the factor p_ℓ^2 in the estimate. In the last bin that is a bias of −31%, the 32% power suppression computed in Section 8.2.1.

- *Noise level 2% too high.* The last bin is biased by −5.4%; the prediction $-0.02\, N_\ell/(T_\ell^2 C_\ell)$ of (8.24), averaged over the bin, is −5.5%.

8.6.2 The width: does the variance follow the formula?

For each multipole, divide the sample variance of the 2000 simulated $\widehat{C}_\ell$ by the prediction (8.29). If the formula is right, the ratio is one up to Monte Carlo error. How large is that error? For Gaussian draws the sample variance of n of them is a scaled χ^2_{n-1}, and the variance $2(n-1)$ of that χ^2 gives it a relative error $\sqrt{2/(n-1)}$ (7.97). A heavier tail than the Gaussian makes occasional draws large, and their squares then move the sample variance further; the fourth moment enters, and the relative error becomes $\sqrt{2/(n-1) + \kappa/n}$, with κ the excess kurtosis of one draw (7.99). Our draws are values of $\widehat{C}_\ell$, a scaled and shifted χ^2_ν, so $\kappa = 12/\nu$ by (8.33). With $\nu = 2\ell + 1$ this is negligible except at the lowest multipoles; at $\ell = 100$ the expected spread of the ratio is 3.2%.

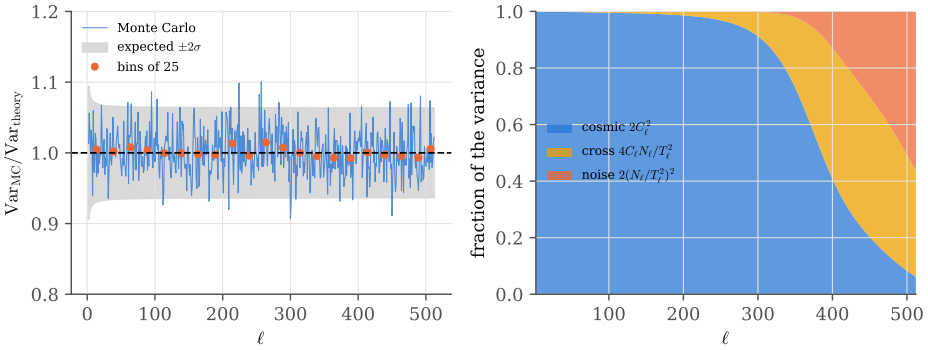


Figure 8.14: The scatter of the debiased estimator. Left: Monte Carlo variance divided by the prediction $2(C_\ell + N_\ell/T_\ell^2)^2/(2\ell+1)$ at every multipole (blue), with the band in which 95% of the points should fall (grey) and bin averages (orange). Right: the share of the three terms of (8.30) in the variance. The cosmic variance (blue) gives way to the sky–noise cross term (yellow) and then to the pure noise term (orange).

The left panel of Figure 8.14 shows the result. Averaged over $\ell = 2$–512, the ratio is 1.001 ± 0.001. The sum of the squared standardised deviations is 498 for 511 multipoles ($p = 0.65$), and 5.1% of the multipoles fall outside the two-standard-deviation band, against the 4.6% expected. The sky-only spectra, observed with no instrument, give a variance ratio of 1.001 against the cosmic variance of Chapter 7. The right panel splits the prediction into the three terms of (8.30). At $\ell = 512$ the pure noise term carries 56% of the variance and the cross term 37%; the noise terms pass half of the total at $\ell = 386$, the multipole where $r_\ell = \sqrt{2} - 1$.

Example: the price of the cross-spectrum. Section 8.3.3 predicted that the cross-spectrum of two half maps, which needs no noise model, scatters more than the auto-spectrum, by the factor $\sqrt{[(S+2N)^2 + S^2]/[2(S+N)^2]}$ with $S = T_\ell^2 C_\ell$ and $N = N_\ell$. Because both estimators are computed on the same simulated skies, the ratio of their scatters is measured with the sky's fluctuation cancelled. Figure 8.15 shows it following the prediction from 1 at low ℓ towards $\sqrt{2}$; at the top of the band the predicted factor is 1.25 and the simulations give 1.24.

Example: the amplitude of the whole spectrum. The cumulative signal-to-noise of Section 8.4.3 predicts that the inverse-variance weighted amplitude $\hat{A}$ of (8.36), built from all multipoles up to 512, has an error $1/\mathrm{SNR}(\le 512) = 0.00358$. On the simulated skies its scatter is 0.00364, with a Monte Carlo error of 1.6%, and its mean is 1.00004 ± 0.00008. The points in the right panel of Figure 8.10 repeat the comparison for other $\ell_{\max}$.

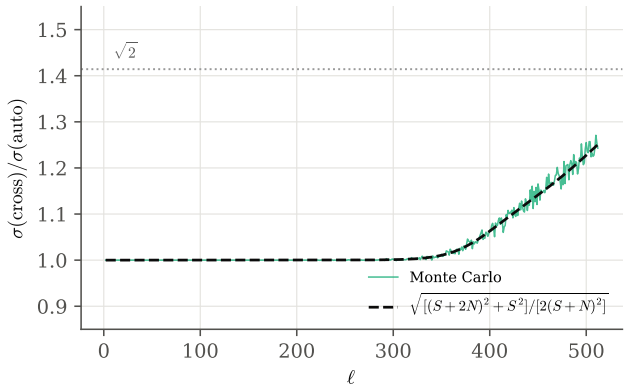


Figure 8.15: Scatter of the cross-spectrum of two half maps divided by that of the auto-spectrum of the full map, on the same simulated skies (green), and the prediction (dashed). The ratio is one where the sky dominates and climbs towards $\sqrt{2}$ (dotted) where the noise does.

8.6.3 The shape: shifted χ^2 or Gaussian?

Figure 8.16 compares the histograms of $\widehat{C}_\ell/C_\ell$ at four multipoles with the shifted, scaled χ^2 of (8.32) and with a Gaussian of the same mean and width. The Kolmogorov–Smirnov test of Chapter 4 measures the largest distance between the cumulative distribution of the simulations and that of a model; its p-values for the shifted χ^2 are 0.09, 0.19, 0.64 and 0.10 at $\ell = 3$, 30, 436 and 512, none of them small. At $\ell = 3$ the measured skewness is 1.07, the χ^2_7 value is 1.07, and the Gaussian, which has no skewness, is rejected with $p = 1 \times 10^{-5}$. By $\ell = 30$ the two models are hard to tell apart by eye, and by ℓ_{eq} they coincide.

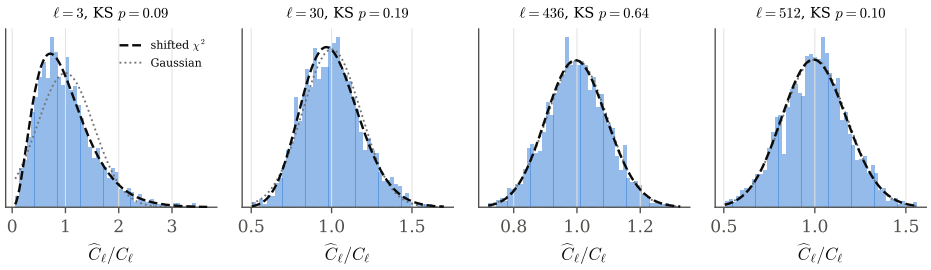


Figure 8.16: Histograms of $\widehat{C}_\ell/C_\ell$ over the simulated skies at four multipoles, with the shifted χ^2 (dashed) and the Gaussian of the same mean and width (dotted). At $\ell = 3$ the distribution is strongly skewed and only the χ^2 follows it; at higher ℓ the two models merge.

8.6.4 When the noise is not uniform

Real scans revisit some parts of the sky more often than others. The simulations include a noise map whose pixel variance follows the pattern $\sigma_p^2 \propto 1/(1 + 4\cos^2\theta)$, the inverse of a hit count that is five times larger at the poles than at the equator, normalised so that its sky average equals the uniform σ^2. The relative variance then runs from 0.36 at the poles to 1.81 at the equator (Figure 8.17, left).

Two predictions were made in Section 8.2.2. The *bias* is Ω_{pix} times the mean pixel variance (8.17), whatever the pattern: the simulations give 0.9999 ± 0.0001 times the uniform value. The *scatter* depends on the full pattern. For the noise alone, write its harmonic covariance $N_{mm'} \equiv \langle n_{\ell m} n^*_{\ell m'} \rangle$ at fixed ℓ, from (8.14). A pattern that depends only on θ does not couple

different m, because the ϕ integral of $e^{i(m'-m)\phi}$ vanishes, so $N_{mm'} = \delta_{mm'} N_{mm}$. The modes are still independent Gaussians, but with different variances N_{mm}, and

$$\operatorname{Var}\Big(\frac{1}{2\ell+1}\sum_m |n_{\ell m}|^2\Big) \overset{(1)}{=} \frac{2}{(2\ell+1)^2}\sum_m N_{mm}^2 \overset{(2)}{\geq} \frac{2}{(2\ell+1)^3}\Big(\sum_m N_{mm}\Big)^2 \overset{(3)}{=} \frac{2\,\bar{N}_\ell^2}{2\ell+1}. \tag{8.37}$$

(1) Each of the $2\ell+1$ real degrees of freedom of mode m is Gaussian with variance set by N_{mm}, so, as in (8.29), each contributes $2N_{mm}^2$; the reality condition $n_{\ell,-m} = (-1)^m n_{\ell m}^*$ pairs m with $-m$ and is already counted in the sum over all m.

(2) Cauchy–Schwarz, $\big(\sum_m 1 \cdot N_{mm}\big)^2 \leq (2\ell+1)\sum_m N_{mm}^2$.

(3) $\bar{N}_\ell = \sum_m N_{mm}/(2\ell+1)$ is the average noise power of (8.17).

The right-hand side is the white-noise variance with the same mean power. Non-uniform noise therefore never decreases the scatter at fixed bias, and increases it by the factor $(2\ell+1)\sum_m N_{mm}^2/(\sum_m N_{mm})^2$, a measure of how unequal the mode variances are. Script `09_inhomogeneous.py` computes N_{mm} exactly for our pattern from (8.14), ring by ring, and predicts a factor 1.053 above $\ell = 100$; the simulations give 1.055. With the sky added, each mode has variance $T_\ell^2 C_\ell + N_{mm}$, the prediction becomes $\sum_m (T_\ell^2 C_\ell + N_{mm})^2$ over its uniform value, 1.030 at $\ell = 512$, and the simulations give 1.024; below $\ell = 100$, where the sky dominates, the ratio is 1.000.

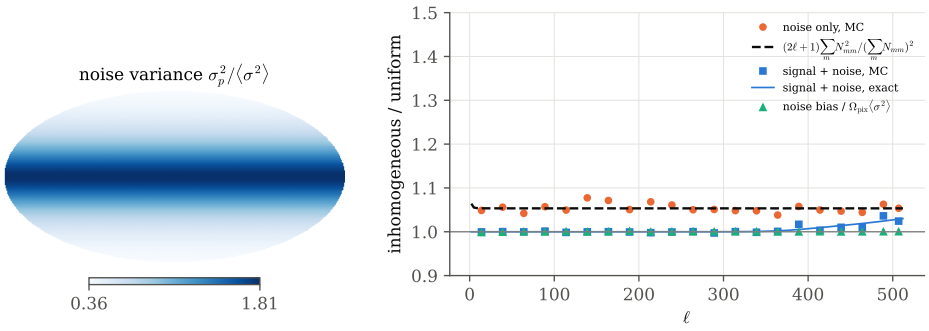


Figure 8.17: Inhomogeneous noise. Left: the relative pixel noise variance of the simulated scan, low at the well-observed poles. Right: inhomogeneous over uniform noise, for the noise bias (green, equal to one), the scatter of the noise spectrum (orange points against the exact prediction, dashed) and the scatter of the full spectrum (blue). The scatter rises by a few per cent only, because each harmonic mode averages the noise over a large part of the sky.

Example: why the increase is only five per cent. A naive guess would be that the variance of the noise spectrum grows like the sky average of σ_p^4 over $\langle\sigma^2\rangle^2$, which for this pattern is 1.230. The exact factor $(2\ell+1)\sum_m N_{mm}^2/(\sum_m N_{mm})^2$ is much closer to one, because N_{mm} is not a pixel variance. By (8.14) it is $N_{mm} = \Omega_{\rm pix}^2 \sum_p \sigma_p^2 |Y_{\ell m}(\hat{\boldsymbol{n}}_p)|^2$: the pixel variance weighted by the mode's own $|Y_{\ell m}|^2$, which integrates to one over the sky, so each N_{mm} is $\Omega_{\rm pix}$ times an average of σ^2 over the sky. Measure σ^2 in units of its sky mean, $\sigma^2(\theta) = c/(1+4\cos^2\theta)$ with $c = 1.81$, and let $x = \cos\theta$.

- The $m = 0$ mode spreads along whole meridians: averaged over its oscillations, $|Y_{\ell 0}|^2$ weights x by $1/(\pi\sqrt{1-x^2})$, which reaches the poles. Writing $x = \sin\phi$ turns the average into $\frac{1}{\pi}\int_0^\pi c\,\mathrm{d}\phi/(1+4\sin^2\phi) = c/\sqrt{5}$, so $N_{00} \approx 0.81$ in these units, although the pole pixels themselves have 0.36.

- The $|m| = \ell$ modes are confined to a narrow band around the equator, so they see only the equatorial value, $N_{\ell\ell} \approx c = 1.81$.
- In between, a mode with $\mu = |m|/\ell$ reaches up to $|x| = \sqrt{1-\mu^2}$, and the same average gives $N_{mm} \approx c/\sqrt{5-4\mu^2}$.

So the mode variances run only from 0.81 to 1.81, while the pixel variances run from 0.36 to 1.81. For large ℓ the m are spread evenly in μ, the sums over m become integrals over μ, and

$$\frac{(2\ell+1)\sum_m N_{mm}^2}{\left(\sum_m N_{mm}\right)^2} \approx \frac{\int_0^1 c^2\,\mathrm{d}\mu/(5-4\mu^2)}{\left[\int_0^1 c\,\mathrm{d}\mu/\sqrt{5-4\mu^2}\right]^2} = \frac{3.262\times 0.323}{1^2} = 1.053,$$

the denominator being one because the average of N_{mm} over m is the sky mean (8.17). This is the exact prediction 1.053 quoted above. The naive $\langle\sigma^4\rangle/\langle\sigma^2\rangle^2$ is larger because it treats every pixel as a separate mode with its own variance; a harmonic mode instead averages many pixels, quiet and noisy together, and only the leftover difference between polar-reaching and equatorial modes counts.

8.6.5 How many simulations did it take?

The last sub-question of Section 8.1 is about the method itself. Figure 8.18 follows the estimates in one bin, $\ell = 400$–424, as the number of simulated skies N grows. The running estimate of the bias wanders inside the envelope $\pm\sigma/\sqrt{N}$, with $\sigma = 1.68\%$ the scatter of one binned estimate. The relative error of the estimated error bar, measured as the spread over 200 random reorderings of the skies, falls as $1/\sqrt{2(N-1)}$: at $N = 107$ it is 6.9%, against the predicted 6.9%. An error bar good to 10% needs about 51 skies, a bias measured to 0.1% about 281.

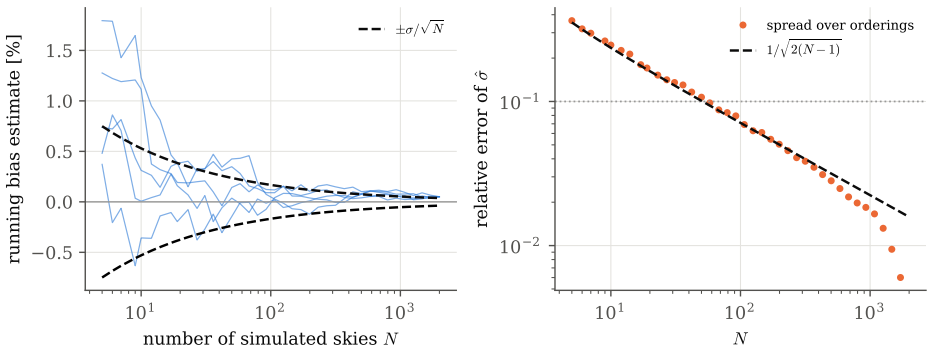


Figure 8.18: Convergence with the number of simulations, in the bin $\ell = 400$–424. Left: five running estimates of the fractional bias, each from a different ordering of the skies, inside the $\pm\sigma/\sqrt{N}$ envelope (dashed). Right: the relative error of the estimated error bar, from the spread over 200 orderings (points), and $1/\sqrt{2(N-1)}$ (dashed); the dotted line marks 10%. Near the 2000 skies available the points drop below the line, because every ordering then draws nearly the same skies.

Key idea: The full-sky verdict

On the full sky with a beam, a pixel window and white or inhomogeneous noise, the simulations reproduce the bias (8.19), the variance (8.29), the distribution (8.32) and the inhomogeneous-noise scatter (8.37), each to its Monte Carlo precision. The machine is therefore trusted where no formula exists.

8.7 Reproducing the literature

Two papers set up the full-sky problem. Knox [166] wrote down the error bar of a spectrum measured from a full-sky map with a Gaussian beam and white noise, and used it to forecast what a satellite could learn about inflation. Hivon et al. [168] built the MASTER method, the pseudo-C_ℓ estimator calibrated by Monte Carlo, and tested it on simulated observations of the Boomerang balloon. Both papers quote numbers that our code can recompute. Agreement is the check that the machine of Section 8.5 measures what the field measures; a disagreement has to be explained, not ignored.

8.7.1 Knox (1995): the error bar of a beamed, noisy map

His question. In 1995 COBE had measured the large-angle anisotropy, and satellites that would map the whole sky at much finer resolution were being proposed. Their promise was to measure C_ℓ out to $\ell \sim 1000$, and through it the shape of the primordial spectrum that inflation predicts. Knox asked how well such a map could actually pin down C_ℓ, given that any real instrument blurs the sky with a beam of finite width and adds detector noise. He wanted one formula that gives the error bar at every ℓ from a few numbers describing the instrument, so that designs could be compared before any of them was built.

His setup. Four ingredients enter, and each has a plain meaning.

- *The sky.* A Gaussian random field with a spectrum C_ℓ from a cosmological model. Knox used the standard cold-dark-matter model of the time, normalised to the quadrupole COBE measured; we use the Planck 2018 best fit (Table 8.1).
- *The beam.* The telescope sees each direction blurred by a Gaussian beam, given by its full width at half maximum (FWHM).
- *The pixel noise.* The map is stored in pixels of solid angle $\Omega_{\rm pix}$, and each pixel carries detector noise of standard deviation $\sigma_{\rm pix}$, independent from pixel to pixel. For a year of observation he took $\sigma_{\rm pix} = 22\,\mu$K in $20' \times 20'$ pixels [Knox1995 p. 7].
- *The sky fraction.* A satellite sees the whole sky, $f_{\rm sky} = 1$.

What he forecasts is the standard deviation ΔC_ℓ of the estimated spectrum at each ℓ, and from it the errors on the inflationary parameters. The noise of a pixel and its size combine into one number, $w^{-1} = \sigma_{\rm pix}^2 \Omega_{\rm pix}$: a pixel twice as large has half the noise variance per unit area, so only this product matters. For his satellite $w^{-1} = 22^2 \times (1/3)^2\,\mu\text{K}^2\,\text{deg}^2$, so $w^{-1/2} = 7.33\,\mu\text{K deg}$; he quotes 7.5, the same number to the precision of his $\sigma_{\rm pix}$.

The sky rms, and the signal-to-noise per pixel. How large is the signal in one pixel, compared with the noise? Knox gives the rms of the sky seen through his $20'$ beam as $92.5\,\mu$K. It follows in three moves. The temperature in direction $\hat{\boldsymbol{n}}$ is the beamed sum of the harmonics, $T(\hat{\boldsymbol{n}}) = \sum_{\ell m} B_\ell a_{\ell m} Y_{\ell m}(\hat{\boldsymbol{n}})$ by the convolution theorem (8.6). The map is real, so $T^2 = TT^*$, and

$$\langle T^2 \rangle \overset{(1)}{=} \sum_{\ell m} \sum_{\ell' m'} B_\ell B_{\ell'} \langle a_{\ell m} a^*_{\ell' m'} \rangle\, Y_{\ell m}(\hat{\boldsymbol{n}}) Y^*_{\ell' m'}(\hat{\boldsymbol{n}}) \overset{(2)}{=} \sum_\ell B_\ell^2 C_\ell \sum_m |Y_{\ell m}(\hat{\boldsymbol{n}})|^2 \overset{(3)}{=} \sum_\ell \frac{(2\ell+1)\, C_\ell\, W_\ell}{4\pi},$$

(1) Multiply the sum by its complex conjugate and average over skies; only the $a_{\ell m}$ are random.
(2) Different (ℓ, m) are uncorrelated and $\langle |a_{\ell m}|^2 \rangle = C_\ell$, so only $\ell' = \ell$, $m' = m$ survives.

	Knox (1995)	ours
cosmology	standard CDM: $h = 0.5$, $\Omega_b h^2 = 0.0125$, no Λ, $n_S = 0.94$, tensors $r = 0.28$, COBE-normalised [Knox1995 p. 4–5]	Planck 2018 best fit
sky fraction	1	1
pixel noise	22 μK in $20' \times 20'$ pixels	the same
noise weight $w^{-1/2}$	7.5 μK deg	7.33 μK deg
beam FWHM	$20'$	$20'$
Fig. 2 experiments	$20'$ and $40'$, $w^{-1/2} = 15$ and 30 μK deg	the same; maps for $w^{-1/2} = 15$

Table 8.1: Knox's experiment and ours. Only the cosmology differs.

(3) The addition theorem $\sum_m |Y_{\ell m}|^2 = (2\ell + 1)/4\pi$, used in (8.17); Knox writes the beam window B_ℓ^2 as W_ℓ.

The direction has dropped out: every pixel has the same signal variance. Our fiducial spectrum gives $\langle T^2 \rangle^{1/2} = 95\,\mu$K (script `11_knox_reproduction.py`). Divided by the noise per pixel, his rms gives $92.5/22 = 4.2$, the "signal-to-noise of about 4 per pixel" he quotes; ours gives 4.3. The number says that the map is signal-dominated pixel by pixel, but not by a wide margin: the noise is a quarter of the signal. Most of the pixel rms comes from the large scales, so at the small scales the beam has already turned down, the noise takes over, and that is why the beam and the noise set the error bars at high ℓ. That two such different cosmologies give rms values within a few per cent is a consequence of both being normalised to the measured large-scale amplitude.

His notation, and ours. Knox's map has $N_{\rm pix}$ pixels of solid angle $\Omega_{\rm pix} = 4\pi/N_{\rm pix}$, each carrying the beamed sky plus a noise of variance $\sigma^2_{\rm pix}$, independent from pixel to pixel. His map coefficients $a_{\ell m}^{\rm MAP}$ are our observed $d_{\ell m}$ of Section 8.3; his $C_\ell^{\rm MAP} = \sum_m |a_{\ell m}^{\rm MAP}|^2/(2\ell+1)$ is our $\widehat{C}_\ell^{\rm obs}$ of (8.18); his $C_\ell^{\rm est}$ is our debiased, deconvolved $\widehat{C}_\ell$ of (8.20); and his ΔC_ℓ is its standard deviation [Knox1995 (4), (25)–(30)]. Three instrument quantities appear in his formulas.

- The noise variance of each harmonic coefficient. By (8.15) it is $\sigma^2_{\rm pix}\Omega_{\rm pix}$, which Knox writes $4\pi\sigma^2_{\rm pix}/N_{\rm pix}$ (his eq. 27), the same number. He calls its inverse the weight per solid angle, $w \equiv (\sigma^2_{\rm pix}\Omega_{\rm pix})^{-1}$, so his w^{-1} is our noise power N_ℓ of (8.16).

- The beam window. His map is smoothed by a Gaussian beam of width $\sigma_b = {\rm FWHM}/\sqrt{8\ln 2}$ (8.7), so the sky power is multiplied by $B_\ell^2 = e^{-\ell(\ell+1)\sigma_b^2}$ (8.10), which he writes as $e^{-\ell^2\sigma_b^2}$.

- No pixel window: his footnote states that the formula ignores pixelisation, so $p_\ell = 1$ and our transfer function (8.12) is $T_\ell = B_\ell$.

His equation (5), derived. Knox's eq. (5) is our variance (8.29) rewritten in these quantities. That variance came from counting: at each ℓ the map supplies $\nu = 2\ell + 1$ independent Gaussian numbers u_k (8.21) of variance $T_\ell^2 C_\ell + N_\ell$, and the scaled sum of their squares is a χ^2_ν variable

with variance 2ν. Then

$$(\Delta C_\ell)^2 \equiv \mathrm{Var}\big(\widehat{C}_\ell\big) \overset{(1)}{=} \frac{2}{2\ell+1}\Big(C_\ell + \frac{N_\ell}{T_\ell^2}\Big)^2 \overset{(2)}{=} \frac{2}{2\ell+1}\Big(C_\ell + \frac{w^{-1}}{B_\ell^2}\Big)^2$$
$$\overset{(3)}{=} \frac{2}{2\ell+1}\Big(C_\ell + w^{-1}e^{\ell^2\sigma_b^2}\Big)^2, \tag{8.38}$$

$$\frac{\Delta C_\ell}{C_\ell} \overset{(4)}{=} \sqrt{\frac{2}{2\ell+1}}\left(1 + \frac{w^{-1}e^{\ell^2\sigma_b^2}}{C_\ell}\right) \overset{(5)}{=} \sqrt{\frac{2}{2\ell+1}}\left(1 + \frac{\ell^2\, w^{-1}e^{\ell^2\sigma_b^2}}{\ell^2 C_\ell}\right). \tag{8.39}$$

(1) The full-sky variance (8.29).

(2) $N_\ell = \sigma_{\mathrm{pix}}^2\Omega_{\mathrm{pix}} = w^{-1}$ by (8.16), and $T_\ell = B_\ell$ because the pixel window is ignored.

(3) $1/B_\ell^2 = e^{\ell(\ell+1)\sigma_b^2}$, with $\ell(\ell+1)$ replaced by ℓ^2. The two differ by a factor $e^{\ell\sigma_b^2}$, which for a 20′ beam ($\sigma_b = 8.5' = 2.47\times10^{-3}$ rad) is 1.006 at $\ell = 1000$, and 1.025 for a 40′ beam. This line is Knox's eq. (4).

(4) Take the square root (both factors are positive) and divide by C_ℓ.

(5) Multiply the numerator and the denominator of the noise term by ℓ^2. Knox chose this form because $\ell^2 C_\ell$ changes by less than a factor of ten between $\ell = 2$ and $\ell \approx 1000$, so the ℓ-dependence of the noise term is carried visibly by $\ell^2 e^{\ell^2\sigma_b^2}$ [Knox1995 p. 8].

Equation (8.39) is his eq. (5). The first factor is the cosmic variance of Chapter 7; the bracket says by how much the detector inflates it, and the inflation grows like $e^{\ell^2\sigma_b^2}$ because the beam turns down the sky and leaves the noise alone (Section 8.2.2). On part of the sky the mode count of Section 8.4.2 replaces $\nu = 2\ell+1$ in step 1 by $(2\ell+1)f_{\mathrm{sky}}$, which multiplies (8.39) by $1/\sqrt{f_{\mathrm{sky}}}$:

$$\frac{\Delta C_\ell}{C_\ell} \approx \sqrt{\frac{2}{(2\ell+1)\,f_{\mathrm{sky}}}}\left(1 + \frac{w^{-1}e^{\ell^2\sigma_b^2}}{C_\ell}\right), \tag{8.40}$$

the single-multipole case $\Delta\ell = 1$ of (8.34). Knox's satellite sees the whole sky, $f_{\mathrm{sky}} = 1$, and his figure uses (8.39) as it stands.

His Fig. 2. Knox plotted (8.39) for four experiments: beams of 20′ and 40′, each with $w^{-1/2} = 15$ and 30 μK deg (the noise degraded by a factor of two to four by foreground removal) [Knox1995 p. 9–10]. The arXiv version of the paper carries the captions but not the plotted curves, so the comparison is made with his formula itself, evaluated for our fiducial spectrum. The left panel of Figure 8.19 redraws his figure: for each pair (FWHM, $w^{-1/2}$) it plots the right-hand side of (8.39) with C_ℓ the Planck 2018 fiducial spectrum, σ_b from the FWHM, and $w^{-1} = (w^{-1/2}\times\pi/180)^2\,\mu\mathrm{K}^2\,\mathrm{sr}$, for $2 \le \ell \le 1000$. His conclusion, that at fixed w the smaller beam is more precise at every ℓ, holds at every multipole up to 1000 for both noise levels. It is visible in (8.39) directly: at fixed w, a smaller σ_b lowers $e^{\ell^2\sigma_b^2}$ at every ℓ and changes nothing else.

His formula, tested with maps. Knox never made a map. He drew each C_ℓ^{est} directly from its $\chi^2_{2\ell+1}$ distribution [Knox1995 p. 8]. A map contains things a formula can leave out, so we simulated the two $w^{-1/2} = 15\,\mu$K deg experiments as full maps. Script `11_knox_reproduction.py` does five things for each experiment:

1. draws a sky from the fiducial C_ℓ and smooths it with the 20′ or 40′ beam;
2. stores it in pixels at $N_{\mathrm{side}} = 512$, which applies the pixel window p_ℓ;

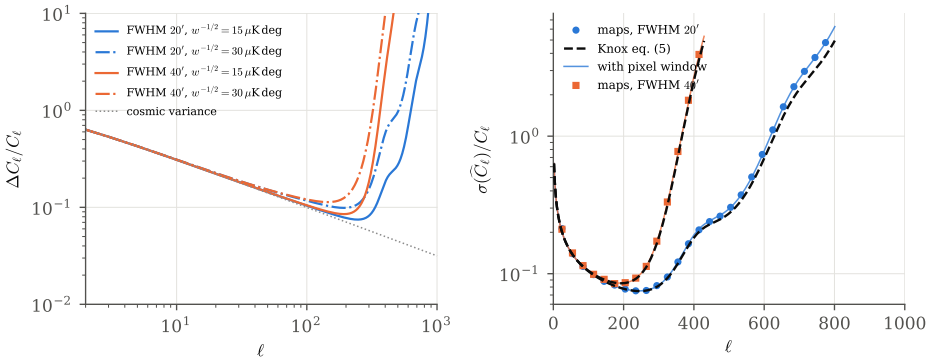


Figure 8.19: Knox's Fig. 2, redone. Left: $\Delta C_\ell / C_\ell$ of (8.39) (his eq. 5) for his four experiments, with our fiducial spectrum: the 20′ beam (blue) beats the 40′ beam (orange) at every ℓ for both noise levels (solid and dash-dotted). Right: the two $w^{-1/2} = 15\,\mu$K deg experiments simulated with maps at $N_{\rm side} = 512$. The simulated scatter (points) follows Knox's formula (dashed) until the pixel window, which his formula leaves out, becomes visible; the formula with the pixel window divided out (thin solid) follows the points everywhere.

3. adds white noise of depth $w^{-1/2} = 15\,\mu$K deg $= 900\,\mu$K arcmin;

4. computes $\widehat{C}_\ell$ with (8.20) up to $\ell = 1000$, dividing by the full $T_\ell^2 = B_\ell^2 p_\ell^2$;

5. repeats this for 300 skies (in 5.0 minutes), takes the standard deviation of $\widehat{C}_\ell$ over the skies at each ℓ, and divides it by two predictions: (8.29) with the pixel window included, and Knox's (8.38) as printed.

The right panel of Figure 8.19 shows the result, and Table 8.3 lists the numbers. Against (8.29), the simulated scatter averaged over ℓ is 1.000 times the prediction for the 20′ beam and 0.999 for the 40′ beam, where the Monte Carlo error is 4.1% per multipole. Against Knox's formula as printed, the median ratio is 1.053 for the 20′ beam and 1.013 for the 40′ beam, and for the 20′ beam it climbs to 1.25 near $\ell = 801$.

Why his formula and the maps differ. The cause is the pixel window. Knox's footnote says that his formula ignores pixelisation, which is harmless when the pixels are several times smaller than the beam. Our 6.9′ pixels are three times smaller than the 20′ beam, yet by $\ell \approx 800$ the pixel window has turned the signal down by a further $p_\ell^2 \approx 0.8$. Dividing it out turns the noise on the sky, N_ℓ / T_ℓ^2, up by $1/p_\ell^2 \approx 1.25$. Near $\ell = 801$ the noise term dominates the bracket of (8.29), so the error bar rises by about the same factor, which is the 1.25 the maps show. The 40′ experiment loses its signal to the beam well before the pixel window matters, so its ratio stays close to one. A formula can be correct for the assumptions it states and still not describe a pixelised map.

His shortcut, and what it teaches about cost. Drawing 300 sets of spectra directly from the shifted χ^2 of (8.32) takes 23 ms, against minutes for the maps, and gives the same scatter to a ratio of 0.999. On the full sky with white noise the map step adds nothing, because the distribution of $\widehat{C}_\ell$ is known exactly. Maps become necessary as soon as something breaks that distribution: a mask, inhomogeneous or correlated noise, an asymmetric beam. The cheapest simulation that still contains the effect under study is the right one, and that choice is what keeps a research pipeline on a laptop.

8.7.2 Hivon et al. (2002): MASTER in the full-sky limit

Their question. By 2002 balloon experiments such as Boomerang had maps with hundreds of thousands of pixels, far beyond the reach of the likelihood of Section 8.1.1. The cheap alternative is the **pseudo-spectrum** $\tilde{C}_\ell$: compute the spectrum of the observed map as if it were a complete sky. But the map covers only a patch, has been filtered to remove slow detector drifts, is blurred by the beam and contains noise, so on average $\tilde{C}_\ell$ is not C_ℓ: it is biased, and power from neighbouring multipoles is mixed into it. Hivon et al. asked how to correct this cheap estimator using nothing but simulations of the experiment, and how many simulations that takes.

Their setup. Each thing the observation does to the sky gets one symbol, and it helps to meet them one at a time.

- *The mask.* Only a patch of sky is kept. Cutting the sky mixes multipoles, and the mixing is described by a **mode-coupling matrix** $M_{\ell\ell'}$: how much of the true power at ℓ' ends up in the pseudo-spectrum at ℓ.
- *The filter.* The time-ordered data are high-pass filtered before the map is made, which removes large-scale power by a factor F_ℓ at each ℓ, measured from noise-free simulations.
- *The beam and the pixel.* Their window B_ℓ is the beam times the pixel window, our T_ℓ of (8.12).
- *The noise.* Its average power in the pseudo-spectrum, $\langle \tilde{N}_\ell \rangle$, is measured from noise-only simulations.

Their test simulated the Boomerang long-duration flight (Table 8.2) [Hivon §4.1].

	Hivon et al. (2002)	ours
beam FWHM	10′	30′
map resolution	$N_{\rm side} = 512$	$N_{\rm side} = 256$
noise	measured spectrum of one detector, with its $1/f$ rise and scan lines	white, uniform
filter	sharp high-pass	none
sky region	ellipse, $f_{\rm sky} = 1.82\%$	full sky
simulations	450 noise-only, 250 noise-free, 1350 signal plus noise	2000 signal plus noise

Table 8.2: The MASTER test and ours.

Their notation, and ours. Their pseudo-spectrum $\tilde{C}_\ell$ of the observed map is our $\widehat{C}_\ell^{\rm obs}$ of (8.18) when the sky is complete; their B_ℓ is our $T_\ell = B_\ell p_\ell$; their $\langle \tilde{N}_\ell \rangle$ is our average noise power $\bar{N}_\ell$ of (8.17); their C_b and N_b are our spectrum and noise power averaged over a bin b of $\Delta\ell$ multipoles; their F_ℓ and $M_{\ell\ell'}$ have no counterpart on our full, unfiltered sky.

Their master equation. Why does the average pseudo-spectrum take the form of a matrix acting on C_ℓ? Three moves give it.

1. The observed sky is the true sky times the mask, $s(\hat{\boldsymbol{n}})M(\hat{\boldsymbol{n}})$, with $M = 1$ on the patch and 0 elsewhere.

2. In harmonic space a product becomes a sum: each coefficient of the masked sky is a sum of the true coefficients $a_{\ell'm'}$ over all ℓ', weighted by the harmonics of the mask itself (Section 8.4.2). So one ℓ of the pseudo-spectrum receives power from several ℓ'.

3. Square the coefficients and average over skies. Different true coefficients are uncorrelated, so only the diagonal terms $\langle|a_{\ell'm'}|^2\rangle = C_{\ell'}$ survive. What remains is a fixed matrix, set by the mask alone, acting on the vector of $C_{\ell'}$: $M_{\ell\ell'}$.

The filter and the beam act on the sky before the mask, so they multiply each $C_{\ell'}$ by $F_{\ell'}B_{\ell'}^2$, and the noise adds its own average power. This is Hivon's eq. (15), and the form of $M_{\ell\ell'}$ is derived for the masked sky in Section 8.10.1. With no mask ($M = 1$ everywhere) and no filter,

$$\langle\tilde{C}_\ell\rangle \overset{(1)}{=} \sum_{\ell'} M_{\ell\ell'}F_{\ell'}B_{\ell'}^2\langle C_{\ell'}\rangle + \langle\tilde{N}_\ell\rangle \overset{(2)}{=} T_\ell^2\, C_\ell + \bar{N}_\ell,$$

(1) Hivon's eq. (15).
(2) With $M = 1$ the masked sky is the sky, no multipoles are mixed and $M_{\ell\ell'} = \delta_{\ell\ell'}$; no filter means $F_\ell = 1$; $B_\ell = T_\ell$ and $\langle\tilde{N}_\ell\rangle = \bar{N}_\ell$ of (8.17).

which is our (8.19): what remains is the beam, the pixel window and the noise. Every statement of the paper about those three can therefore be tested on our full-sky simulations; the statements about the mask are taken up with the masked sky.

Noise-bias removal. Hivon et al. found the recovered spectrum unbiased both where the signal dominates and where the signal-to-noise is very low [Hivon §4.2, Fig. 5]. Our debiased estimator passes the corresponding test with $\chi^2 = 20.5$ for 21 bins (Section 8.6.1).

The analytic error bar. Their rule of thumb [Hivon (16)–(17)] gives the error of a bin of $\Delta\ell$ multipoles as $(C_b + N_b/T_b^2)\sqrt{2/\nu_b}$ with $\nu_b = (2\ell_b+1)\Delta\ell\, f_{\rm sky} w_2^2/w_4$, and they found their simulated errors "almost identical" to it. The rule follows from (8.29) in two steps. The binned estimate is the plain average $\widehat{C}_b = \Delta\ell^{-1}\sum_{\ell\in b}\widehat{C}_\ell$ over the $\Delta\ell$ multipoles of bin b, centred on ℓ_b. On the full sky,

$$\begin{aligned}\mathrm{Var}\big(\widehat{C}_b\big) &\overset{(1)}{=} \frac{1}{\Delta\ell^2}\sum_{\ell\in b}\mathrm{Var}\big(\widehat{C}_\ell\big) \overset{(2)}{=} \frac{1}{\Delta\ell^2}\sum_{\ell\in b}\frac{2}{2\ell+1}\Big(C_\ell + \frac{N_\ell}{T_\ell^2}\Big)^2 \\ &\overset{(3)}{\approx} \frac{2}{(2\ell_b+1)\,\Delta\ell}\Big(C_b + \frac{N_b}{T_b^2}\Big)^2. \end{aligned} \tag{8.41}$$

(1) Different multipoles of a full sky are independent, so the variances of the $\Delta\ell$ terms add, and the factor $1/\Delta\ell$ comes out squared.
(2) The single-multipole variance (8.29).
(3) If C_ℓ, N_ℓ/T_ℓ^2 and $2\ell+1$ change little across the bin, the $\Delta\ell$ terms are all close to the one at ℓ_b, and their sum is $\Delta\ell$ times it.

The square root of (8.41) is $(C_b + N_b/T_b^2)\sqrt{2/\nu_b}$ with $\nu_b = (2\ell_b+1)\Delta\ell$, the number of modes in the bin. On part of the sky the mode count of Section 8.4.2 multiplies ν_b by $f_{\rm sky}$, and by w_2^2/w_4 for a weighted mask, which gives their eqs. (16)–(17). The quantity compared is therefore our simulated standard deviation of $\widehat{C}_b$ divided by the square root of (8.41) with $f_{\rm sky} = 1$, and, separately, divided by the square root of line 2, which makes no assumption about the bin. On the full sky, in bins of $\Delta\ell = 50$, our simulated error divided by the rule lies between 0.989 and 1.098 in the bins above $\ell = 63$, with a Monte Carlo error of 1.6%. The first bin, $\ell = 13$–62, is the

exception, with a ratio of 1.85. The rule evaluates the spectrum at the bin average, but C_ℓ falls by a factor of about twenty across that bin, and the variance of a bin average, $\sum_\ell \mathrm{Var}(\widehat{C}_\ell)/\Delta\ell^2$, is dominated by the largest C_ℓ in it. The exact expression, line 2 of (8.41), gives ratios between 0.987 and 1.040 in every bin. The rule fails exactly at its step 3: it needs a spectrum that is nearly constant across the bin.

How fast the calibrations converge. MASTER estimates three things by simulation: the transfer function, the noise power, and the error bars. Each estimate comes from a finite number N of simulations, so each carries a Monte Carlo error of its own. The question is: *how large is that error, and how does it shrink with N?* Hivon et al. give the answers as their eqs. (38)–(42) [Hivon (38)–(42)]. We derive them from the mode count of (8.41).

The transfer function and the noise power. Each is an average over the N simulations of one binned spectrum, call it X_b: the noise-free spectrum divided by the input spectrum for the transfer function, the noise-only spectrum for the noise power. In one simulation, X_b is built from $\nu_b = (2\ell_b + 1)\Delta\ell$ independent modes, so by (8.41) it scatters by $\sigma(X_b) = \sqrt{2/\nu_b}\,\langle X_b\rangle$. Averaging N independent simulations reduces that scatter by $\sqrt{N}$.

The error bar. It is the standard deviation s of $\widehat{C}_b$ over the N simulations. With many modes in a bin, $\widehat{C}_b$ is close to Gaussian, and then $(N-1)s^2/\sigma^2$ is a χ^2_{N-1} variable (Chapter 3), whose variance we know.

Then

$$\frac{\delta \bar{X}_b}{\langle X_b\rangle} \overset{(1)}{=} \frac{\sigma(X_b)/\sqrt{N}}{\langle X_b\rangle} \overset{(2)}{=} \sqrt{\frac{2}{(2\ell_b+1)\,\Delta\ell\,N}}, \tag{8.42}$$

$$\frac{\delta s}{s} \overset{(3)}{=} \frac{1}{2}\,\frac{\delta(s^2)}{\sigma^2} \overset{(4)}{=} \frac{1}{2}\sqrt{\frac{2}{N-1}} = \frac{1}{\sqrt{2(N-1)}}. \tag{8.43}$$

(1) The average of N independent copies has $1/N$ times the variance of one.

(2) $\sigma(X_b) = \sqrt{2/\nu_b}\,\langle X_b\rangle$ with $\nu_b = (2\ell_b+1)\Delta\ell$; on part of the sky ν_b carries the factor f_{sky}, which gives [Hivon (19), (38)–(39)].

(3) Error propagation for $s = \sqrt{s^2}$: $\delta s = \delta(s^2)/(2s)$, and $s \approx \sigma$.

(4) $s^2 = \sigma^2\chi^2_{N-1}/(N-1)$ and $\mathrm{Var}(\chi^2_{N-1}) = 2(N-1)$, so $\delta(s^2) = \sigma^2\sqrt{2/(N-1)}$.

Third, the error that an N-simulation noise estimate puts on $\widehat{C}_\ell$, relative to its statistical error, is $N_\ell/(T_\ell^2 C_\ell + N_\ell)\cdot N^{-1/2}$, derived in (8.25). What is compared is the scatter of each Monte Carlo estimate between disjoint groups of N of our skies, divided by (8.42), (8.43) or (8.25). Script `12_hivon_convergence.py` measures all four by splitting our 2000 skies into disjoint groups of $N = 25$ to 400 (Figure 8.20). At $N = 100$ the measured scatter divided by our formula is 1.00 for the transfer function, 0.99 for the noise power, 1.01 for the error bar and 0.98 for the noise-induced error, each ±0.01.

The published coefficients are larger, and most of the difference can be traced. Take the transfer function and the noise power at $\ell \approx 400$, $\Delta\ell = 50$ and $N = 100$. Our full-sky formula (8.42) gives 0.072%; Hivon's numbers are 0.6–0.9%. Two causes act through the mode count ν_b, and the error goes as $1/\sqrt{\nu_b}$.

- *The sky fraction.* Their region is 1.82% of the sky, so ν_b carries the factor $f_{\mathrm{sky}} = 0.0182$ and the error rises by $1/\sqrt{0.0182} = 7.4$: from 0.072% to 0.53%. After this step their numbers are still $0.6/0.53 = 1.13$ to $0.9/0.53 = 1.70$ times ours.

- *The weighting.* They weighted their region with a cosine-apodised window, for which

$w_2^2/w_4 = 0.5$. In (8.34) this factor multiplies f_{sky}, so it halves ν_b and raises the error by $\sqrt{2} = 1.41$: from 0.53% to 0.75%.

Their range now sits on either side of ours: $0.6/0.75 = 0.8$ and $0.9/0.75 = 1.2$. What is left, a factor between 0.8 and 1.2, depends on details of their region that our full-sky model does not contain, such as the uneven number of observations across the patch and the correlated noise of their detector. We do not reproduce it, and with a full sky we cannot. For the error bar, their "$0.1\sqrt{100/N}$" is a rounded upper value of $1/\sqrt{2(N-1)} = 7.1\%$ at $N = 100$, which our groups reproduce. For the noise-induced error, their $0.1\sqrt{100/N}$ is exactly $1/\sqrt{N}$, our (8.25) in the deeply noise-dominated limit; in our last bin the noise is 0.68 of the power, so the formula gives 6.8% and the groups give 6.6%.

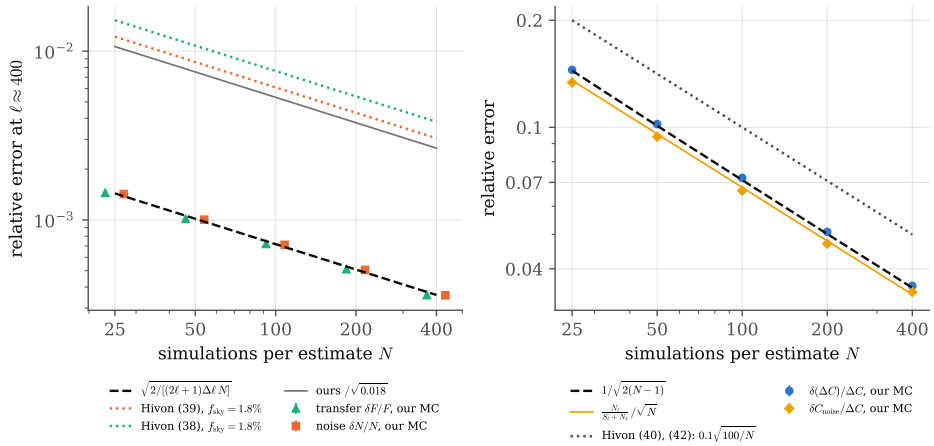


Figure 8.20: Hivon et al.'s convergence laws against our simulations. Left: the relative Monte Carlo error of the transfer function (green) and of the noise power (orange) at $\ell \approx 400$ in bins of 50, from disjoint groups of N skies, on the full-sky law (dashed). Hivon's published laws for 1.8% of the sky (dotted) lie above it by about $1/\sqrt{f_{\mathrm{sky}}}$ (grey). Right: the relative error of an estimated error bar (blue) and the error a noise estimate adds to $\widehat{C}_\ell$ (yellow), against $1/\sqrt{2(N-1)}$ (dashed), its noise-weighted version (yellow line) and Hivon's $0.1\sqrt{100/N}$ (dotted).

The shape of the distribution. Hivon et al. showed that at small numbers of modes a χ^2_ν with no free parameter describes the simulated estimates far better than a Gaussian of the same mean and width [Hivon §4.2, Fig. 6]. On the full sky the number of modes is small only at the lowest multipoles. At $\ell = 3$, $\nu = 7$, the shifted χ^2 passes the Kolmogorov–Smirnov test with $p = 0.09$ and the Gaussian fails with $p = 1 \times 10^{-5}$ (Section 8.6.3). Their small f_{sky} brought the same effect up to $\ell \sim 100$; that version needs the mask.

What reproducing it taught us. On the full sky every claim MASTER makes about the beam, the pixel window and the noise is checked by our machine, and the one failure we found, in the first bin, shows that the binned rule of thumb needs a spectrum that is flat across the bin. The claims that depend on the mask are not reproduced here.

8.7.3 Planck's pivot: why A_s and n_s are quoted at $k_* = 0.05\,\mathrm{Mpc}^{-1}$

Planck reports two numbers for the primordial fluctuations: their amplitude, $\ln(10^{10}A_s) = 3.044 \pm 0.014$, and their tilt, $n_s = 0.9649 \pm 0.0042$ (Chapter T1). The amplitude is the one "which we define at the pivot scale $k_0 = 0.05\,\mathrm{Mpc}^{-1}$" [25]. WMAP quoted its amplitude at a different

quantity	published	ours	why they differ
Knox (1995)			
$w^{-1/2}$ for 22 μK in 20′ pixels	7.5 μK deg	7.33 μK deg	his rounding
sky rms through a 20′ beam	92.5 μK	95 μK	sCDM vs Planck 2018
signal-to-noise per pixel	$\simeq 4$	4.3	same
error formula, eq. (4)	exact	maps: 1.000	with p_ℓ divided out
as printed, 20′ beam	—	median 1.053	no pixel window in eq. (4)
smaller beam wins at fixed w	every ℓ	every $\ell \le 1000$	—
N for σ to 10%	“100 suffice”	51 (7.1% at 100)	—
Hivon et al. (2002), full-sky limit			
bias after noise removal	none (Fig. 5)	$\chi^2 = 20.5/21$	—
MC error / rule (16)	“almost identical”	0.989–1.098	first bin 1.85: C_ℓ not flat
$\delta F/F$, $\delta N/N$ at $N = 100$	0.6–0.9%	0.072%	$f_{\rm sky} = 1.8\%$; ours $/\sqrt{f_{\rm sky}}$: 0.53%
$\delta(\Delta C)/\Delta C$ at $N = 100$	$\approx 10\%$	7.2%	their value is rounded up
noise error / stat. error	10%	6.6%	our bin only 0.68 noise
χ^2 vs Gaussian, few modes	χ^2 wins	KS p: 0.09 vs 1×10^{-5}	full sky: only at low ℓ
simulations	$450 + 250 + 1350$	2000	ours are cheaper, simpler

Table 8.3: Published numbers and the same quantities from our code. “Maps” is the simulated scatter divided by the formula, averaged over ℓ.

scale, $k = 0.002\,\mathrm{Mpc}^{-1}$, and Planck still quotes the tensor-to-scalar ratio there, $r_{0.002} < 0.10$ [25]. Why does an amplitude need a scale attached to it at all, and why this one?

The spectrum is a straight line. The primordial spectrum is a power law,

$$\mathcal{P}(k) = A_s \left(\frac{k}{k_0}\right)^{n_s - 1}, \tag{8.44}$$

so $\ln \mathcal{P} = \ln A_s + (n_s - 1)\ln(k/k_0)$. Plotted as $\ln \mathcal{P}$ against $\ln k$ it is a straight line. Its slope is $n_s - 1$. Its height at the point $k = k_0$ is $\ln A_s$. That point, the **pivot**, is our choice: we may describe the same line by its height at any k we like. The slope stays the same whichever point we choose. The height changes, and so does something less obvious: how strongly the measured height and the measured slope are correlated.

Think of fitting a straight line through points. If we report the height of the line far to the left of the data, any error in the slope swings that height up or down, so height and slope come out strongly correlated. If we report the height in the middle of the data, a small change of slope just pivots the line about that point, and the height hardly moves: the two errors are independent. The pivot scale is the same idea. *At which k do the measured amplitude and the measured tilt become uncorrelated, and where does that k sit?*

Step 1: what moving the pivot does. Call $A(k_0)$ the amplitude quoted at k_0, and write n for n_s. The same spectrum, evaluated at another point k_1, has height $A(k_1) = A(k_0)\,(k_1/k_0)^{n-1}$; this is just (8.44) at $k = k_1$. Write $L = \ln(k_1/k_0)$ for the distance between the two points in $\ln k$. It is a fixed number, not a measured one. Then

$$\begin{aligned} \ln A(k_1) &\overset{(1)}{=} \ln A(k_0) + (n-1)\,L, \\ \operatorname{Cov}\big(\ln A(k_1), n\big) &\overset{(2)}{=} \operatorname{Cov}\big(\ln A(k_0), n\big) + L\,\operatorname{Var}(n), \\ \operatorname{Var}\big(\ln A(k_1)\big) &\overset{(3)}{=} \operatorname{Var}\big(\ln A(k_0)\big) + 2L\,\operatorname{Cov}\big(\ln A(k_0), n\big) + L^2\,\operatorname{Var}(n). \end{aligned} \tag{8.45}$$

(1) Take the logarithm of $A(k_1) = A(k_0)(k_1/k_0)^{n-1}$.
(2) Take the covariance of both sides of line 1 with n. A covariance splits over a sum. The constant $-L$ hidden in $(n-1)L$ does not fluctuate, so it drops out, and the covariance of n with itself is its variance.
(3) The variance of a sum $X + LY$ is $\operatorname{Var} X + 2L\operatorname{Cov}(X,Y) + L^2 \operatorname{Var} Y$; here $X = \ln A(k_0)$ and $Y = n$.

Line 2 says that moving the pivot by L shifts the covariance by $L\ \operatorname{Var}(n)$. So there is exactly one pivot at which the covariance is zero. Setting line 2 to zero and solving for L gives the **decorrelation pivot**:

$$L_{\mathrm{dec}} = -\frac{\operatorname{Cov}\big(\ln A(k_0), n\big)}{\operatorname{Var}(n)}, \qquad k_{\mathrm{dec}} = k_0 \exp\!\left[-\frac{\operatorname{Cov}\big(\ln A(k_0), n\big)}{\operatorname{Var}(n)}\right]. \tag{8.46}$$

Line 3 gives a bonus. Its derivative with respect to L is $2\operatorname{Cov} + 2L\operatorname{Var}(n)$, which vanishes at the same L_{dec}. So the pivot that makes the two numbers independent is also the pivot at which the amplitude has its smallest error. All (8.46) needs is the covariance of the two numbers at one pivot, however it was obtained: from a Fisher matrix, a Markov chain or a fit.

Step 2: from wavenumber to multipole. The data are multipoles, the pivot is a wavenumber, so first we need to know which multipoles see which wavenumbers. The CMB is a picture of the last-scattering sphere, at comoving distance D_* from us. A wave of wavenumber k on that sphere has wavelength $2\pi/k$. Seen from a distance D_*, that length spans an angle of $2\pi/(kD_*)$ radians. A spherical harmonic of degree ℓ has an angular wavelength of $2\pi/\ell$. Matching the two, $\ell \approx kD_*$. CAMB gives $D_* = 13.87\,\mathrm{Gpc}$ for the fiducial model (see the cosmology notes). So the Planck pivot $0.05\,\mathrm{Mpc}^{-1}$ sits at $\ell \approx 694$, and the WMAP pivot $0.002\,\mathrm{Mpc}^{-1}$ at $\ell \approx 28$. The match is approximate: a wave whose crests are not perpendicular to our line of sight looks longer on the sky, so one k also feeds every ℓ below kD_*. Turned around: multipole ℓ receives power from a range of wavenumbers, through a kernel $G_\ell(k)$ that peaks near $k_\ell \equiv \ell/D_*$ and has a tail towards larger k. The temperature spectrum is linear in the primordial one,

$$C_\ell = \int G_\ell(k)\,\mathcal{P}(k)\,\mathrm{d}\ln k, \tag{8.47}$$

with $G_\ell \geq 0$ fixed by the physics between the early universe and last scattering, which CAMB computes.

Step 3: how much each multipole knows. To use (8.46) we need that covariance for a temperature measurement, and the Fisher matrix of Chapter 10 gives it. We already have the Fisher information of one C_ℓ, (8.31). We need it for the parameters θ_i that the spectrum depends on. Two facts make this a short step. On the full sky different multipoles are independent, so

their informations add. And the parameters reach the data only through C_ℓ, so the chain rule converts information on C_ℓ into information on θ_i:

$$F_{ij} \equiv \left\langle \frac{\partial^2(-\ln \mathcal{L})}{\partial\theta_i \partial\theta_j} \right\rangle \overset{(1)}{=} \sum_\ell \left\langle \frac{\partial^2(-\ln \mathcal{L}_\ell)}{\partial C_\ell^2} \right\rangle \frac{\partial C_\ell}{\partial\theta_i}\frac{\partial C_\ell}{\partial\theta_j} \overset{(2)}{=} \sum_\ell \frac{(2\ell+1) f_{\rm sky}}{2\,(C_\ell + N_\ell/T_\ell^2)^2} \frac{\partial C_\ell}{\partial\theta_i}\frac{\partial C_\ell}{\partial\theta_j}. \tag{8.48}$$

(1) Write $\ln\mathcal{L} = \sum_\ell \ln\mathcal{L}_\ell$ and differentiate twice with the chain rule. One term contains $\partial^2 C_\ell/\partial\theta_i\partial\theta_j$, but it is multiplied by the average slope of $-\ln\mathcal{L}_\ell$, which is zero at the true parameters (Chapter 3).
(2) The information on one C_ℓ is (8.31). On part of the sky there are only $(2\ell+1)f_{\rm sky}$ modes instead of $2\ell+1$ (Section 8.4.2).

Now the two parameters. How does C_ℓ change when each of them is nudged? Differentiate (8.47) under the integral; the kernel does not depend on A or n. The amplitude multiplies $\mathcal{P}(k)$ at every k, so $\partial\mathcal{P}/\partial\ln A = \mathcal{P}$. The tilt enters through $\ln\mathcal{P} = \ln A + (n-1)\ln(k/k_0)$, so $\partial\mathcal{P}/\partial n = \mathcal{P}\ln(k/k_0)$: raising n raises the spectrum more at large k and lowers it at $k < k_0$. Then

$$\frac{\partial C_\ell}{\partial \ln A} \overset{(1)}{=} \int G_\ell\,\mathcal{P}\,\mathrm{d}\ln k = C_\ell,$$
$$\frac{\partial C_\ell}{\partial n} \overset{(2)}{=} \int G_\ell\,\mathcal{P}\,\ln\frac{k}{k_0}\,\mathrm{d}\ln k \overset{(3)}{=} C_\ell \left\langle \ln\frac{k}{k_0}\right\rangle_\ell \overset{(4)}{\approx} C_\ell\, g_\ell, \qquad g_\ell \equiv \ln\frac{k_\ell}{k_0}.$$

(1) $\partial\mathcal{P}/\partial\ln A = \mathcal{P}$, and the integral is (8.47).
(2) $\partial\mathcal{P}/\partial n = \mathcal{P}\ln(k/k_0)$.
(3) Divide and multiply by C_ℓ: $\langle\cdot\rangle_\ell$ is the average over $\ln k$ with weight $G_\ell\mathcal{P}/C_\ell$, which integrates to one.
(4) Replace the kernel by a spike at its peak $k_\ell = \ell/D_*$ (Step 2). The tail of the kernel towards larger k makes this a shortcut, not an exact step; the script below keeps the full kernel.

So g_ℓ is the distance of multipole ℓ from the pivot in $\ln k$. Put the two derivatives into (8.48). Every term carries the same factor, which we name the **information weight** of multipole ℓ,

$$\mathcal{I}_\ell \equiv \frac{(2\ell+1)f_{\rm sky}}{2}\left[\frac{C_\ell}{C_\ell + N_\ell/T_\ell^2}\right]^2,$$

the information that multipole ℓ carries about $\ln A$: its number of modes, cut down where the noise on the sky N_ℓ/T_ℓ^2 exceeds the signal. Since $\frac{(2\ell+1)f_{\rm sky}}{2(C_\ell+N_\ell/T_\ell^2)^2} = \mathcal{I}_\ell/C_\ell^2$, the three entries are

$$F_{AA} = \sum_\ell \frac{\mathcal{I}_\ell}{C_\ell^2} C_\ell^2 = \sum_\ell \mathcal{I}_\ell, \qquad F_{An} = \sum_\ell \frac{\mathcal{I}_\ell}{C_\ell^2} C_\ell \cdot C_\ell g_\ell = \sum_\ell \mathcal{I}_\ell\, g_\ell, \qquad F_{nn} = \sum_\ell \mathcal{I}_\ell\, g_\ell^2,$$

so that

$$F = \begin{pmatrix} \sum_\ell \mathcal{I}_\ell & \sum_\ell \mathcal{I}_\ell g_\ell \\ \sum_\ell \mathcal{I}_\ell g_\ell & \sum_\ell \mathcal{I}_\ell g_\ell^2 \end{pmatrix}.$$

Its inverse is the covariance. For a 2×2 matrix $\left(\begin{smallmatrix} a & b \\ b & c \end{smallmatrix}\right)$ the inverse is $\left(\begin{smallmatrix} c & -b \\ -b & a \end{smallmatrix}\right)/\det F$, so the covariance of the two numbers is $-b/\det F$ and the variance of n is $a/\det F$. Put them into (8.46):

$$\ln\frac{k_{\rm dec}}{k_0} = -\frac{\mathrm{Cov}(\ln A, n)}{\mathrm{Var}(n)} \overset{(1)}{=} \frac{\sum_\ell \mathcal{I}_\ell g_\ell}{\sum_\ell \mathcal{I}_\ell} \overset{(2)}{=} \frac{\sum_\ell \mathcal{I}_\ell\,\ln(k_\ell/k_0)}{\sum_\ell \mathcal{I}_\ell},$$

(1) $-(-b/\det F)/(a/\det F) = b/a$; the determinant cancels.
(2) Write out g_ℓ.

Adding $\ln k_0$ to both sides, $\ln k_{\mathrm{dec}} = \sum_\ell \mathcal{I}_\ell \ln k_\ell / \sum_\ell \mathcal{I}_\ell$. In words: *the decorrelation pivot is the average of* $\ln k$ *over the scales the data measure, each scale counted with the weight of what it knows.* That is the straight-line picture made exact: the height and the slope are independent when the height is quoted in the middle of the data, and "middle" means the weighted middle. The sky fraction multiplies every $\mathcal{I}_\ell$, so it cancels: observing less sky widens the errors but does not move k_{dec}.

Step 4: the numbers. Script `14_pivot.py` does the calculation without the spike shortcut, for the Planck-like experiment of Table 8.4. It computes the temperature spectrum with CAMB; finds the derivatives of C_ℓ with respect to $\ln A$ and n by nudging each up and down, at both pivots; builds the Fisher matrix (8.48); inverts it to get the covariance; and puts that covariance into (8.46) to get k_{dec}.

beam FWHM	$7'$
white-noise depth	$33\,\mu$K arcmin
sky fraction	$f_{\mathrm{sky}} = 0.6$
multipoles	$2 \le \ell \le 2500$, temperature only
free parameters	$\ln A$ and n; then all six ΛCDM parameters, with a Gaussian prior of width 0.0073 on τ standing in for the large-angle polarisation that fixes it

Table 8.4: The idealised Planck-like experiment of the pivot calculation.

quantity	at $0.05\,\mathrm{Mpc}^{-1}$	at $0.002\,\mathrm{Mpc}^{-1}$	reason
correlation of $\ln A$ and n	-0.74	-0.99	0.002 lies far from the weighted middle of the data
moved from 0.05 by (8.45)	—	-0.99	the shift is exact
k_{dec} [Mpc^{-1}]	0.083	0.083	a property of the data, not of the pivot
$\sigma(\ln A)$	0.0017	0.0094	the $L^2\,\mathrm{Var}(n)$ term of line 3
correlation, six parameters free	-0.21	-0.81	a common amplitude error from τ

Table 8.5: The pivot calculation at the two pivots.

Table 8.5 collects the results, and Figure 8.21 shows the correlation for every pivot together with the weights $\mathcal{I}_\ell$. Each difference has a cause that the steps above can size.

The correlation. With only $\ln A$ and n free, $k_{\mathrm{dec}} = 0.083\,\mathrm{Mpc}^{-1}$, that is $\ell_{\mathrm{dec}} \approx 1155$. The Planck pivot is at $\ell \approx 694$, not far below it, and the correlation there is moderate. The WMAP pivot is at $\ell \approx 28$, far to the left of almost all the information, and there the two numbers are almost perfectly anti-correlated: the straight-line picture of the height read off far from the data.

The check of Step 1. Computing the Fisher matrix directly at $0.002\,\mathrm{Mpc}^{-1}$ and moving the covariance there from $0.05\,\mathrm{Mpc}^{-1}$ with (8.45) give the same correlation, and both pivots give the same k_{dec}: the decorrelation scale belongs to the data. The spike shortcut of Step 3, the $\mathcal{I}_\ell$-weighted average of $\ln \ell$, gives $\ell \approx 972$. It falls short of the exact 1155 because the kernel's tail towards larger k makes each multipole see slightly larger k than ℓ/D_*.

The amplitude error. Moving from 0.05 to $0.002\,\mathrm{Mpc}^{-1}$ is a step $L = \ln(0.002/0.05) = -3.22$. Line 3 of (8.45) then gives $\mathrm{Var}(\ln A)$ at 0.002 as $0.0017^2 + 2L\,\mathrm{Cov} + L^2\,\mathrm{Var}(n)$. With $\sigma(n) = 0.0025$

and the correlation -0.74, the three terms are 0.29, 2.0 and 6.5 in units of 10^{-5}; the square root of their sum is 0.0094, the 0.0094 of the table. The $L^2\,\mathrm{Var}(n)$ term carries three quarters of it: quoted far from the data, the height inherits the uncertainty of the slope, magnified by the distance.

Marginalising. With all six ΛCDM parameters free, the covariance of $\ln A$ and n is read from the inverse of the full six-by-six Fisher matrix, so each error now includes what the other parameters can absorb. On all but the largest scales the temperature spectrum fixes only the combination $A\,e^{-2\tau}$, so the prior of width 0.0073 on τ leaves $\ln A$ an error of about $2 \times 0.0073 = 0.015$ at every pivot. That common error is much larger than the slope term $L\,\sigma(n)$ for any pivot near the data, so the correlation, which is the slope term over the total, shrinks: -0.21 at $0.05\,\mathrm{Mpc}^{-1}$. At $0.002\,\mathrm{Mpc}^{-1}$, L is large enough that the slope term still wins, and the correlation stays at -0.81. The tilt, for its part, trades with the matter densities that shape the acoustic peaks, and those trades do not weigh all multipoles equally; the weighted middle in (8.46) moves to $k_{\mathrm{dec}} = 0.091\,\mathrm{Mpc}^{-1}$. The size of that shift comes from the full matrix, not from a formula we can write by hand.

What this says about the choices. For this idealised experiment the exact decorrelation scale is near $\ell \approx 1155$, where the weight $\mathcal{I}_\ell$ peaks before the beam and the noise cut it off. The Planck pivot, $0.05\,\mathrm{Mpc}^{-1}$ at $\ell \approx 694$, is not exactly there. It is a round number inside the band of scales where the temperature information is concentrated, and there the correlation is modest. The Planck paper fixes the pivot by definition and calls the tilt measured "around the pivot scale of $k \simeq 0.05\,\mathrm{Mpc}^{-1}$" robust [25]; it does not quote a decorrelation scale of its own, which for the real likelihood also depends on the polarisation spectra and on the foreground model.

WMAP's choice answered a different need. Its temperature spectrum was limited by cosmic variance only over the first few hundred multipoles. For a noiseless spectrum that stops at $\ell = 500$, (8.46) gives $k_{\mathrm{dec}} = 0.030\,\mathrm{Mpc}^{-1}$ ($\ell \approx 416$). WMAP's pivot, $0.002\,\mathrm{Mpc}^{-1}$ at $\ell \approx 28$, is far to the left of that. It sits at the large-scale end, where tensor modes contribute (they fade above $\ell \sim 100$) and where COBE had fixed the normalisation. There the amplitude and the tilt are almost perfectly anti-correlated, -0.98 for the same noiseless spectrum. Neither pivot is wrong. A pivot is a choice of where to read off the height of a line, and (8.45) converts one choice into another. A pivot near k_{dec} simply gives two numbers that can be read separately.

8.8 How it is done in research

The toy pipeline of a beamed, noisy full sky has every structural element of a real temperature analysis and almost none of its detail. It is worth seeing the real thing once, in the order in which it runs, because every later chapter that uses $\widehat{C}_\ell$ inherits its choices. The description follows the Planck 2018 temperature analysis [75] and the MASTER method [Hivon §3.6], with modern implementations such as NaMaster [169].

The data products. An analysis starts from the time-ordered data of every detector, with its pointing, and from three calibrations made on the same data: the beam of each detector, measured by scanning planets; the noise spectrum of each detector, measured from the timestream itself; and the gain. Map-making turns the timestream into maps at $N_{\mathrm{side}} = 2048$ for each high-frequency channel and for each *split* of the data (the two half-missions, two sets of detectors), together with hit counts and pixel noise variances. Masks remove the Galaxy and bright point sources and are apodised. The *effective* beam window function of each map, which combines the beams of the detectors that went into it with the way the scan crossed each pixel, is computed from the beam measurements and the scan. Finally a large set of simulations is

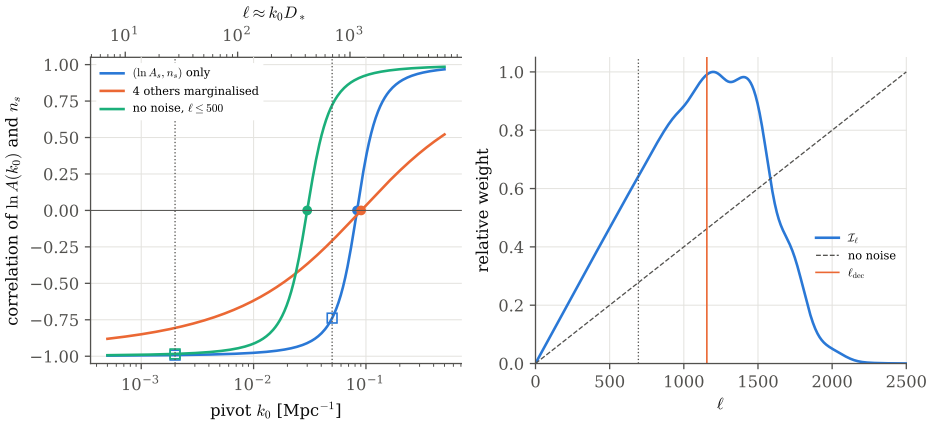


Figure 8.21: The pivot scale. Left: the correlation of $\ln A(k_0)$ and n_s as a function of the pivot k_0, for the Planck-like experiment with $(\ln A_s, n_s)$ alone (blue) and with the other four parameters marginalised (orange), and for a noiseless spectrum cut at $\ell = 500$ (green). The curves follow from one Fisher matrix at $0.05\,\mathrm{Mpc}^{-1}$ moved along by (8.45); the open squares are Fisher matrices computed directly at each pivot, and the dots mark $k_{\rm dec}$, where each curve crosses zero. The top axis converts k_0 to $\ell \approx k_0 D_*$. Right: the weight $\mathcal{I}_\ell$, the information on $\ln A_s$ per multipole, rises like $2\ell + 1$ until the beam and the noise cut it off; its weighted centre in $\ln \ell$ is $\ell_{\rm dec}$ (orange). The dotted lines mark 0.002 and $0.05\,\mathrm{Mpc}^{-1}$ (left) and $0.05\,\mathrm{Mpc}^{-1}$ (right).

produced end to end: CMB skies, foregrounds and noise timestreams, run through the same map-maker. In Planck 2018 these are the 300 FFP10 realisations [75].

The steps.

1. *Maps.* Generalised least squares, or a destriper that removes the $1/f$ drifts, solves $\boldsymbol{d} = \mathsf{A}\boldsymbol{s} + \boldsymbol{n}$ for the map [HobsonBMC (10.10)–(10.13)].

2. *Pseudo-spectra.* Each masked, apodised map is transformed; spectra are formed between *different* splits, so that the noise bias never enters (Section 8.3.3).

3. *Deconvolution.* The mode coupling of the mask is computed exactly from the mask's own spectrum and inverted in bins of ℓ; the effective beam and pixel window are divided out [Hivon (14)–(26)]. A transfer function measured on noise-free simulations corrects for any filtering [Hivon (18)].

4. *Covariance.* The covariance of the binned spectra, which a mask makes non-diagonal, is computed with analytic approximations for masked, noisy pseudo-spectra and checked on simulations, or estimated directly from simulations with the correction for a finite number of them [6].

5. *Likelihood.* Above $\ell = 30$ the binned cross-spectra of the frequency maps enter a Gaussian likelihood together with foreground and instrument nuisance parameters, for $30 \leq \ell \leq 2508$ [75]. Below $\ell = 30$, where the distribution is visibly a skewed χ^2 (Section 8.4.1), the likelihood is computed from the low-resolution map itself, with Gibbs sampling of the sky and the spectrum together [75].

The checks. None of this is trusted until it has passed tests designed to fail if something is wrong. *Null tests* difference two splits of the same sky: the signal cancels, and the spectrum of the difference must be consistent with the noise model alone. *Consistency tests* compare spectra from different frequencies, detector sets and masks, which must agree after foregrounds are

accounted for. *End-to-end tests* run the entire pipeline on the simulations as if they were data and check that the input spectrum and parameters come back without bias and with error bars that cover the truth as often as they claim [75]. The full-sky checks of Section 8.6 are the smallest version of the last kind.

Taxonomy: Research pipeline and its laptop version

step	research version	here	cost of the reduction
sky model	Gaussian CMB + foregrounds, $N_{\rm side} = 2048$	Gaussian CMB, $N_{\rm side} = 256$	$\ell \le 512$: 4.2% of the modes
beam	measured, asymmetric, per detector, with errors	circular Gaussian, 30′	no beam-error term
noise	correlated $1/f$ timestream through map-maker	white pixel noise (and one pattern)	no stripes or noise correlations
spectra	half-mission cross-spectra, masked	full sky, auto and cross	mask in the masked-sky analysis
simulations	300 end-to-end FFP10	2000 Gaussian skies, 6 min	MC error 1.6% vs 4.1%, but less realism
full-scale run	computing cluster	extrapolated: 5 h for 1000 Gaussian skies at 2048	breaks the 10-minute budget
storage	maps on disk	spectra only, 29 MB	1000 maps at 2048: 403 GB

The reductions in the table are deliberate. Each removes something expensive whose effect on the statistics of $\widehat{C}_\ell$ is either known exactly on the full sky or is not the subject here. What cannot be removed is the structure: random sky, deterministic instrument, random noise, the same analysis on data and simulations, and statistics read off the simulations. That structure is the same at $N_{\rm side} = 256$ and at 2048.

How this is used in real research: Verification and research with the same machine

On the full sky the simulations *verify*: every number they produced had an analytic prediction, and agreement tested the code. In research the same loop is the *instrument*, used where no prediction exists: the transfer function of a filtered scan [Hivon (18)], the noise power of coloured, inhomogeneous, filtered noise [Hivon §3.3], and the covariance and distribution of the binned spectrum of a masked sky, which the likelihood of Chapter 9 and the Fisher forecast of Chapter 10 need.

The beam and the noise also change the map in ways the spectrum does not show. The beam removes the smallest hot and cold spots, and the noise adds a grain of tiny new ones. Counts of spots, their connectivity and the other shape statistics of Chapter 12 respond to both, and the same simulations, which already produce beamed, noisy maps, will provide their sampling distributions.

Recap

- An instrument multiplies each $a_{\ell m}$ by $T_\ell = B_\ell p_\ell$ (beam times pixel window, $B_\ell = e^{-\ell(\ell+1)\sigma_b^2/2}$ for a Gaussian beam) and adds noise after the smoothing; uniform white noise has $N_\ell = \sigma^2 \Omega_{\rm pix}$ at every ℓ, and any noise pattern has mean power $\Omega_{\rm pix}\langle\sigma^2\rangle$.
- The observed spectrum averages to $T_\ell^2 C_\ell + N_\ell$. Subtracting N_ℓ and dividing by

T_ℓ^2 gives an unbiased estimator, which on the full sky is the maximum-likelihood estimator and reaches the Cramér–Rao bound. Cross-spectra of independent halves need no noise subtraction and cost at most $\sqrt{2}$ in error.

- Its variance is $2(C_\ell + N_\ell/T_\ell^2)^2/(2\ell+1)$ (Knox), its distribution a shifted, scaled $\chi^2_{2\ell+1}$, skewed at low ℓ and straddling zero where the noise dominates. The noise supplies half the variance once $N_\ell/(T_\ell^2 C_\ell) > \sqrt{2}-1$.
- The f_{sky} version of the formula counts modes; it holds for wide bins with many modes and fails at low ℓ and small patches.
- The cumulative signal-to-noise $[\sum_\ell \mathrm{SNR}_\ell^2]^{1/2}$ is one over the error of an overall amplitude; it saturates a little beyond the multipole where noise and signal are equal.
- 2000 simulated skies at $N_{\mathrm{side}} = 256$, six minutes on a laptop, reproduce all of these to their Monte Carlo precision, and reproduce the full-sky content of Knox (1995) and Hivon et al. (2002) (Table 8.3).
- An error bar needs about 50 simulations for 10% precision; a bias needs $(\sigma/\mathrm{target})^2$.

8.9 A sky with holes in it

The full-sky experiment of Section 8.2 could see every direction. No real temperature map can be used that way. The disc of the Milky Way shines in microwaves through synchrotron radiation, free–free emission and warm dust, and a band of the sky around the Galactic plane is brighter in foregrounds than in CMB; a few thousand radio galaxies and dusty galaxies add bright points everywhere else (see the cosmology notes, and Chapter T1 for the map of these objects). A ground-based or balloon telescope sees only a patch to begin with. The analysis therefore multiplies the map by a **mask** that is zero where the sky cannot be trusted and one (or a smooth value between zero and one) where it can.

The beam and the pixel window of Section 8.2.1 were easy to undo because they act on each multipole separately: they multiply $a_{\ell m}$ by a number T_ℓ. A mask acts on each *direction* separately: it multiplies the map, point by point. A product in one space is a convolution in the other, so the mask mixes the multipoles. Every $a_{\ell m}$ of the masked sky is a weighted sum of many $a_{\ell' m'}$ of the true sky, and every measured power is a blend of the true powers at neighbouring ℓ. The full-sky facts of Chapter 7 and Section 8.4 (independent $a_{\ell m}$, a diagonal covariance, a $\chi^2_{2\ell+1}$ law) all fail at once, and with them every formula we have so far.

In words, the questions are these:

1. What does multiplying the map by a mask do to each $a_{\ell m}$, and to the measured spectrum on average?
2. Can the average be undone, so that an unbiased estimate of C_ℓ comes out, and what does undoing it cost?
3. What does the mask do to the scatter of the estimates, to the correlations between multipoles and to the shape of their distribution?
4. How many simulated skies does it take to measure all of that well enough to put into a likelihood?

The first two have closed-form answers. The last two have none in general, and that is where the Monte Carlo loop of Section 8.5 stops verifying and starts measuring.

8.9.1 A one-dimensional warm-up: where the power of a wave goes

The mechanism is easiest to see on a line. Take a periodic box of length L and a pure wave $f(x) = \cos(2\pi k_0 x/L)$ with $k_0 = 40$. Its Fourier spectrum has all of its power at $k = \pm k_0$ and none anywhere else. Now "observe" it only on 40% of the box: multiply it by a window $W(x)$ that is one on an interval of length $a = 0.4L$ and zero outside.

With Fourier coefficients $\tilde{g}_k = \frac{1}{L}\int_0^L g(x)e^{-2\pi ikx/L}\mathrm{d}x$ and the expansion $W(x) = \sum_q \tilde{W}_q e^{2\pi iqx/L}$, the coefficients of the product are

$$\widetilde{(Wf)}_k \overset{(1)}{=} \frac{1}{L}\int_0^L \sum_q \tilde{W}_q e^{2\pi iqx/L}\,\tfrac{1}{2}\big(e^{2\pi ik_0x/L} + e^{-2\pi ik_0x/L}\big)e^{-2\pi ikx/L}\mathrm{d}x$$
$$\overset{(2)}{=} \tfrac{1}{2}\big(\tilde{W}_{k-k_0} + \tilde{W}_{k+k_0}\big). \tag{8.49}$$

(1) Insert the Fourier series of W and write the cosine as two exponentials.

(2) $\frac{1}{L}\int_0^L e^{2\pi i(q\pm k_0-k)x/L}\mathrm{d}x$ is one when $q = k \mp k_0$ and zero otherwise, so each sum over q keeps a single term.

The single spike at k_0 has become a copy of the window's own spectrum, centred on k_0. Whatever power the window has at a distance q from zero, the masked wave has at a distance q from k_0. This is **leakage**, and the window's spectrum is its kernel.

Example: how much leaks far away. For the sharp-edged window centred at x_c,

$$\tilde{W}_q \overset{(1)}{=} \frac{1}{L}\int_{x_c-a/2}^{x_c+a/2} e^{-2\pi iqx/L}\mathrm{d}x \overset{(2)}{=} e^{-2\pi iqx_c/L}\,\frac{\sin(\pi qa/L)}{\pi q}, \qquad |\tilde{W}_q|^2 = \frac{\sin^2(\pi qa/L)}{\pi^2 q^2}. \tag{8.50}$$

(1) The window is one on the interval and zero elsewhere.

(2) Integrate the exponential and use $e^{i\alpha} - e^{-i\alpha} = 2i\sin\alpha$.

The tail falls only as $1/q^2$, because the window jumps. The fraction of the power that lands more than Q modes away from the spike is the tail of this sum divided by the total. The total is $\sum_q |\tilde{W}_q|^2 = \frac{1}{L}\int W^2\mathrm{d}x = a/L$ by Parseval. In the tail $\sin^2$ averages to $\frac{1}{2}$ and there are two sides, so the tail is about $2\times\frac{1}{2}\sum_{q>Q} 1/(\pi^2q^2) \approx 1/(\pi^2 Q)$. For $Q = 10$ and $a/L = 0.4$ the leaked fraction is $1/(\pi^2\times 10\times 0.4) = 0.025$. The script `20_leakage_1d.py` measures 0.026.

Example: the same window with soft edges. Replace each jump by a half-period of a cosine, rising from zero to one over a few per cent of the box, keeping the same total length. A function whose derivatives are all continuous has Fourier coefficients that fall faster than any power of q, so the far tail nearly disappears: the script measures 1.6×10^{-3}, sixteen times less (Figure 8.22). The price is visible in the left panel: part of the observed interval is now down-weighted, so less of the wave is used at full weight. Softening the edge of a window to suppress leakage is called **apodisation** (from the Greek "removing the feet" of the kernel).

Everything that follows on the sphere is this picture with spherical harmonics in place of plane waves. One difference matters: on the line the window was fixed and the wave was unique, whereas the sky has $2\ell+1$ modes at each ℓ and the window couples each of them to many others, so the bookkeeping needs the angular-momentum coupling rules of quantum mechanics. The physics is identical: a mask with sharp edges has a slowly falling spectrum and leaks power far away; a tapered mask leaks less and uses the sky less efficiently.

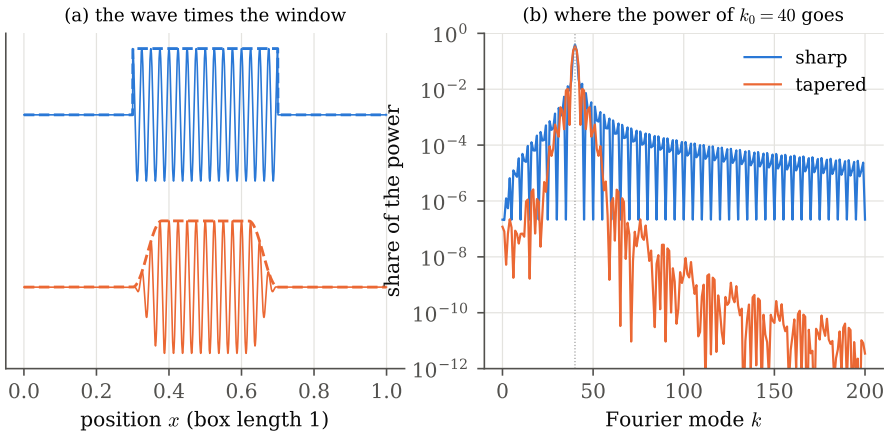


Figure 8.22: A single Fourier mode seen through 40% of the box. Left: the wave times a sharp window (blue) and times a cosine-tapered window (orange), offset for clarity. Right: the share of the power that lands at each Fourier mode. The sharp window spreads the spike at $k_0 = 40$ into a slowly falling comb of side lobes; the tapered one keeps it within a few modes and then falls by ten more decades.

8.9.2 The window and its moments

Working definition 8.5 (window, mask, sky fraction and moments)

A **window** $W(\hat{\boldsymbol{n}})$ assigns to each direction $\hat{\boldsymbol{n}}$ a weight between zero and one [one where the sky is used at full weight, zero where it is thrown away]. A binary window is a **mask** in the narrow sense. Its observed **sky fraction** $f_{\rm sky}$ is the fraction of the sphere where $W \neq 0$, and its **moments** w_i are defined by

$$f_{\rm sky} w_i = \frac{1}{4\pi} \int_{4\pi} \mathrm{d}\Omega \, W^i(\hat{\boldsymbol{n}}) \qquad \text{[Hivon (9)]},$$

so that w_i is the average of W^i over the observed region and $w_i = 1$ for every i when W is binary. The window's harmonic coefficients and power spectrum are

$$w_{\ell m} = \int \mathrm{d}\Omega \, W(\hat{\boldsymbol{n}}) Y^*_{\ell m}(\hat{\boldsymbol{n}}), \qquad \mathcal{W}_\ell = \frac{1}{2\ell+1} \sum_m |w_{\ell m}|^2 \qquad \text{[Hivon (10)]}.$$

Operationally: on HEALPix, $f_{\rm sky}$ is the fraction of pixels with $W > 0$, w_i is the mean of W^i over those pixels, and $\mathcal{W}_\ell$ is `anafast` applied to the mask itself.

The window's spectrum $\mathcal{W}_\ell$ will turn out to be the kernel of the mode coupling, the spherical version of $|\tilde{W}_q|^2$ in (8.50). Two of its values are fixed by the moments. The monopole is the square of the mean of the window,

$$\mathcal{W}_0 \overset{(1)}{=} |w_{00}|^2 \overset{(2)}{=} \left| \frac{1}{\sqrt{4\pi}} \int \mathrm{d}\Omega \, W \right|^2 \overset{(3)}{=} \left| \frac{4\pi f_{\rm sky} w_1}{\sqrt{4\pi}} \right|^2 = 4\pi f_{\rm sky}^2 w_1^2, \tag{8.51}$$

(1) At $\ell = 0$ there is one value of m, so the average over m is that single term.
(2) $Y_{00} = 1/\sqrt{4\pi}$, a constant.
(3) The definition of w_1 with $i = 1$.

and the weighted sum of all of them is the total of W^2 over the sky:

$$\sum_{\ell\geq 0}(2\ell+1)\mathcal{W}_\ell \overset{(1)}{=} \sum_{\ell m}|w_{\ell m}|^2 \overset{(2)}{=} \int \mathrm{d}\Omega\, W^2(\hat{\boldsymbol{n}}) \overset{(3)}{=} 4\pi f_{\mathrm{sky}} w_2. \tag{8.52}$$

(1) The definition of $\mathcal{W}_\ell$, multiplied by $2\ell+1$.
(2) Parseval's theorem on the sphere: expand $W = \sum w_{\ell m} Y_{\ell m}$ in $\int WW^*$ and use the orthonormality $\int Y_{\ell m} Y^*_{\ell' m'} = \delta_{\ell\ell'}\delta_{mm'}$ (Section 7.8.1). W is real, so $W^2 = |W|^2$.
(3) The definition of w_2 with $i = 2$.

These are the two identities quoted after [Hivon (10)]. The first says how much of the window is a constant offset; the second, how much power it has in total. The second is the one that will reappear as the white-noise level of a masked map.

Example: the controlled masks. To study the effect of sky coverage and of shape separately we use eight windows on $N_{\mathrm{side}} = 256$ maps (Figure 8.23, script `21_masks.py`): the full sky; polar caps keeping 10, 30 and 70% of the sky; the 10% cap with a 10° cosine taper; a Galactic-like band cut $|b| < 20°$ around the equator, binary and with a 5° taper; and 400 randomly placed holes of radius 1°, the shape of a point-source mask. A cap of half-opening θ_c has area $2\pi(1-\cos\theta_c)$, so a cap holding a fraction f_{sky} of the sphere has $\cos\theta_c = 1 - 2f_{\mathrm{sky}}$: $\theta_c = 36.87°$, 66.42° and 113.58°. The band keeps $f_{\mathrm{sky}} = 1 - \sin 20° = 0.6580$, and the pixels give 0.6576. Each hole removes $2\pi(1-\cos 1°)$ steradians, and 400 of them would remove $400 \times 2\pi(1-\cos 1°)/4\pi = 0.0305$ of the sky if they never touched; the pixels report 0.0301, because a few random holes overlap.

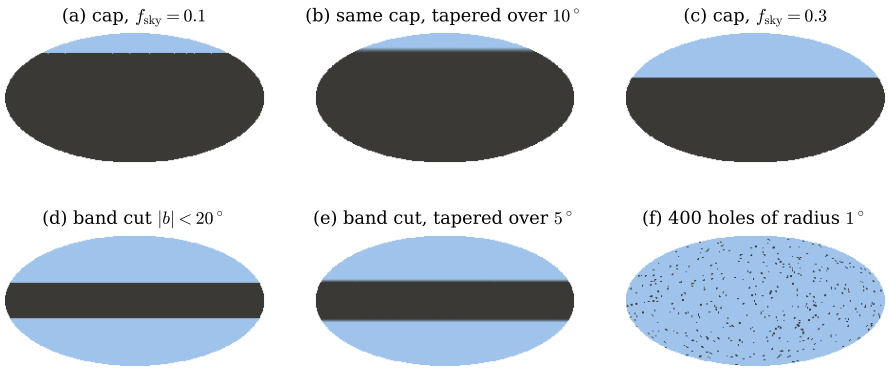


Figure 8.23: The controlled masks, in Mollweide projection with the north pole at the top. Blue is the observed sky, dark grey the removed sky. The tapers of (b) and (e) are visible as a paler rim inside the edge.

Example: the moments of a tapered cap. The taper of the 10% cap is $W = \sin^2(\pi d/2\Delta)$ for a distance $d < \Delta = 10°$ inside the edge, and $W = 1$ deeper in. How much statistical weight does such a taper cost? The moments need the averages of $W^2 = \sin^4(\pi d/2\Delta)$ and $W^4 = \sin^8(\pi d/2\Delta)$ over the cap, and the cap splits into two parts where these are easy: a core where $W = 1$, and a rim where W rises from zero.

How much of the cap is rim? The cap's edge is at $\theta_c = 36.87°$, where $\cos\theta_c = 0.8$, so the cap has area $2\pi(1-0.8) = 1.257$ sr. The rim runs from $\theta = 26.87°$ to 36.87° and has area $2\pi(\cos 26.87° - \cos 36.87°) = 2\pi(0.892 - 0.800) = 0.578$ sr. That is $0.578/1.257 = 0.46$ of the cap, and the core is the remaining $1 - 0.46 = 0.54$.

The average over the rim. Across the rim the variable $u = \pi d/2\Delta$ runs over a quarter period, $0 \leq u \leq \pi/2$. If the rim's area were spread evenly in d, the rim averages would be the

averages of $\sin^4 u$ and $\sin^8 u$ over that quarter period. For $\sin^4$ the double-angle identity $\sin^2 u = \frac{1}{2}(1 - \cos 2u)$ does it:

$$\langle \sin^4 u \rangle \overset{(1)}{=} \tfrac{1}{4}\big\langle 1 - 2\cos 2u + \cos^2 2u \big\rangle \overset{(2)}{=} \tfrac{1}{4}\big(1 - 0 + \tfrac{1}{2}\big) = \tfrac{3}{8}.$$

(1) Square $\frac{1}{2}(1 - \cos 2u)$.

(2) Over $0 \le u \le \pi/2$, $2u$ covers half a period of the cosine: $\int_0^{\pi/2} \cos 2u \, du = 0$, and $\cos^2 2u = \frac{1}{2}(1 + \cos 4u)$ averages to $\frac{1}{2}$.

For higher powers, integrating $\int_0^{\pi/2} \sin^k u \, du$ by parts once gives the step $I_k = \frac{k-1}{k} I_{k-2}$, so the average of $\sin^{2n} u$ is $\frac{(2n-1)(2n-3)\cdots 1}{(2n)(2n-2)\cdots 2}$: $\frac{3}{4} \cdot \frac{1}{2} = \frac{3}{8}$ for $n = 2$, as above, and $\frac{7}{8} \cdot \frac{5}{6} \cdot \frac{3}{4} \cdot \frac{1}{2} = \frac{35}{128}$ for $n = 4$. Then

$$w_2 \approx 0.54 \times 1 + 0.46 \times \tfrac{3}{8} = 0.71, \qquad w_4 \approx 0.54 \times 1 + 0.46 \times \tfrac{35}{128} = 0.67.$$

The approximation is the even spread of area in d. The area of a thin ring is proportional to $\sin\theta$, which grows from 0.45 at the inner edge of the rim to 0.60 at the outer edge, so the outer part of the rim, where W is small, has about a third more area per degree than the inner part. The pixel averages, $w_2 = 0.698$ and $w_4 = 0.653$, are a little lower than the estimate for exactly that reason. A taper this wide changes the effective sky fraction of [Hivon (17)], $f_{\rm sky} w_2^2 / w_4$, from 0.100 to 0.075: a quarter of the sky's statistical weight is given up to soften the edge.

Table 8.6 collects the moments of every mask. For a binary mask all the w_i are one and only $f_{\rm sky}$ matters; a taper lowers w_2 and w_4 and with them the effective sky fraction. Both identities, (8.51) and (8.52), hold for every mask: the first to within 10^{-15} (rounding), the second to within 5×10^{-4}, the power the mask has above the $\ell = 767$ at which the sum was stopped.

mask	$f_{\rm sky}$	w_2	w_4	w_2^2/w_4	$f_{\rm sky} w_2^2/w_4$
cap 0.1	0.100	1.000	1.000	1.000	0.100
cap 0.1, 10° taper	0.100	0.698	0.653	0.746	0.075
cap 0.3	0.300	1.000	1.000	1.000	0.300
cap 0.7	0.700	1.000	1.000	1.000	0.700
band $\lvert b\rvert < 20°$	0.658	1.000	1.000	1.000	0.658
band, 5° taper	0.658	0.924	0.911	0.936	0.616
400 holes of 1°	0.970	1.000	1.000	1.000	0.970

Table 8.6: Sky fraction and moments of the controlled masks at $N_{\rm side} = 256$. The last column is the effective sky fraction that enters the mode count of [Hivon (17)].

The spectrum of a mask. The kernel of the leakage is $\mathcal{W}_\ell$, and its tail decides how far power can travel. For a cap the tail can be worked out. A cap is symmetric about the pole, so only $m = 0$ survives, and with $x = \cos\theta$, $x_c = \cos\theta_c$,

$$w_{\ell 0} \overset{(1)}{=} 2\pi \sqrt{\frac{2\ell+1}{4\pi}} \int_{x_c}^{1} P_\ell(x)\, dx \overset{(2)}{=} 2\pi \sqrt{\frac{2\ell+1}{4\pi}} \, \frac{P_{\ell-1}(x_c) - P_{\ell+1}(x_c)}{2\ell+1}, \tag{8.53}$$

(1) $Y_{\ell 0} = \sqrt{(2\ell+1)/4\pi}\, P_\ell(\cos\theta)$, $W = 1$ for $\theta < \theta_c$, and the ϕ integral gives 2π.

(2) The Legendre identity $(2\ell+1)P_\ell = P'_{\ell+1} - P'_{\ell-1}$ integrates to $\int_{x_c}^1 P_\ell \, dx = [P_{\ell+1} - P_{\ell-1}]_{x_c}^1/(2\ell+1)$, and $P_\ell(1) = 1$ for every ℓ.

How fast does this fall at large ℓ? There the Legendre polynomial is a slowly decaying cosine, $P_\ell(\cos\theta_c) \approx \sqrt{2/(\pi\ell\sin\theta_c)}\cos[(\ell+\tfrac{1}{2})\theta_c - \pi/4]$. Three factors of (8.53) depend on ℓ, and they are best counted one at a time.

$$\begin{aligned}
\sqrt{\frac{2\ell+1}{4\pi}} &\overset{(1)}{\propto} \ell^{1/2}, \\
P_{\ell-1}(x_c) - P_{\ell+1}(x_c) &\overset{(2)}{=} \sqrt{\frac{2}{\pi\ell\sin\theta_c}}\left[\cos(\varphi-\theta_c) - \cos(\varphi+\theta_c)\right] \\
&\overset{(3)}{=} \sqrt{\frac{2}{\pi\ell\sin\theta_c}}\, 2\sin\theta_c\,\sin\varphi \;\propto\; \ell^{-1/2}, \\
\frac{1}{2\ell+1} &\overset{(4)}{\propto} \ell^{-1}.
\end{aligned}$$

(1) The normalisation of $Y_{\ell 0}$ grows as the square root of ℓ.
(2) The asymptotic form for $\ell \mp 1$, with $\varphi \equiv (\ell+\tfrac{1}{2})\theta_c - \pi/4$; the amplitudes for $\ell-1$ and $\ell+1$ differ from the one at ℓ only by a relative $O(1/\ell)$.
(3) $\cos(\varphi-\theta_c) - \cos(\varphi+\theta_c) = 2\sin\varphi\sin\theta_c$. The phases of the two polynomials differ by $2\theta_c$, which is not small for a cap, so the difference is as large as each term: no cancellation.
(4) The factor from the integral.

The product is $\ell^{1/2}\times\ell^{-1/2}\times\ell^{-1}$, so $w_{\ell 0} \propto \ell^{-1}$, oscillating with $\sin\varphi$, and $\mathcal{W}_\ell = |w_{\ell 0}|^2/(2\ell+1) \propto \ell^{-2}/\ell = \ell^{-3}$. Weighted by the number of modes, $(2\ell+1)\mathcal{W}_\ell \propto \ell^{-2}$: the same $1/q^2$ tail as the sharp window on the line. Figure 8.24 shows it: the binary cap and band fall by about three decades per decade of ℓ, the tapered versions drop by ten decades by $\ell \approx 200$ and then sit on the floor of the numerical precision. The holes behave differently. Each hole is a small disc of radius 1°, nearly a point on large scales, and 400 randomly placed points have a nearly flat spectrum up to $\ell \sim 180°/1°$; their $\mathcal{W}_\ell$ is small but does not fall. That will make the holes a poor neighbour for the high-ℓ spectrum.

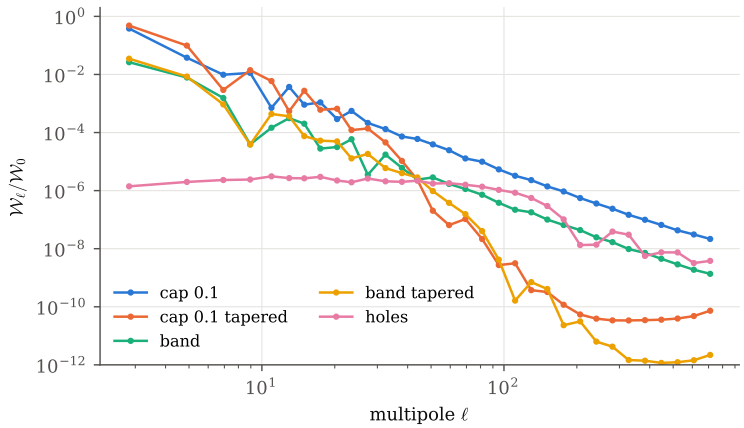


Figure 8.24: The power spectra of the masks, normalised to their monopole. The sharp cap and band fall slowly, about as ℓ^{-3}; their tapered versions collapse by ten decades and then stop at the numerical floor. The holes have a low, flat spectrum that stays up to the highest ℓ.

8.9.3 The pseudo-spectrum

The quantity we can compute from a masked map is the obvious one: transform the masked map and average the squared coefficients over m. With the masked coefficients

$$\tilde{a}_{\ell m} = \int \mathrm{d}\Omega\, \Delta T(\hat{\boldsymbol{n}}) W(\hat{\boldsymbol{n}}) Y^*_{\ell m}(\hat{\boldsymbol{n}}) \approx \Omega_{\mathrm{pix}} \sum_p W(p)\, \Delta T(p)\, Y^*_{\ell m}(p), \tag{8.54}$$

$$\widetilde{C}_\ell = \frac{1}{2\ell+1} \sum_{m=-\ell}^{\ell} |\tilde{a}_{\ell m}|^2, \tag{8.55}$$

[Hivon (11)–(13)], where the sum runs over the pixels p of area Ω_{pix} and ΔT is the observed map of Section 8.2, sky plus noise. $\widetilde{C}_\ell$ is the **pseudo-spectrum** (or pseudo-C_ℓ), and an estimator built on it is a **pseudo-C_ℓ estimator**. The cosmology notes write the same construction as $\widetilde{\Theta} = \Theta W$ with the mask expanded in $w_{\ell m}$ (see the cosmology notes). On the full sky $W = 1$ and $\widetilde{C}_\ell$ is the observed spectrum of Section 8.3.

Computing (8.54) for every (ℓ, m) by brute force costs $N_{\mathrm{pix}} \ell^2_{\mathrm{max}}$ operations, but on a grid of iso-latitude rings, like HEALPix, the sum over ϕ along each ring is a fast Fourier transform and the cost drops to $N_{\mathrm{pix}}^{1/2} \ell^2_{\mathrm{max}}$ [Hivon §2.2]. That is the reason the pseudo-spectrum, and not a likelihood, is the workhorse at high ℓ: for a Planck map it is a few seconds, where the exact likelihood of Section 8.1.1 needs N^3_{pix} operations. Our transforms use `map2alm` with `iter=0`, which is exactly the pixel sum (8.54); iterating would try to correct the quadrature error, which is negligible below $\ell = 2N_{\mathrm{side}}$ (Section 7.11.3) and is part of what the coupling calculation of Section 8.10 has to reproduce anyway.

Key idea: A mask is a product in the sky, a convolution in ℓ

Multiplying the map by W multiplies nothing in harmonic space; it convolves. Each pseudo-coefficient $\tilde{a}_{\ell m}$ is a combination of true coefficients at many (ℓ', m'), with weights set by the window's own coefficients $w_{\ell m}$. A sharp edge gives the window a tail $\propto \ell^{-3}$ and moves power far; a smooth edge confines it to neighbouring multipoles.

The code. The masks are a few lines each. A cap is a test on the colatitude of each pixel; the taper is applied to the distance d from the edge, measured inwards:

Listing 8.4: Masks and their cosine taper (`code/ch08/lib_masks.py`)

```
def taper(d, width):
    """Cosine taper of the distance d (rad) from the mask edge: 0 outside, 1 deep inside.

    W = 0 for d <= 0;  W = (1 - cos(pi d / width)) / 2  for 0 < d < width;  W = 1 beyond.
    width = 0 gives the binary (top-hat) mask.
    """
    d = np.asarray(d, dtype=float)
    if width <= 0:
        return (d > 0).astype(float)
    x = np.clip(d / width, 0.0, 1.0)
    return 0.5 * (1.0 - np.cos(np.pi * x))


def _angles(nside):
    theta, phi = hp.pix2ang(nside, np.arange(hp.nside2npix(nside)))
    return theta, phi


```

```
def cap(nside, fsky, apo_deg=0.0):
    """Observed region = a cap around the north pole holding a fraction fsky of the sphere.

    A cap of half-opening theta_c has area 2 pi (1 - cos theta_c), so cos theta_c = 1 - 2
    fsky.
    The taper (if any) eats into the cap from its edge inwards.
    """
    theta, _ = _angles(nside)
    theta_c = np.arccos(1.0 - 2.0 * fsky)
    return taper(theta_c - theta, apo_deg * DEG)
```

`taper` returns 0 outside the mask, rises as $\sin^2(\pi d/2\Delta) = \frac{1}{2}[1-\cos(\pi d/\Delta)]$ across a rim of width Δ, and returns 1 deeper in; with $\Delta = 0$ it is the binary mask. `cap` converts the requested sky fraction into the half-opening angle with $\cos\theta_c = 1 - 2f_{\rm sky}$ and applies the taper to $\theta_c - \theta$. Because the taper eats into the cap from its edge, the area with $W > 0$, and hence $f_{\rm sky}$, is the same with and without it, and only the moments change (Table 8.6).

The simulation itself is the loop of Section 8.5 with one extra line. Each of the 2000 skies is drawn from the fiducial spectrum, smoothed by the 30′ beam and the pixel window, and given white noise of 200 μK arcmin (the experiment of Section 8.2); then the same map is multiplied by each of the eight masks and transformed:

Listing 8.5: The cut-sky Monte Carlo: one sky, eight masks (`code/ch08/23_mc_masked.py`)

```
def pcl(m):
    return hp.alm2cl(hp.map2alm(m, lmax=L, iter=0))


def run_chunk(c):
    path = DATA / f"mc_masked_{c}.npz"
    if path.exists():
        return
    rng = rng_for("ch08", "23_mc_masked", stream=c)
    out = {k: np.empty((PER_CHUNK, L + 1), np.float32) for k in
           ["sky"] + ["pcl_" + k for k in KEYS] + ["filt_s", "filt_sn", "filt_n"]}
    t0 = time.time()
    for i in range(PER_CHUNK):
        a = cs.synalm(cl[: exp.lmax_sim + 1], exp.lmax_sim, rng)    # the sky, l <= 767
        s = cs.observe_signal(a, exp)                                # beam and pixel window
        n = cs.noise_map(exp, rng)                                   # white pixel noise
        out["sky"][i] = hp.alm2cl(a)
        d = s + n
        for k in KEYS:
            out["pcl_" + k][i] = pcl(MASKS[k] * d)
```

`pcl` is (8.54)–(8.55). Observing the *same* sky through all eight masks (line 67) is the common-random-numbers trick of Section 8.5: differences between masks are then differences of geometry, not of luck. The sky's own full-sky spectrum (line 64) is stored as well, so that every estimate can also be compared with the spectrum of the very sky it came from. The run was split into eight chunks of 250 skies with their own random streams, so that an interruption costs at most one chunk; it took 22 CPU-minutes, 0.66 s per sky for one synthesis and nine transforms. The full research version of this loop simulates maps at $N_{\rm side} = 2048$ up to $\ell \approx 2500$, with the real scan and noise; ours stops at $N_{\rm side} = 256$ and $\ell = 767$, which keeps the acoustic peaks and the start of the damping tail and gives up the high-ℓ multipoles where the beam has removed the signal anyway.

Watch out: Which sky fraction?

Three different "sky fractions" appear in cut-sky formulas, and they are equal only for a binary mask. f_{sky} counts the pixels with $W \neq 0$; $f_{\mathrm{sky}} w_2$ is the factor by which a mask lowers the power of white noise (Section 8.10); $f_{\mathrm{sky}} w_2^2 / w_4$ counts the effective number of modes (Section 8.12). Using f_{sky} for a tapered mask overestimates the number of modes, by a factor $0.100/0.075 = 1.33$ for our tapered cap.

8.10 The mode-coupling matrix

On the line, the masked wave's spectrum was the window's spectrum shifted to k_0 (Section 8.9.1). On the sphere the pseudo-spectrum $\widetilde{C}_\ell$ of (8.55), the m-average of the squared coefficients of the masked map, is again fixed on average by the window's spectrum $\mathcal{W}_\ell$ (Section 8.9.2) and the true spectrum C_ℓ. For a given sky spectrum C_ℓ and window W, we therefore seek the ensemble mean $\langle \widetilde{C}_\ell \rangle$. In MASTER the **mode-coupling matrix** $M_{\ell\ell'}$ gives that mean by specifying how much power from each true multipole ℓ' enters the measured multipole ℓ.

8.10.1 From the window to the coupling kernel

Masked coefficients are linear in the true ones. Expand the true sky as $\Delta T = \sum_{\ell' m'} a_{\ell' m'} Y_{\ell' m'}$ inside the integral (8.54):

$$\tilde{a}_{\ell m} = \sum_{\ell' m'} K_{\ell m, \ell' m'}\, a_{\ell' m'}, \qquad K_{\ell m, \ell' m'} \equiv \int \mathrm{d}\Omega\, Y^*_{\ell m}(\hat{\boldsymbol{n}})\, W(\hat{\boldsymbol{n}})\, Y_{\ell' m'}(\hat{\boldsymbol{n}}) \tag{8.56}$$

[Hivon (A16)–(A18)], [134, eqs 5a–6]. A physicist will recognise K: it is the matrix element $\langle \ell m | \hat{W} | \ell' m' \rangle$ of the operator "multiply by W" in the basis of angular-momentum eigenstates. A potential $V(\theta, \phi)$ in quantum mechanics couples the states $|\ell' m'\rangle$ in exactly this way, and the selection rules of atomic transitions are the selection rules of mode coupling. On the full sky $W = 1$, the operator is the identity, and $K_{\ell m, \ell' m'} = \delta_{\ell \ell'} \delta_{m m'}$. Because $\tilde{a}_{\ell m}$ is a linear combination of Gaussian variables it is Gaussian, but the $\tilde{a}_{\ell m}$ are no longer independent [Hivon §A.2]. Their covariance takes two moves:

$$\langle \tilde{a}_{\ell m} \tilde{a}^*_{\ell'' m''} \rangle \overset{(1)}{=} \sum_{\ell' m'} \sum_{\ell''' m'''} K_{\ell m, \ell' m'} K^*_{\ell'' m'', \ell''' m'''} \langle a_{\ell' m'} a^*_{\ell''' m'''} \rangle \overset{(2)}{=} \sum_{\ell' m'} C_{\ell'}\, K_{\ell m, \ell' m'} K^*_{\ell'' m'', \ell' m'}. \tag{8.57}$$

(1) Write both masked coefficients with (8.56), each with its own summation indices. The matrix K is fixed by the window, so the average over skies acts only on the true coefficients.

(2) Full-sky isotropy, $\langle a_{\ell' m'} a^*_{\ell''' m'''} \rangle = C_{\ell'} \delta_{\ell' \ell'''} \delta_{m' m'''}$ (Section 7.9): the Kronecker deltas collapse the second double sum onto the first.

Three spherical harmonics and the $3j$ symbols. To evaluate K we expand the window as well, $W = \sum_{LM} w_{LM} Y_{LM}$, which leaves an integral of three harmonics. It is the **Gaunt integral**:

$$\int \mathrm{d}\Omega\, Y_{\ell_1 m_1} Y_{\ell_2 m_2} Y_{\ell_3 m_3} = \sqrt{\frac{(2\ell_1+1)(2\ell_2+1)(2\ell_3+1)}{4\pi}} \begin{pmatrix} \ell_1 & \ell_2 & \ell_3 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \ell_1 & \ell_2 & \ell_3 \\ m_1 & m_2 & m_3 \end{pmatrix}, \tag{8.58}$$

with $Y^*_{\ell m} = (-1)^m Y_{\ell, -m}$ for the conjugate [Hivon (A20)]. The formula comes from the quantum theory of angular momentum, where the product of two harmonics is expanded in single harmonics with Clebsch–Gordan coefficients; we take it from there. Two of its consequences can be checked by hand, and they are the ones that matter physically. Each harmonic carries

a factor $e^{im_i\phi}$, so the ϕ integral of the product is $\int_0^{2\pi} e^{i(m_1+m_2+m_3)\phi}\mathrm{d}\phi$, which vanishes unless $m_1+m_2+m_3=0$: the projections are conserved. And $Y_{\ell_1 m_1}Y_{\ell_2 m_2}$ is a polynomial of degree $\ell_1+\ell_2$ in the components of $\hat{\boldsymbol{n}}$, so it contains no harmonic above $\ell_1+\ell_2$, and is orthogonal to $Y_{\ell_3 m_3}$ when $\ell_3 > \ell_1+\ell_2$; the same argument for each pair says that the three ℓ's must form a triangle. The arrays in brackets are **Wigner $3j$ symbols** [the amplitude for three angular momenta ℓ_1, ℓ_2, ℓ_3 with projections m_1, m_2, m_3 to add up to zero total angular momentum], the symmetric form of the Clebsch–Gordan coefficients of quantum mechanics, $\begin{pmatrix} j_1 & j_2 & j_3 \\ m_1 & m_2 & m_3 \end{pmatrix} = \frac{(-1)^{j_1-j_2-m_3}}{\sqrt{2j_3+1}}\langle j_1 m_1\, j_2 m_2 | j_3\, -m_3\rangle$. Four of their properties do all the work [Hivon (A21)–(A24)]:

1. *Triangle rule:* the symbol vanishes unless $|\ell_1-\ell_2| \le \ell_3 \le \ell_1+\ell_2$, as for adding two angular momenta in quantum mechanics.
2. *Projections:* it vanishes unless $m_1+m_2+m_3=0$.
3. *Parity:* with all $m_i=0$ it vanishes when $\ell_1+\ell_2+\ell_3$ is odd (the product of three harmonics with $m=0$ is odd under $\hat{\boldsymbol{n}}\to-\hat{\boldsymbol{n}}$ and integrates to zero).
4. *Orthogonality:*

$$\sum_{\ell_3 m_3}(2\ell_3+1)\begin{pmatrix}\ell_1 & \ell_2 & \ell_3\\ m_1 & m_2 & m_3\end{pmatrix}\begin{pmatrix}\ell_1 & \ell_2 & \ell_3\\ m_1' & m_2' & m_3\end{pmatrix} = \delta_{m_1 m_1'}\delta_{m_2 m_2'},$$

$$\sum_{m_1 m_2}\begin{pmatrix}\ell_1 & \ell_2 & \ell_3\\ m_1 & m_2 & m_3\end{pmatrix}\begin{pmatrix}\ell_1 & \ell_2 & \ell_3'\\ m_1 & m_2 & m_3'\end{pmatrix} = \frac{\delta_{\ell_3\ell_3'}\delta_{m_3 m_3'}}{2\ell_3+1},$$

   the second for ℓ_3 inside the triangle.

Finally, an even permutation of the columns leaves a $3j$ symbol unchanged and an odd one multiplies it by $(-1)^{\ell_1+\ell_2+\ell_3}$, so its *square* is symmetric in its columns.

The one sum that does the work. Before the main calculation, one double sum that it will meet. Fix ℓ and ℓ', take any numbers c_{LM}, and ask for the total over all projections m, m' of $\left|\sum_{LM} c_{LM}\,(\ell\ L\ \ell';\, -m\ M\ m')\right|^2$, where $(\ell\ L\ \ell';\, -m\ M\ m')$ is the $3j$ symbol with that top and bottom row. Then

$$\sum_{mm'}\left|\sum_{LM} c_{LM}\begin{pmatrix}\ell & L & \ell'\\ -m & M & m'\end{pmatrix}\right|^2 \overset{(1)}{=} \sum_{LM}\sum_{L'M'} c_{LM}\,c^*_{L'M'}\sum_{mm'}\begin{pmatrix}\ell' & \ell & L\\ m' & -m & M\end{pmatrix}\begin{pmatrix}\ell' & \ell & L'\\ m' & -m & M'\end{pmatrix}$$
$$\overset{(2)}{=} \sum_{LM}\sum_{L'M'} c_{LM}\,c^*_{L'M'}\,\frac{\delta_{LL'}\delta_{MM'}}{2L+1} \overset{(3)}{=} \sum_{LM}\frac{|c_{LM}|^2}{2L+1}. \tag{8.59}$$

(1) Write $|z|^2 = zz^*$ as a double sum over (L,M) and (L',M'); the $3j$ symbols are real, and the order of the sums can be exchanged. Move the columns cyclically, $(\ell\ L\ \ell') \to (\ell'\ \ell\ L)$, an even permutation that leaves each symbol unchanged.
(2) The second orthogonality relation, with $\ell_1=\ell'$, $\ell_2=\ell$ and the summed projections $m_1=m'$, $m_2=-m$ (as m runs from $-\ell$ to ℓ, so does $-m$); L, M, L', M' are held fixed, in the roles of $\ell_3, m_3, \ell_3', m_3'$. Terms with L outside the triangle vanish on both sides.
(3) The Kronecker deltas leave only $L'=L$, $M'=M$.

All the cross terms between different (L,M) have dropped out: the squared modulus of a sum has become a sum of squared moduli.

The average pseudo-spectrum. Now everything is in place. Insert (8.56) into (8.55) and average:

$$\begin{aligned}
\langle \widetilde{C}_\ell \rangle &\overset{(1)}{=} \frac{1}{2\ell+1} \sum_m \sum_{\ell' m'} C_{\ell'} \, |K_{\ell m, \ell' m'}|^2 \\
&\overset{(2)}{=} \frac{1}{2\ell+1} \sum_{\ell'} C_{\ell'} \sum_{mm'} \Big| \sum_{LM} w_{LM} (-1)^m \sqrt{\tfrac{(2\ell+1)(2L+1)(2\ell'+1)}{4\pi}} \begin{pmatrix} \ell & L & \ell' \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \ell & L & \ell' \\ -m & M & m' \end{pmatrix} \Big|^2 \\
&\overset{(3)}{=} \sum_{\ell'} C_{\ell'} \frac{2\ell'+1}{4\pi} \sum_{LM} |w_{LM}|^2 (2L+1) \begin{pmatrix} \ell & L & \ell' \\ 0 & 0 & 0 \end{pmatrix}^2 \frac{1}{2L+1} \\
&\overset{(4)}{=} \sum_{\ell'} C_{\ell'} \underbrace{\frac{2\ell'+1}{4\pi} \sum_L (2L+1) \mathcal{W}_L \begin{pmatrix} \ell & \ell' & L \\ 0 & 0 & 0 \end{pmatrix}^2}_{M_{\ell\ell'}} .
\end{aligned} \tag{8.60}$$

(1) $\widetilde{C}_\ell$ is the m-average of $|\tilde{a}_{\ell m}|^2$; with (8.57) at $\ell'' = \ell$, $m'' = m$, the average of $|\tilde{a}_{\ell m}|^2$ is $\sum_{\ell' m'} C_{\ell'} |K_{\ell m, \ell' m'}|^2$.
(2) Expand W in K and use the Gaunt integral (8.58) with $Y^*_{\ell m} = (-1)^m Y_{\ell,-m}$.
(3) The sign $(-1)^m$ does not depend on (L, M) and has modulus one, so it drops out of the squared modulus. Take the factor $\sqrt{(2\ell+1)(2\ell'+1)/4\pi}$ out of the square too, and apply (8.59) with $c_{LM} = w_{LM}\sqrt{2L+1}\,(\ell\ L\ \ell';0\ 0\ 0)$. The $(2\ell+1)$ that comes out cancels the prefactor $1/(2\ell+1)$.
(4) $\sum_M |w_{LM}|^2 = (2L+1)\mathcal{W}_L$ by the definition of the window's spectrum, and the squared symbol is symmetric in its columns.

This is [Hivon (A27)–(A31)], $\langle \widetilde{C}_\ell \rangle = \sum_{\ell'} M_{\ell\ell'} C_{\ell'}$ [Hivon (14)], with

$$M_{\ell\ell'} = \frac{2\ell'+1}{4\pi} \sum_L (2L+1) \mathcal{W}_L \begin{pmatrix} \ell & \ell' & L \\ 0 & 0 & 0 \end{pmatrix}^2 . \tag{8.61}$$

The matrix depends only on the window, through its spectrum, and not on the sky. It is not symmetric because of the factor $2\ell'+1$; Efstathiou [134] writes it as $M_{\ell\ell'} = (2\ell'+1)\,\Xi(\ell, \ell', \mathcal{W})$ and uses the symmetric matrix $G_{\ell\ell'} = M_{\ell\ell'}/(2\ell'+1)$ [134, eqs 9–11]. NaMaster writes the same expression with $P_L^{vw} = (2L+1)\,\mathrm{PCL}_L(v, w)$, the window's spectrum times its mode count, in place of $(2L+1)\mathcal{W}_L$ [169, eq. 12]; for two different windows v and w on two maps it uses their cross-spectrum, which reduces to ours when $v = w = W$.

Beam, noise and filtering. The beam and the pixel window act on the sky *before* the mask is applied: the instrument records the smoothed sky, and the mask is applied to the recorded map. The true spectrum in (8.60) is therefore $T_{\ell'}^2 C_{\ell'}$, with $T_\ell = B_\ell p_\ell$ of Section 8.2.1. A filter on the timestream or the map, which suppresses some scales, is modelled by a **transfer function** F_ℓ in the same place. Noise is added before masking too, and its masked power $\langle \widetilde{N}_\ell \rangle$ simply adds:

Key idea: The master equation of the cut sky

$$\langle \widetilde{C}_\ell \rangle = \sum_{\ell'} M_{\ell\ell'} F_{\ell'} T_{\ell'}^2 \langle C_{\ell'} \rangle + \langle \widetilde{N}_\ell \rangle \qquad \text{[Hivon (15)]},$$

with $M_{\ell\ell'}$ from (8.61). The coupling depends on the window alone; the beam and filter sit inside the sum, at the true multipole ℓ'.

Example: the full sky. For $W = 1$, $w_{00} = \sqrt{4\pi}$ and every other w_{LM} vanishes, so $\mathcal{W}_0 = 4\pi$ and $\mathcal{W}_{L>0} = 0$. Then

$$M_{\ell\ell'} \overset{(1)}{=} \frac{2\ell'+1}{4\pi} 4\pi \begin{pmatrix} \ell & \ell' & 0 \\ 0 & 0 & 0 \end{pmatrix}^2 \overset{(2)}{=} (2\ell'+1)\frac{\delta_{\ell\ell'}}{2\ell+1} = \delta_{\ell\ell'}.$$

(1) Only $L = 0$ survives the sum in (8.61).

(2) With $\ell_3 = 0$ the triangle rule forces $\ell = \ell'$, and $\begin{pmatrix} \ell & \ell & 0 \\ 0 & 0 & 0 \end{pmatrix} = (-1)^\ell/\sqrt{2\ell+1}$.

No coupling, as it must be.

Example: white noise through a mask. Uniform white noise has the same power N at every multipole. Its masked power is

$$\begin{aligned} \langle \widetilde{N}_\ell \rangle &\overset{(1)}{=} N \sum_{\ell'} M_{\ell\ell'} \overset{(2)}{=} \frac{N}{4\pi} \sum_L (2L+1)\mathcal{W}_L \sum_{\ell'} (2\ell'+1) \begin{pmatrix} \ell & \ell' & L \\ 0 & 0 & 0 \end{pmatrix}^2 \\ &\overset{(3)}{=} \frac{N}{4\pi} \sum_L (2L+1)\mathcal{W}_L \overset{(4)}{=} N f_{\rm sky} w_2. \end{aligned} \tag{8.62}$$

(1) The master equation with $C_{\ell'} = N$ for every ℓ' and no beam (noise is not smoothed).

(2) Insert (8.61) and exchange the sums.

(3) The first orthogonality relation with all projections zero: only $m_3 = 0$ survives and $\sum_{\ell'}(2\ell'+1)(\ell\ \ell'\ L;0\ 0\ 0)^2 = 1$.

(4) Parseval for the window, (8.52).

This is [Hivon (A32)]. Masked white noise stays white, at a level lowered by the average of W^2 over the sky: each pixel's noise is multiplied by W, so its variance by W^2. The same steps with the roles of ℓ and ℓ' exchanged give the column rule $\sum_\ell (2\ell+1)M_{\ell\ell'} = (2\ell'+1) f_{\rm sky} w_2$: all the power of one true multipole is redistributed, and its total is lowered by the same factor. These two *sum rules* are the first test of any coupling code.

Example: the flat-sky limit. On a small patch the sphere looks flat and the harmonics become plane waves [Hivon (A1)–(A15)]. The coupling of two Fourier modes is then simply the window's Fourier coefficient at their difference, $K_{\boldsymbol{k}_1\boldsymbol{k}_2} = \tilde{w}(\boldsymbol{k}_2 - \boldsymbol{k}_1)$, the two-dimensional version of (8.49), and the average power becomes $\langle \widetilde{C}_{k_1} \rangle = \int k_2 \mathrm{d}k_2\, M_{k_1k_2} \langle C_{k_2} \rangle$ with $M_{k_1k_2} = 2\pi \int k_3 \mathrm{d}k_3\, \mathcal{W}(k_3) J(k_1,k_2,k_3)$ [Hivon (A14)]. Here $J(k_1,k_2,k_3) = \frac{2}{\pi}(2k_1^2k_2^2 + 2k_1^2k_3^2 + 2k_2^2k_3^2 - k_1^4 - k_2^4 - k_3^4)^{-1/2}$ [Hivon (A10)] is the density of ways in which three vectors of lengths k_1, k_2, k_3 can close into a triangle, $\boldsymbol{k}_1 = \boldsymbol{k}_2 + \boldsymbol{k}_3$, which is momentum conservation. For large multipoles the squared $3j$ symbol tends to exactly the same expression in (ℓ_1, ℓ_2, ℓ_3) [Hivon (A26)]: the triangle rule of angular momentum is the triangle of momentum conservation, and the curved-sky kernel reduces to the flat one.

8.10.2 Computing the $3j$ symbols and the matrix

Only symbols with all projections zero enter (8.61), and for those there is a closed form. For even $J = \ell_1 + \ell_2 + \ell_3$ and $g = J/2$,

$$\begin{pmatrix} \ell_1 & \ell_2 & \ell_3 \\ 0 & 0 & 0 \end{pmatrix}^2 = \frac{(J-2\ell_1)!\,(J-2\ell_2)!\,(J-2\ell_3)!}{(J+1)!} \left[\frac{g!}{(g-\ell_1)!\,(g-\ell_2)!\,(g-\ell_3)!} \right]^2 \tag{8.63}$$

[Hivon (A25)], and zero for odd J.

Example: one symbol by hand. For $(2,2,2)$: $J=6$, $g=3$. The first factor is $2!\,2!\,2!/7! = 8/5040$ and the bracket is $3!/(1!\,1!\,1!) = 6$, so the square is $8\times 36/5040 = 2/35 = 0.057143$. The code returns 0.057143.

The factorials overflow a double-precision number beyond 170!, long before $\ell = 767$. The standard cure is to work with logarithms: $\ln n!$ is the log-gamma function $\ln\Gamma(n+1)$, the logarithm of (8.63) is a sum of eight such terms, and one exponential at the end gives a number of ordinary size. No recursion is needed for this family of symbols, and the closed form is exact up to rounding. Checked against the exact rational values from `sympy.physics.wigner` for six triples up to $(400, 380, 30)$, the worst relative error is 1.026×10^{-12}, and the orthogonality sum $\sum_{\ell_3}(2\ell_3+1)(100\ 140\ \ell_3;0\ 0\ 0)^2$ comes out as 1.000000000000.

Listing 8.6: The squared $3j$ symbol and the coupling matrix (`code/ch08/lib_masks.py`)

```
@njit(cache=True)
def _w3j000_sq(l1, l2, l3, lf):
    """(l1 l2 l3; 0 0 0)^2 from the closed form (Hivon et al. eq. A25), evaluated in
    logs."""
    L = l1 + l2 + l3
    if L % 2 == 1 or l3 < abs(l1 - l2) or l3 > l1 + l2:
        return 0.0
    g = L // 2
    lnv = (lf[L - 2 * l1] + lf[L - 2 * l2] + lf[L - 2 * l3] - lf[L + 1]
           + 2.0 * (lf[g] - lf[g - l1] - lf[g - l2] - lf[g - l3]))
    return math.exp(lnv)


def three_j_000_sq(l1, l2, l3):
    lf = log_factorials(l1 + l2 + l3 + 2)
    return _w3j000_sq(int(l1), int(l2), int(l3), lf)


@njit(parallel=True, cache=True)
def _coupling(wl, lmax_row, lmax_col, lf):
    nW = wl.shape[0]
    M = np.zeros((lmax_row + 1, lmax_col + 1))
    for l1 in prange(lmax_row + 1):
        for l2 in range(lmax_col + 1):
            s = 0.0
            lo = abs(l1 - l2)
            hi = min(l1 + l2, nW - 1)
            l3 = lo
            if (l1 + l2 + l3) % 2 == 1:      # only even l1 + l2 + l3 survive
                l3 += 1
            while l3 <= hi:
                s += (2 * l3 + 1) * wl[l3] * _w3j000_sq(l1, l2, l3, lf)
                l3 += 2
            M[l1, l2] = (2 * l2 + 1) / (4 * np.pi) * s
    return M
```

`_w3j000_sq` first applies the selection rules (odd J, triangle) and returns zero without computing anything; otherwise it evaluates the logarithm of (8.63) from a precomputed table `lf` of $\ln n!$. `_coupling` is (8.61) literally: for each pair (ℓ,ℓ') it sums over the multipoles L of the window allowed by the triangle rule, starting at $|\ell-\ell'|$ and stepping by two to skip the odd-parity terms that vanish. The loop over rows runs in parallel (`prange`, compiled by `numba`). For $\ell,\ell' \le 767$ a matrix takes a few seconds. It needs the window's spectrum up to $L = \ell+\ell'$, so the masks' spectra were computed to $L = 1535$ (script `22_coupling.py`).

The sum rules, numerically. The white-noise rule $\sum_{\ell'} M_{\ell\ell'} = f_{\rm sky} w_2$ holds for every row $\ell \le 400$ to within 8.0×10^{-4} for the sharp 10% cap, 2.4×10^{-5} for its tapered version and 7.0×10^{-4} for the holes; the residue is the part of each row that would lie beyond $\ell' = 767$ and is cut off. Sharp masks have the slowest tails and lose the most.

8.10.3 Checking the matrix against the sky

A formula with this many indices deserves two independent checks, one without random numbers and one with them.

A direct check, mode by mode. A sky whose power is all in one multipole ℓ' is a sum of its $2\ell'+1$ real modes: the real $Y_{\ell'0}$ with variance $C_{\ell'}$, and the real and imaginary parts of $a_{\ell'm'}$ for $m' > 0$, each with variance $C_{\ell'}/2$ (Section 7.10.1). Because the modes are independent, the average pseudo-spectrum of such a sky is the variance-weighted sum of the pseudo-spectra of each masked mode. Setting $C_{\ell'} = 1$ that sum is the column $M_{\ell\ell'}$ of the matrix, computed with $2\ell'+1$ map transforms and no $3j$ symbol and no random number. Script `24_coupling_check.py` does this for $\ell' = 20$ and 100 through the 10% cap and the band. The two computations agree to 8.5×10^{-13} wherever the column carries power (Figure 8.25). This is also the "Monte Carlo of the mask response" in its cheapest form: one transform per mode, nothing random.

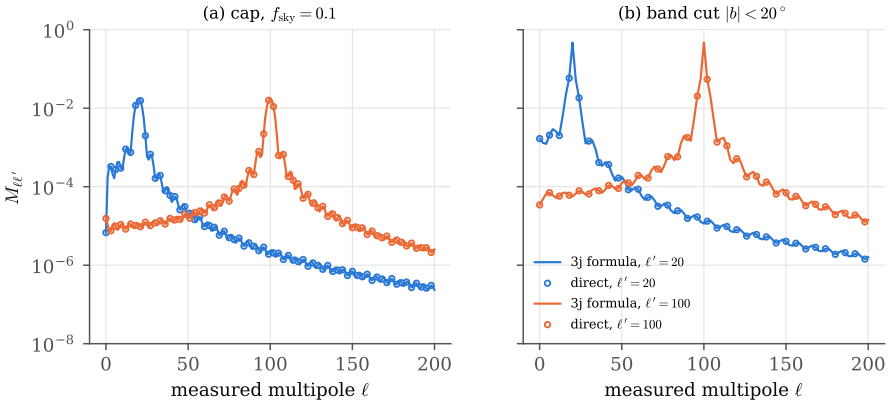


Figure 8.25: Two columns of the coupling matrix, $\ell' = 20$ (blue) and $\ell' = 100$ (orange), from the $3j$ formula (lines) and from transforming each masked real mode of that multipole (circles). (a) The 10% cap. (b) The band cut, shown at the multipoles of the same parity as ℓ' only: the band is symmetric about the equator and does not couple multipoles of opposite parity.

The band and parity. The right panel of Figure 8.25 shows a selection rule at work. The band is symmetric under $\theta \to \pi - \theta$, the reflection through the equator. Its coefficients w_{LM} are nonzero only for even L (and $M = 0$), because Y_{L0} is odd under the reflection for odd L. The parity rule of the $3j$ symbol then requires $\ell + \ell' + L$ even, so $\ell + \ell'$ is even: the band couples ℓ' only to multipoles of the same parity. The caps have no such symmetry and couple everything.

The Monte Carlo check. The second check uses the 2000 simulated skies of Section 8.9.3, with beam, pixel window and noise. For each mask we compare the mean of the simulated pseudo-spectra with the master equation, $\sum_{\ell'} M_{\ell\ell'} T_{\ell'}^2 C_{\ell'} + N f_{\rm sky} w_2$, and divide the difference by the Monte Carlo error of the mean, the simulated standard deviation over $\sqrt{2000}$. If the formula is right, these pulls are draws of a standard Gaussian, and their mean square over $2 \le \ell \le 600$ should be close to one. It is 1.05 on the full sky, 1.04 for the 30% cap, 1.06 for the band, 1.07

for the holes and 0.87 and 0.81 for the sharp and tapered 10% caps (Figure 8.26). The last two are a little below one, and that is expected: through a small cap neighbouring multipoles are strongly correlated (Section 8.12), so the 599 pulls contain only about 60 independent numbers, and their mean square scatters by about $\sqrt{2/60} \approx 0.2$ around one.

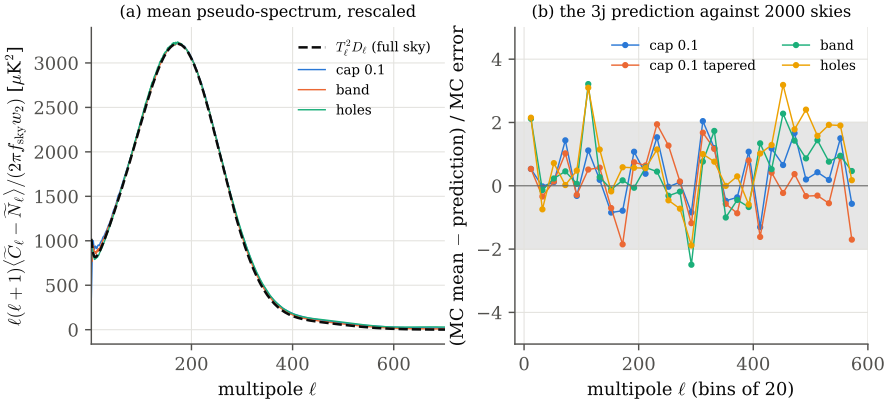


Figure 8.26: The master equation against 2000 simulated skies. (a) The mean pseudo-spectrum through four masks, noise subtracted and divided by $f_{\rm sky} w_2$, on the full-sky spectrum (dashed): at this scale the coupling is invisible except as a smoothing of the curve. (b) The mean minus the prediction of (8.61), in units of its Monte Carlo error, in bins of 20 multipoles: the residuals scatter within ±2 (grey) with no trend.

Watch out: The coupling needs the window to twice the band limit

$M_{\ell\ell'}$ involves $\mathcal{W}_L$ up to $L = \ell + \ell'$. Computing the mask's spectrum only to the analysis band limit $\ell_{\rm max}$ silently drops the coupling between the highest multipoles, worst for sharp masks whose spectra fall slowly. Compute $\mathcal{W}_L$ to $2\ell_{\rm max}$, and if the pixelisation cannot represent the mask that far ($L > 3N_{\rm side}$ on HEALPix), compute the mask at a higher resolution than the data.

8.10.4 Leakage and apodisation on the sphere

Figure 8.27 shows what happens to the power of a single true multipole, $\ell' = 100$, through each mask: the share $(2\ell + 1)M_{\ell,100}/\sum_\ell (2\ell + 1)M_{\ell,100}$ that lands at each measured ℓ. Every mask keeps most of the power near ℓ', in a central peak whose width is set by the size of the observed region, as for any diffraction pattern: a region of angular size Θ cannot resolve multipoles closer than about π/Θ. Outside the peak the masks differ by orders of magnitude, exactly as the one-dimensional windows did. The fraction landing more than 20 multipoles away is 2.9% for the sharp 10% cap and 0.33% for the tapered one; 1.3% for the sharp band and 0.44% for the tapered one; 1.5% and 0.63% for the 30% and 70% caps; and 2.3% for the holes, whose flat window spectrum carries power to every multipole.

Apodisation therefore buys a clean kernel and pays in two ways. The tapered region uses the sky at less than full weight, so the effective sky fraction $f_{\rm sky} w_2^2/w_4$ drops (Table 8.6); and the observed region is effectively smaller, so the central peak is wider and neighbouring multipoles become even more correlated. Whether the trade is worth it depends on the spectrum. A spectrum that falls steeply, like the CMB in its damping tail behind a 30′ beam, is ruined by a tail that carries power from the acoustic peaks into multipoles where there is almost none of their own; there apodisation is indispensable (Section 8.11.1). For a nearly flat spectrum leakage hardly matters, and the sharp mask, which keeps more modes, gives smaller error bars. Hivon et al. make the same point for their patch: a well-chosen apodisation can improve the

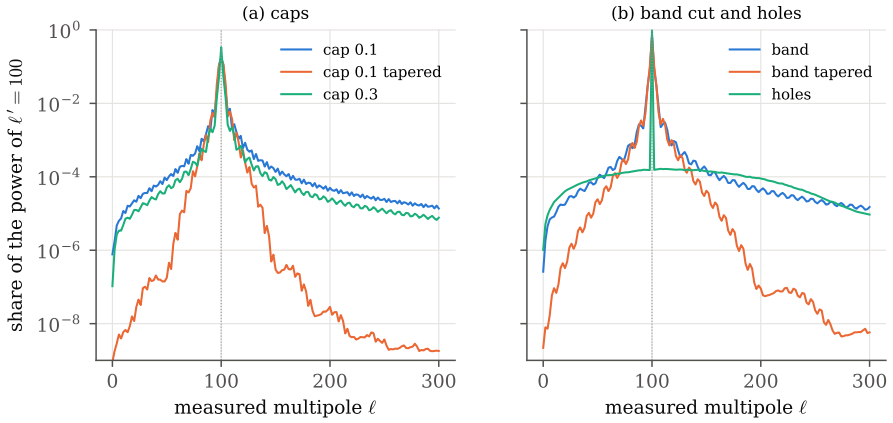


Figure 8.27: Where the power of the true multipole $\ell' = 100$ goes. (a) Caps: the sharp 10% cap (blue) and 30% cap (green) leak a slowly falling tail to all ℓ; the tapered 10% cap (orange) has a wider central peak but tails eight decades lower. (b) The band cut (blue), its tapered version (orange), and the holes (green), whose leakage is nearly flat in ℓ.

achievable spectral resolution of a non-convex survey [Hivon §3.1].

Example: the price of the taper in modes. At $\ell = 200$ in a bin of $\Delta\ell = 20$, the sharp 10% cap offers $(2\ell+1)\Delta\ell\, f_{\rm sky} = 802$ modes by the counting of [Hivon (17)]; the tapered cap offers $(2\ell+1)\Delta\ell\, f_{\rm sky} w_2^2/w_4 = 401 \times 20 \times 0.075 = 600$. If the bin were cosmic-variance limited its error would grow by $\sqrt{802/600} = 1.16$. That is the price of the taper; the next section shows what it buys.

8.11 Undoing the mask

The master equation of Section 8.10 says what the mask does on average: $\langle \widetilde{C}_\ell \rangle = \sum_{\ell'} M_{\ell\ell'} F_{\ell'} T_{\ell'}^2 C_{\ell'} + \langle \widetilde{N}_\ell \rangle$, with $M_{\ell\ell'}$ the coupling matrix of the window, T_ℓ the beam times the pixel window and F_ℓ the transfer function of any filter. We observe $\widetilde{C}_\ell$ and want C_ℓ. Three recipes suggest themselves, in increasing order of effort. Divide by a single number, the fraction of the sky that was used. Invert the matrix, multipole by multipole. Or average the multipoles into bins first and invert a small matrix between bins. We take them one at a time and ask of each the question that matters for an estimator: is it unbiased, and if not, why not?

8.11.1 The naive estimator: divide by the sky fraction

The quickest guess is that a mask removes a fraction of the power and nothing else:

$$\widehat{C}_\ell^{\,\rm naive} = \frac{\widetilde{C}_\ell - \langle \widetilde{N}_\ell \rangle}{f_{\rm sky} w_2\, T_\ell^2}. \tag{8.64}$$

It has a respectable origin. For a flat spectrum it is exact: the row rule $\sum_{\ell'} M_{\ell\ell'} = f_{\rm sky} w_2$ of (8.62) says that a constant spectrum comes out of the mask multiplied by $f_{\rm sky} w_2$ and nothing else. The question is how far a curved spectrum departs from that. Subtracting the exact average,

$$\langle \widehat{C}_\ell^{\,\rm naive} \rangle - C_\ell \overset{(1)}{=} \frac{\sum_{\ell'} M_{\ell\ell'} T_{\ell'}^2 C_{\ell'}}{f_{\rm sky} w_2 T_\ell^2} - C_\ell \overset{(2)}{=} \frac{\sum_{\ell'} M_{\ell\ell'} \left(S_{\ell'} - S_\ell\right)}{f_{\rm sky} w_2\, T_\ell^2}, \qquad S_\ell \equiv T_\ell^2 C_\ell, \tag{8.65}$$

(1) The master equation with $F = 1$; the noise term cancels against the subtraction.
(2) Write $C_\ell = S_\ell/T_\ell^2$ and use the row rule to put S_ℓ inside the sum as $\sum_{\ell'} M_{\ell\ell'} S_\ell/(f_{\text{sky}} w_2)$.

so the bias is the kernel-weighted difference between the smoothed spectrum S at the true multipoles ℓ' and at the measured one. Two mechanisms make it nonzero, one near ℓ and one far from it.

Near ℓ: curvature. Divide by $C_\ell = S_\ell/T_\ell^2$ and call $\kappa_{\ell'} \equiv M_{\ell\ell'}/(f_{\text{sky}} w_2)$ the kernel; by the row rule its weights add up to one. Where the spectrum is smooth over the kernel, expand $S_{\ell'}$ around ℓ:

$$\frac{\langle \widehat{C}_\ell^{\text{naive}} \rangle - C_\ell}{C_\ell} \overset{(1)}{=} \sum_{\ell'} \kappa_{\ell'} \frac{S_{\ell'} - S_\ell}{S_\ell} \overset{(2)}{=} \frac{S'_\ell}{S_\ell} \sum_{\ell'} \kappa_{\ell'} (\ell' - \ell) + \frac{S''_\ell}{2S_\ell} \sum_{\ell'} \kappa_{\ell'} (\ell' - \ell)^2 \overset{(3)}{=} \frac{1}{2} \sigma_K^2 \frac{S''_\ell}{S_\ell}.$$

(1) (8.65) divided by $C_\ell = S_\ell/T_\ell^2$.
(2) Taylor's expansion $S_{\ell'} = S_\ell + (\ell' - \ell) S'_\ell + \frac{1}{2}(\ell' - \ell)^2 S''_\ell$, with the derivatives taken with respect to ℓ; the constant term cancels.
(3) For a kernel symmetric about ℓ the first moment $\sum \kappa_{\ell'} (\ell' - \ell)$ vanishes (the factor $2\ell' + 1$ in M makes it slightly asymmetric, by a relative σ_K/ℓ); the second moment is the kernel's variance, σ_K^2.

At a peak $S'' < 0$ and the estimate comes out low; in a trough $S'' > 0$ and it comes out high. The mask smooths the peaks and fills the troughs.

Far from ℓ: leakage. If a fraction ε of the kernel's weight sits far away, at multipoles where S is near its peak value S_{peak}, the second piece of the bias is about $\varepsilon\, S_{\text{peak}}/S_\ell$. A tiny ε is harmless where S_ℓ is near the peak, and dominant where S_ℓ is a small fraction of it.

Example: why the damping tail is hit hardest. Compare the first peak, $\ell = 220$, with the damping tail at $\ell = 580$, and put a number on S_{peak}/S_ℓ. Behind the 30′ beam ($\sigma_b = 12.7' = 3.7 \times 10^{-3}$ rad, $\sigma_b^2 = 1.37 \times 10^{-5}$) the beam window is $e^{-\ell(\ell+1)\sigma_b^2}$. At $\ell = 220$, $\ell(\ell+1)\sigma_b^2 = 48\,620 \times 1.37 \times 10^{-5} = 0.67$ and the window is $e^{-0.67} = 0.51$; at $\ell = 580$, $336\,980 \times 1.37 \times 10^{-5} = 4.6$ and the window is $e^{-4.6} = 0.010$. The spectrum itself is $C_\ell = 2\pi D_\ell/\ell(\ell+1)$ with $D_{220} \approx 5700$ and $D_{580} \approx 2000\,\mu\text{K}^2$, so

$$\frac{C_{220}}{C_{580}} = \frac{5700/48\,620}{2000/336\,980} = \frac{0.117}{0.0059} \approx 20, \qquad \frac{S_{220}}{S_{580}} = 20 \times \frac{0.51}{0.010} \approx 1000.$$

How much leakage does it take to double the power seen at $\ell = 580$? The leaked power must equal S_{580}, which is 10^{-3} of the peak. The peak is about a hundred multipoles wide, so each of its multipoles needs to send only $10^{-3}/100 = 10^{-5}$ of its power to $\ell = 580$. The sharp 10% cap sends 2.9% of the power of each multipole more than twenty multipoles away (Section 8.10.4). Spread over the several hundred multipoles of the range, that is a few parts in 10^5 per multipole: the same size as the threshold. So the far term should be of the order of the true power at $\ell = 580$. That is what the simulations show (Figure 8.28a): the naive estimate through the sharp cap is biased by +4% in the first bin, rises slowly and reaches 112% near $\ell = 580$; through the holes it reaches 122%, and through the sharp band 49%. Tapering cures most of it: 7% for the tapered cap and 3% for the tapered band at the worst multipole. Even then a few per cent remain, much more than the error of an average over thousands of skies, so the naive estimator cannot be used for a precision spectrum.

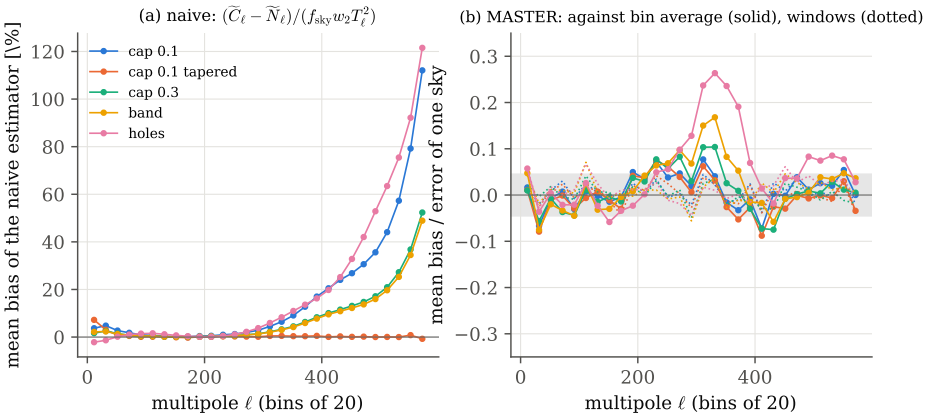


Figure 8.28: Bias of the two estimators over 2000 skies, in bins of 20 multipoles. (a) The naive estimator (8.64), in per cent: small at low ℓ, then rising steeply where the beam has removed most of the signal and leakage from the acoustic peaks takes over; the tapered cap (orange) stays within a few per cent. (b) MASTER, as a fraction of the error of a single sky, against the bin average of the input D_ℓ (solid) and against the bandpower-window prediction (dotted). The grey band is ±2 Monte Carlo errors of the mean.

8.11.2 Inverting the matrix, multipole by multipole

If M is invertible, $\widehat{C}_\ell = \sum_{\ell'} (M^{-1})_{\ell\ell'} (\widetilde{C}_{\ell'} - \langle \widetilde{N}_{\ell'} \rangle)/T_\ell^2$ is unbiased for every spectrum [134, eq. 12]. The condition number of M (the ratio of its largest to its smallest singular value, $2 \le \ell, \ell' \le 767$) tells us whether that is possible. It is 1.1 for the full sky, 1.8 for the 70% cap, 1.8 for the band and 1.1 for the holes, all harmless. For the 30% cap it is 1.3×10^{16} and for the 10% cap 1.3×10^{17}: those matrices are singular to machine precision. Figure 8.29 shows the eigenvalues. For the small caps a fixed fraction of them, 59% and 27%, falls off a cliff to the rounding floor, while for every other mask none does.

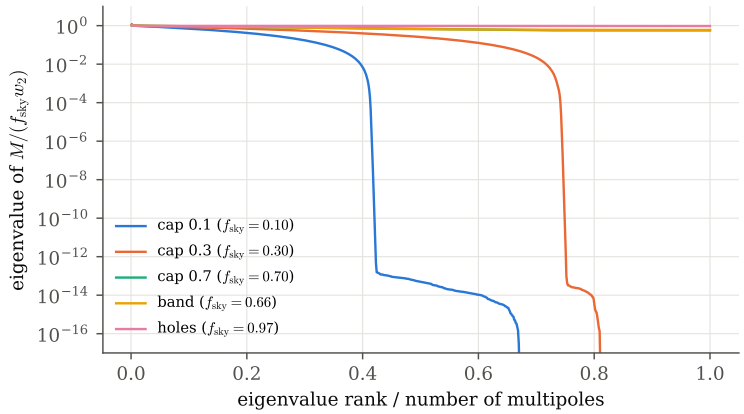


Figure 8.29: Eigenvalues of the coupling matrix (symmetrised and divided by $f_{\rm sky} w_2$), sorted and plotted against their rank as a fraction of the 766 multipoles. The 10% cap (blue) has a cliff at 0.41 and the 30% cap (orange) at 0.74: beyond it the eigenvalues are pure rounding. The 70% cap, the band and the holes have no cliff.

The event, and what could have produced it. The observed fact is sharp: a fixed fraction of the eigenvalues vanishes, 59% for one cap and 27% for the other, and none for the band, which removes far more sky than the 70% cap leaves out. So it is not the amount of sky that matters.

The candidates are the shape of the region, its size, and its smoothness, and the clean way to decide between them is to look at the coupling from the side of real space.

What the masked map measures in real space. Take the masked map $W\Delta T$ and average the product of its values over all pairs of directions $(\hat{\boldsymbol{n}}_1, \hat{\boldsymbol{n}}_2)$ an angle ϑ apart, anywhere on the sphere (where $W = 0$ the product is simply zero). Call the result the pseudo-correlation function $\widetilde{C}(\vartheta)$. Now average over skies. The window is fixed, so for each pair $\langle W_1\Delta T_1\, W_2\Delta T_2\rangle = W_1W_2\langle\Delta T_1\Delta T_2\rangle$, and for a statistically isotropic sky $\langle\Delta T_1\Delta T_2\rangle = C(\vartheta)$ depends only on the separation (Section 7.9.4). It is the same number for every pair in the average, so it comes out:

$$\langle\widetilde{C}(\vartheta)\rangle = C(\vartheta)\,\Xi_W(\vartheta), \qquad \Xi_W(\vartheta) \equiv \text{average of } W(\hat{\boldsymbol{n}}_1)W(\hat{\boldsymbol{n}}_2) \text{ over pairs a distance } \vartheta \text{ apart.}$$

The window's pair function Ξ_W is the fraction of pairs at separation ϑ that the window keeps, weighted by W; on the full sky $\Xi_W = 1$, and the masked map measures $C(\vartheta)$ itself.

The same statement in multipoles. A pair average of any fixed field f on the sphere has a Legendre series whose coefficients are the field's spectrum $f_\ell = \sum_m |f_{\ell m}|^2/(2\ell+1)$. With $x = \hat{\boldsymbol{n}}_1\cdot\hat{\boldsymbol{n}}_2$ and $x_0 = \cos\vartheta$,

$$\begin{aligned}
&\frac{\int \mathrm{d}\Omega_1 \int \mathrm{d}\Omega_2\, f(\hat{\boldsymbol{n}}_1) f(\hat{\boldsymbol{n}}_2)\,\delta(x-x_0)}{\int \mathrm{d}\Omega_1 \int \mathrm{d}\Omega_2\,\delta(x-x_0)} \\
&\overset{(1)}{=} \frac{1}{8\pi^2}\sum_\ell \frac{2\ell+1}{2} P_\ell(x_0) \int \mathrm{d}\Omega_1 \int \mathrm{d}\Omega_2\, f_1 f_2\, P_\ell(x) \\
&\overset{(2)}{=} \frac{1}{8\pi^2}\sum_\ell \frac{2\ell+1}{2} P_\ell(x_0)\,\frac{4\pi}{2\ell+1}\sum_m |f_{\ell m}|^2 \overset{(3)}{=} \sum_\ell \frac{2\ell+1}{4\pi} f_\ell\, P_\ell(\cos\vartheta).
\end{aligned}$$

(1) The denominator counts the pairs: 4π for $\hat{\boldsymbol{n}}_1$ times 2π for the uniform cos of the angle to $\hat{\boldsymbol{n}}_2$. In the numerator, write the delta function by the completeness of the Legendre polynomials, $\delta(x-x_0) = \sum_\ell \frac{2\ell+1}{2} P_\ell(x)P_\ell(x_0)$.

(2) The addition theorem, $P_\ell(\hat{\boldsymbol{n}}_1\cdot\hat{\boldsymbol{n}}_2) = \frac{4\pi}{2\ell+1}\sum_m Y_{\ell m}(\hat{\boldsymbol{n}}_1)Y^*_{\ell m}(\hat{\boldsymbol{n}}_2)$, splits the double integral into the product of $\int f Y_{\ell m}$ and $\int f Y^*_{\ell m}$, which are $f^*_{\ell m}$ and $f_{\ell m}$ for a real field.

(3) $\sum_m |f_{\ell m}|^2 = (2\ell+1) f_\ell$, and the constants combine as $\frac{1}{8\pi^2}\cdot\frac{2\ell+1}{2}\cdot\frac{4\pi}{2\ell+1}\cdot(2\ell+1) = \frac{2\ell+1}{4\pi}$.

With $f = W\Delta T$ and with $f = W$ this gives

$$\widetilde{C}(\vartheta) = \sum_\ell \frac{2\ell+1}{4\pi}\widetilde{C}_\ell P_\ell(\cos\vartheta), \qquad \Xi_W(\vartheta) = \sum_L \frac{2L+1}{4\pi}\mathcal{W}_L P_L(\cos\vartheta).$$

A product in ϑ is a convolution in ℓ, and the Legendre transform of $\langle\widetilde{C}(\vartheta)\rangle = C(\vartheta)\Xi_W(\vartheta)$ gives back the coupling matrix:

$$\begin{aligned}
\langle\widetilde{C}_\ell\rangle &\overset{(1)}{=} 2\pi\int_{-1}^{1}\mathrm{d}x\, P_\ell(x)\, C(x)\,\Xi_W(x) \overset{(2)}{=} 2\pi\sum_{\ell' L}\frac{2\ell'+1}{4\pi}C_{\ell'}\frac{2L+1}{4\pi}\mathcal{W}_L\int_{-1}^{1} P_\ell P_{\ell'} P_L\,\mathrm{d}x \\
&\overset{(3)}{=} \sum_{\ell'} C_{\ell'}\,\frac{2\ell'+1}{4\pi}\sum_L (2L+1)\mathcal{W}_L \begin{pmatrix}\ell & \ell' & L\\ 0 & 0 & 0\end{pmatrix}^2 = \sum_{\ell'} M_{\ell\ell'}C_{\ell'}.
\end{aligned} \tag{8.66}$$

(1) The inverse Legendre transform, $C_\ell = 2\pi\int P_\ell(x)C(x)\mathrm{d}x$ with $x = \cos\vartheta$ (Section 7.9.4), applied to $\widetilde{C}$, and the product rule.

(2) Expand both $C(x)$ and $\Xi_W(x)$ in Legendre series.

(3) $\int_{-1}^{1} P_\ell P_{\ell'} P_L\,\mathrm{d}x = 2\begin{pmatrix}\ell & \ell' & L\\ 0 & 0 & 0\end{pmatrix}^2$, the Gaunt integral (8.58) with all $m = 0$ and $Y_{\ell 0} = \sqrt{(2\ell+1)/4\pi}\,P_\ell$, integrated over ϕ.

The second derivation of (8.61) also explains the cliff. Now go through the candidates. *The cap:* no two points of a cap of half-opening θ_c are more than $2\theta_c$ apart, so $\Xi_W(\vartheta) = 0$ for $\vartheta > 2\theta_c$, and the data say nothing about $C(\vartheta)$ there. Any change of the C_ℓ that changes $C(\vartheta)$ only beyond $2\theta_c$ leaves every $\widetilde{C}_\ell$ untouched, and such changes form the null space of M. How large is it? Count numbers. A spectrum $C_2, \ldots, C_{\ell_{\max}}$ is about $\ell_{\max}$ numbers, and so is the function $C(\vartheta)$ it describes: a Legendre series stopped at $\ell_{\max}$ cannot change on scales finer than about $\pi/\ell_{\max}$ in angle, so it is fixed by about one value per cell of that width, $\ell_{\max}$ cells across $[0, \pi]$. The data fix $C(\vartheta)$ only in the cells with $\vartheta < 2\theta_c$, about $\ell_{\max} \cdot 2\theta_c/\pi$ numbers. The remaining $\ell_{\max}(1 - 2\theta_c/\pi)$ combinations of the C_ℓ change $C(\vartheta)$ only where the window has no pairs; they leave every $\widetilde{C}_\ell$ unchanged, so they are directions in which M gives zero. The predicted fraction of zero singular values is $1 - 2\theta_c/\pi$. For the 10% cap, $2\theta_c = 73.7°$ and the prediction is 59.0%; for the 30% cap, $2\theta_c = 132.8°$ and 26.2%. The simulations' cliffs are at 59% and 27%. *The* 70% *cap* has $2\theta_c = 227° > 180°$: every separation is observed and nothing is lost. *The band* removes a great deal of sky but keeps two caps, which contain pairs at every separation up to antipodal, so Ξ_W never vanishes. *The holes* remove small discs and leave pairs at every separation.

Key idea: When the coupling matrix can be inverted

$M_{\ell\ell'}$ is invertible exactly when the window contains pairs of directions at every separation $0 \le \vartheta \le \pi$, so that the true correlation function is seen, through Ξ_W, at all angles. A patch of angular diameter $\Theta < \pi$ loses a fraction $1 - \Theta/\pi$ of the multipole information to null modes. Large Galactic cuts are invertible; small patches never are.

This is the content of Efstathiou's rule of thumb, that M remains invertible for Galactic cuts up to about 30° above and below the plane, "if the two-point correlation function can be determined on all angular scales from the data within the uncut sky" [134, §2.1]. He also points out the difference between M and the coupling K of the coefficients themselves: K is singular for *any* cut, because combinations of $a_{\ell m}$ that live entirely inside the cut region leave no trace in the observed sky. The spectrum, a much smaller set of numbers, can be recovered when the coefficients cannot.

8.11.3 Binning first: the MASTER estimator

When M cannot be inverted, or when its inverse would amplify noise, the cure is to ask for less. The counting of Section 8.11.2 says how much less. A patch of diameter Θ ($2\theta_c$ for a cap) fixes $C(\vartheta)$ only for $\vartheta < \Theta$, about $\ell_{\max}\Theta/\pi$ numbers out of $\ell_{\max}$: one number for every π/Θ multipoles. What the data can determine is therefore not each C_ℓ but averages of the spectrum over blocks of about $\delta\ell \sim \pi/\Theta$ multipoles, and the null combinations are the ones that change sign within such a block, so that the block average does not see them. So we estimate averages over bins of multipoles, at least π/Θ wide, and assume that the spectrum is constant inside each.

Which spectrum is nearly constant across a bin? Not C_ℓ. On the large-scale plateau $\ell(\ell+1)C_\ell$ is about constant, so $C_\ell \propto 1/\ell(\ell+1)$: across a bin from $\ell = 22$ to 41, C_ℓ falls by a factor $(41 \cdot 42)/(22 \cdot 23) = 3.4$ while $\ell(\ell+1)C_\ell$ stays flat. The quantity to hold constant is therefore the "flattened" spectrum $D_\ell = \ell(\ell+1)C_\ell/2\pi$ [Hivon §3.4].

Working definition 8.6 (binning operators and the MASTER estimator)

For bins b with edges $\ell_{\rm low}^{(b)} \le \ell < \ell_{\rm low}^{(b+1)}$, the **binning operator** and its reciprocal are

$$P_{b\ell} = \frac{1}{2\pi}\frac{\ell(\ell+1)}{\ell_{\rm low}^{(b+1)} - \ell_{\rm low}^{(b)}}, \qquad Q_{\ell b} = \frac{2\pi}{\ell(\ell+1)} \qquad \text{for } \ell \text{ in bin } b\text{, zero otherwise}$$

[Hivon (20)–(21)]: P turns a spectrum C_ℓ into bin averages of D_ℓ, and Q turns a constant D_b back into a C_ℓ that is flat in D across the bin. The binned coupling and the **MASTER estimator** are

$$K_{bb'} = P_{b\ell} M_{\ell\ell'} F_{\ell'} T_{\ell'}^2 Q_{\ell' b'}, \qquad \widehat{D}_b = K^{-1}_{bb'} P_{b'\ell} \big(\widetilde{C}_\ell - \langle \widetilde{N}_\ell \rangle_{\rm MC}\big)$$

[Hivon (25)–(26)], with sums over repeated indices. $\langle \widetilde{N}_\ell \rangle_{\rm MC}$ is the noise pseudo-spectrum from noise-only simulations, and $\widehat{\mathcal{N}}_b = K^{-1}_{bb'} P_{b'\ell} \langle \widetilde{N}_\ell \rangle_{\rm MC}$ is the **noise on the sky** [Hivon (27)].

Why the recipe is unbiased, and for what, takes one chain. Start from the master equation and bin it:

$$P_{b\ell}\big(\langle \widetilde{C}_\ell \rangle - \langle \widetilde{N}_\ell \rangle\big) \overset{(1)}{=} P_{b\ell} M_{\ell\ell'} F_{\ell'} T_{\ell'}^2 C_{\ell'} \overset{(2)}{=} P_{b\ell} M_{\ell\ell'} F_{\ell'} T_{\ell'}^2 Q_{\ell' b'} D_{b'} \overset{(3)}{=} K_{bb'} D_{b'}, \tag{8.67}$$

(1) The master equation, multiplied by P (Hivon's (22)–(23)).
(2) *The assumption:* inside each bin D_ℓ is a constant $D_{b'}$, so $C_{\ell'} = Q_{\ell' b'} D_{b'}$.
(3) The definition of K.

A square system of $n_{\rm bins}$ equations, which a handful of bins makes small and well conditioned. Its solution is [Hivon (24)], and replacing the average by the measurement gives the estimator. The estimator is linear in the data, so its mean is $K^{-1}P(\langle \widetilde{C} \rangle - \langle \widetilde{N} \rangle)$ exactly; step (2) is the only approximation, and it is a statement about the *sky*, not about the estimator. For the real spectrum,

$$\langle \widehat{D}_b \rangle = \sum_\ell \mathcal{F}_{b\ell} C_\ell, \qquad \mathcal{F}_{b\ell} = \sum_{b'} K^{-1}_{bb'} \sum_{\ell'} P_{b'\ell'} M_{\ell'\ell} F_\ell T_\ell^2, \tag{8.68}$$

the **bandpower window functions** of NaMaster [169, eqs 18–19]. If D_ℓ really is flat in every bin, $\mathcal{F}_{b\ell} C_\ell$ is the plain bin average; otherwise the two differ by a known, deterministic amount, and the honest comparison of theory with data passes the theory through $\mathcal{F}$.

Example: the full sky. With $M = \mathbb{1}$ and $F = 1$, $K_{bb'}$ is diagonal and $K_{bb} = \langle T_\ell^2 \rangle_b$, the average of the squared beam over the bin (the factors $\ell(\ell+1)/2\pi$ of P and $2\pi/\ell(\ell+1)$ of Q cancel). Then $\widehat{D}_b = \langle T_\ell^2 \widetilde{D}_\ell \rangle_b / \langle T_\ell^2 \rangle_b$: a beam-weighted bin average, not the plain one. Where D_ℓ changes across a bin the two differ. Over 2000 skies the full-sky MASTER mean differs from the plain bin average by up to 10.8 Monte Carlo errors of the mean, but from the window prediction (8.68) by at most 2.7, which is what thirty Gaussian pulls give by chance. In units of the error bar of a *single* sky the bin-average offset is $0.24\,\sigma$ at most.

The cut-sky results. We ran MASTER with bins of 20 from $\ell = 2$ on every mask, with the noise on the sky subtracted (Figure 8.28b, script `25_master.py`). Against the bandpower windows the mean over 2000 skies agrees to within 2.0 Monte Carlo errors for the 10% cap, 2.2 for the 30% cap, 3.2 for the band and 3.0 for the holes, in the worst of the thirty bins below $\ell = 600$: MASTER is unbiased. Against the plain bin average the largest offset is $0.08\,\sigma$ (sharp 10% cap), $0.17\,\sigma$ (band) and $0.26\,\sigma$ (holes) of a single sky, a fraction of the error bar that a single measurement would not notice, but which a likelihood built on the plain bin average would

turn into a systematic shift. The same recipe works for the singular 10% cap. Its diameter is $2\theta_c = 73.7^\circ = 1.29\,\mathrm{rad}$, so the data fix one number per $\pi/1.29 \approx 2.4$ multipoles. A bin of 20 multipoles averages over about eight such blocks; every null combination changes sign within a block and averages to nearly zero across the bin, so it drops out of the binned equations, and $K_{bb'}$ is invertible: its condition number is 1.2×10^4, against 1.3×10^{17} for M, and close to the full sky's $7.3e+03$, which comes from the beam and the pixel window alone.

Watch out: Compare theory through the bandpower windows

A MASTER bandpower is an average of C_ℓ with weights $\mathcal{F}_{b\ell}$, not the plain bin average of D_ℓ. Comparing data with the binned theory $P_{b\ell}C_\ell$ introduces a bias that grows with the curvature of the spectrum across the bin and with the number of simulations used to test it. Binning should be chosen after the simulations are done [Hivon §3.4], and the theory compared through (8.68).

8.11.4 A filter on the map: the transfer function

Real data are filtered before the spectrum is taken. Timestreams are high-pass filtered to suppress the $1/f$ drift of the detectors, signals synchronous with the scan are removed, and the largest scales, poorly constrained by a small survey and liable to contaminate everything else, are cut away [Hivon §3.2]. A filter acts along the scan direction and therefore breaks isotropy: its effect on the spectrum is not a function of ℓ alone, and writing it as a transfer function F_ℓ in the master equation is an ansatz to be tested. Hivon et al. measure it by simulation. They run $N_{\mathrm{MC}}^{(\mathrm{s})}$ noise-free skies of a known spectrum C_ℓ^{th} through the full filtered observation, average their pseudo-spectra, and solve the master equation for F:

$$\langle \widetilde{C}_\ell \rangle_{\mathrm{MC}} = \sum_{\ell'} M_{\ell\ell'} F_{\ell'} T_{\ell'}^2 C_{\ell'}^{\mathrm{th}}.$$

A first guess. If $F_{\ell'} T_{\ell'}^2 C_{\ell'}^{\mathrm{th}}$ hardly changes across the kernel, it can be taken out of the sum at $\ell' = \ell$, and the row rule $\sum_{\ell'} M_{\ell\ell'} = f_{\mathrm{sky}} w_2$ leaves $\langle \widetilde{C}_\ell \rangle_{\mathrm{MC}} \approx f_{\mathrm{sky}} w_2 F_\ell T_\ell^2 C_\ell^{\mathrm{th}}$, so $F_\ell^{(0)} = \langle \widetilde{C}_\ell \rangle_{\mathrm{MC}} / (f_{\mathrm{sky}} w_2 T_\ell^2 C_\ell^{\mathrm{th}})$. This is the naive estimator of Section 8.11.1 applied to the filter, and it inherits its bias wherever the spectrum is curved.

A correction. Write the true transfer function as the guess plus a correction, $F = F^{(0)} + \delta F$, and insert it into the master equation:

$$\langle \widetilde{C}_\ell \rangle_{\mathrm{MC}} \overset{(1)}{=} \sum_{\ell'} M_{\ell\ell'} F_{\ell'}^{(0)} T_{\ell'}^2 C_{\ell'}^{\mathrm{th}} + \sum_{\ell'} M_{\ell\ell'}\, \delta F_{\ell'} T_{\ell'}^2 C_{\ell'}^{\mathrm{th}} \overset{(2)}{\approx} \sum_{\ell'} M_{\ell\ell'} F_{\ell'}^{(0)} T_{\ell'}^2 C_{\ell'}^{\mathrm{th}} + f_{\mathrm{sky}} w_2\, \delta F_\ell T_\ell^2 C_\ell^{\mathrm{th}},$$

(1) The master equation is linear in F.

(2) The first-guess approximation, now applied only to the correction: if $\delta F_{\ell'} T_{\ell'}^2 C_{\ell'}^{\mathrm{th}}$ varies slowly across the kernel, the row rule turns its sum into $f_{\mathrm{sky}} w_2$ times its value at ℓ.

Solving for δF_ℓ, and replacing the simulated average by a running average S over about 50 multipoles to smooth its Monte Carlo noise,

$$F_\ell = F_\ell^{(0)} + \frac{S\big(\langle \widetilde{C}_\ell \rangle_{\mathrm{MC}}\big) - \sum_{\ell'} M_{\ell\ell'} F_{\ell'}^{(0)} T_{\ell'}^2 C_{\ell'}^{\mathrm{th}}}{T_\ell^2 C_\ell^{\mathrm{th}} f_{\mathrm{sky}} w_2} \tag{8.69}$$

[Hivon (18)]. The exact coupled term now carries the curvature of the spectrum, and only the correction is treated approximately; one such step was enough for their patch. The step relies

on δF being smooth. Where $F^{(0)}$ has inherited a large leakage bias with the shape of the acoustic peaks, δF has that shape too, and the step can overshoot (we meet this below).

How well F is known. F_ℓ is, up to fixed factors, the average of $N_{\rm MC}^{(\rm s)}$ independent simulated pseudo-spectra, smoothed over about $\Delta\ell$ multipoles. Each pseudo-spectrum, so smoothed, carries about $\nu = (2\ell+1)\Delta\ell\, f_{\rm sky}$ modes and has a relative variance $2/\nu$ (Section 8.4.2); an average of $N_{\rm MC}^{(\rm s)}$ of them divides that by $N_{\rm MC}^{(\rm s)}$. So $\delta F_\ell/F_\ell \approx \sqrt{2/[(2\ell+1)\Delta\ell\, f_{\rm sky}]}/\sqrt{N_{\rm MC}^{(\rm s)}}$ [Hivon (16), (19)]. They also showed that F barely depends on the assumed spectrum: for a flat and an open universe, whose first peaks are at $\ell \approx 220$ and 420, the two transfer functions differ by less than 1% for $50 < \ell < 1000$ [Hivon Fig. 2].

Example: a ring filter on the 10% cap. Our toy filter imitates destriping: on each ring of constant latitude inside the 10% cap, subtract the mean of the observed pixels (`ring_filter` in `23_mc_masked.py`). It removes, ring by ring, the pattern a scan along that ring cannot distinguish from a detector offset, and with it most of the largest-scale power. The first 500 signal-only skies calibrate F; the other 1500 skies, with noise, test it (Figure 8.30). Two observations stand out.

The ratio $F^{(0)}$ is a poor starting point: at $300 \le \ell \le 600$ its median is 1.26, larger than one, because it inherits the leakage bias of the naive estimator, and one step of (8.69) brings it only to 1.06 and oscillates wildly at low ℓ. For the sharp cap, the running average S and the coupled guess do not cancel: the iteration converges only if $F^{(0)}$ is already close. The robust alternative uses the binning of MASTER for F as well: take F constant in each bin, $F_\ell = R_{\ell b}F_b$ with $R_{\ell b} = 1$ in bin b, and solve the small linear system $P_{b\ell}M_{\ell\ell'}T^2_{\ell'}C^{\rm th}_{\ell'}R_{\ell' b'}F_{b'} = P_{b\ell}\langle\widetilde{C}_\ell\rangle_{\rm MC}$ directly. It gives $F_b = 0.88$ in the first bin ($2 \le \ell < 22$), 0.97 in the second, 0.991 in the fifth and 1.000 at high ℓ: the filter takes out large scales only, as it should.

With this F, MASTER on the independent test skies is unbiased again: in the first bin the error is -1.3%, against -13% if the filter is ignored (-3.6% in the second), and the largest offset from the bin average is 0.13 of a single sky's error bar. How well do 500 skies know F? Solve for F_b separately in 4 disjoint groups of 500 skies. Above $\ell = 100$ their scatter is 1.55 ± 0.13 times the prediction [Hivon (19)]: through a cap this small, deconvolution correlates the bins and the counting formula is optimistic, but even so 500 skies pin F down to well under a per cent above $\ell \approx 100$.

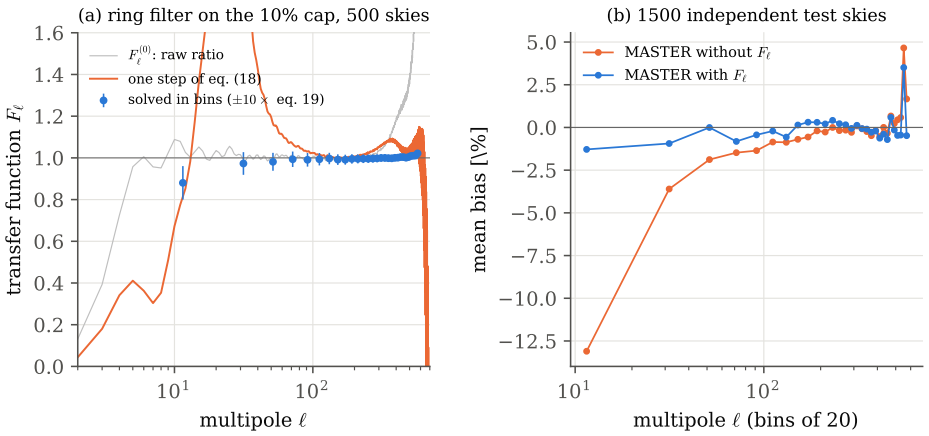


Figure 8.30: The transfer function of the ring filter on the 10% cap. (a) The raw ratio $F^{(0)}$ (grey), one step of (8.69) (orange) and the solution in bins (blue, error bars ten times the prediction of [Hivon (19)]). (b) MASTER on 1500 test skies, mean bias in per cent, without F (orange) and with it (blue): ignoring the filter loses 13% of the power in the first bin.

A full analysis measures F with end-to-end simulations: the sky is turned into timestreams with the real pointing, filtered exactly as the data, and made into maps by the same map-maker. That is what Hivon et al. did for Boomerang, and its cost, about half a CPU-day for 300 simulations on the computers of 2001 [Hivon (33)], is why our version filters the maps instead of timestreams. The reduction keeps the logic (simulate, average, solve for F, test on independent skies) and loses the realism of the scan: our F describes a toy filter, not an instrument.

8.12 The scatter of a cut-sky spectrum

MASTER (Section 8.11.3) gives an unbiased estimate. A likelihood needs much more than a mean: the width of the scatter of every estimate, the correlations between estimates at different multipoles, and the shape of the distribution. On the full sky all three had exact answers (Sections 7.12 and 8.4.1): $(2\ell+1)\widehat{C}_\ell/C_\ell$ is $\chi^2_{2\ell+1}$, the variance is $2C_\ell^2/(2\ell+1)$ plus noise, and different multipoles are independent. A mask breaks all three, and for a general mask none has a closed form. The 2000 simulated skies of Section 8.9.3 are now the instrument, and the formulas are what is being tested.

8.12.1 The shape of the distribution

What the simulations show. Figure 8.31 shows the distribution of the pseudo-spectrum, divided by its mean, at $\ell = 2$, 10 and 50, on the full sky and through the 30% and 10% caps. At $\ell = 50$ every panel is a narrow, nearly symmetric bump. At $\ell = 2$ the distributions are lopsided: a peak well below the mean and a long tail to large values. Through the 10% cap the quadrupole's most probable value is close to zero. The fraction of skies that fall below their own mean, $\frac{1}{2}$ for a symmetric law, is 0.58 on the full sky at $\ell = 2$, 0.60 through the 30% cap and 0.64 through the 10% cap: a typical sky shows less large-scale power than the theory predicts on average, and the smaller the patch, the more so.

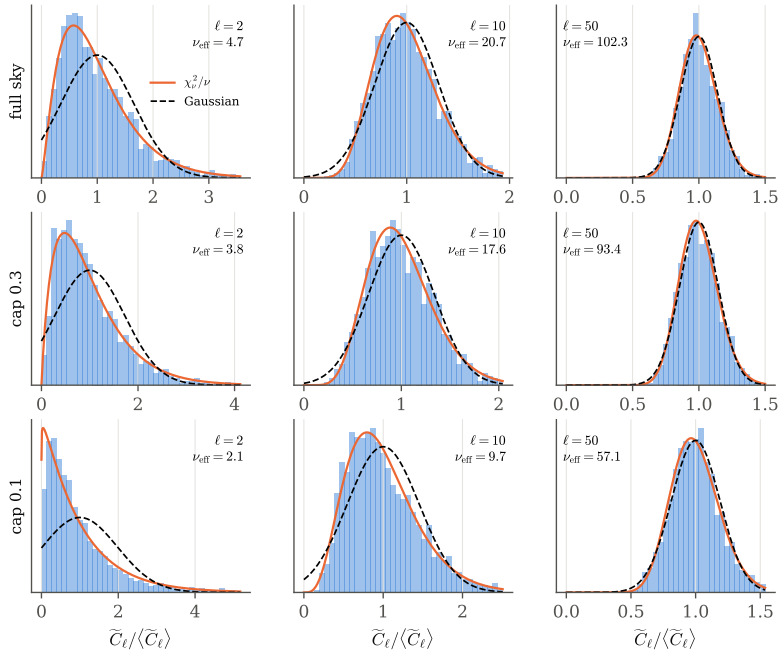


Figure 8.31: Histograms of $\widetilde{C}_\ell/\langle\widetilde{C}_\ell\rangle$ over 2000 skies, at $\ell = 2$, 10 and 50 (columns) for the full sky and the 30% and 10% caps (rows). Orange: a χ^2_ν/ν with $\nu = \nu_{\rm eff}$ fixed by the simulated mean and variance; dashed: the Gaussian of the same mean and variance. The white line marks the mean.

What could have produced that shape? Two candidates come to mind. The first is the central limit theorem: $\widetilde{C}_\ell$ is a sum of $2\ell+1$ squared coefficients, so perhaps it is Gaussian. The histograms rule this out at low ℓ, and the reason is that the sum has few terms. The second is the full-sky law with fewer degrees of freedom: a scaled χ^2_ν with ν the number of independent modes the patch supplies. That candidate has a precise justification. The masked coefficients are jointly Gaussian with the covariance (8.57), and $\widetilde{C}_\ell$ is a quadratic form in them.

The real numbers inside one pseudo-multipole. The masked map is real, so $\tilde{a}_{\ell,-m} = (-1)^m \tilde{a}^*_{\ell m}$: the coefficients with negative m repeat those with positive m, and $\tilde{a}_{\ell 0}$ is real. Collect the independent real numbers as in Section 7.10.1: $z_0 = \tilde{a}_{\ell 0}$, and for each $m = 1, \dots, \ell$ the two numbers $\sqrt{2}\,\mathrm{Re}\,\tilde{a}_{\ell m}$ and $\sqrt{2}\,\mathrm{Im}\,\tilde{a}_{\ell m}$. That is $1 + 2\ell$ real numbers, a vector $\boldsymbol{z}$. Then

$$(2\ell+1)\widetilde{C}_\ell = \sum_{m=-\ell}^{\ell} |\tilde{a}_{\ell m}|^2 = \tilde{a}^2_{\ell 0} + 2\sum_{m=1}^{\ell} \left[(\mathrm{Re}\,\tilde{a}_{\ell m})^2 + (\mathrm{Im}\,\tilde{a}_{\ell m})^2\right] = \boldsymbol{z}^{\mathsf{T}}\boldsymbol{z},$$

because $|\tilde{a}_{\ell,-m}| = |\tilde{a}_{\ell m}|$ counts each positive m twice; the factors $\sqrt{2}$ give every component of $\boldsymbol{z}$ the same weight one. The covariance Σ of $\boldsymbol{z}$ follows from (8.57) and the reality condition, by taking real and imaginary parts. Then

$$(2\ell+1)\widetilde{C}_\ell \overset{(1)}{=} \boldsymbol{u}^{\mathsf{T}}\Sigma\,\boldsymbol{u} \overset{(2)}{=} \boldsymbol{u}^{\mathsf{T}} O\Lambda O^{\mathsf{T}}\boldsymbol{u} \overset{(3)}{=} \sum_k \lambda_k\, v_k^2, \qquad v_k \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \tag{8.70}$$

(1) Write $\boldsymbol{z} = \Sigma^{1/2}\boldsymbol{u}$ with $\boldsymbol{u}$ a vector of independent standard Gaussians; then $\boldsymbol{z}^{\mathsf{T}}\boldsymbol{z} = \boldsymbol{u}^{\mathsf{T}}\Sigma\boldsymbol{u}$.
(2) Diagonalise the symmetric matrix $\Sigma = O\Lambda O^{\mathsf{T}}$ with O orthogonal and $\Lambda = \mathrm{diag}(\lambda_k)$.
(3) $\boldsymbol{v} = O^{\mathsf{T}}\boldsymbol{u}$ is again a vector of independent standard Gaussians, because an orthogonal rotation preserves the identity covariance.

A cut-sky pseudo-multipole is a weighted sum of independent χ^2_1 variables, with weights equal to the eigenvalues of the covariance of its own coefficients. On the full sky all the weights are equal to C_ℓ and the sum is $C_\ell \chi^2_{2\ell+1}$. Through a mask the weights spread out.

Which χ^2 looks most like it? Try the model $Y = c\,\chi^2_\nu/\nu$, a scaled χ^2 with two free numbers, c and ν, and choose them so that Y has the same mean and variance as $X \equiv (2\ell+1)\widetilde{C}_\ell = \sum_k \lambda_k v_k^2$:

$$\langle Y\rangle \overset{(1)}{=} c = \sum_k \lambda_k = \langle X\rangle, \qquad \mathrm{Var}\,Y \overset{(2)}{=} \frac{c^2}{\nu^2}\,2\nu = \frac{2c^2}{\nu} = 2\sum_k \lambda_k^2 = \mathrm{Var}\,X,$$

$$\Longrightarrow \quad \nu_{\text{eff}} \overset{(3)}{=} \frac{2\langle X\rangle^2}{\mathrm{Var}\,X} = \frac{2\langle \widetilde{C}_\ell\rangle^2}{\mathrm{Var}\,\widetilde{C}_\ell} = \frac{\left(\sum_k \lambda_k\right)^2}{\sum_k \lambda_k^2} \overset{(4)}{\le} 2\ell+1. \tag{8.71}$$

(1) A χ^2_ν has mean ν; each v_k^2 is a χ^2_1 with mean one.
(2) A χ^2_ν has variance 2ν, and a constant factor comes out squared; the v_k^2 are independent, so their variances $2\lambda_k^2$ add.
(3) Solve the second equation for ν and insert $c = \langle X\rangle$. The factor $2\ell+1$ cancels in the ratio.
(4) The Cauchy–Schwarz inequality for the vectors $(1,\dots,1)$ and $(\lambda_1,\dots,\lambda_{2\ell+1})$: $\left(\sum_k 1\cdot\lambda_k\right)^2 \le \left(\sum_k 1^2\right)\left(\sum_k \lambda_k^2\right) = (2\ell+1)\sum_k \lambda_k^2$, with equality only when the two vectors are parallel, that is, when all the weights are equal.

A few large weights and many small ones make ν_{eff} small. This moment-matched χ^2 is the Welch–Satterthwaite approximation; it is exact in its first two moments and approximate beyond.

The third moment as a test. Cumulants of independent terms add, and a constant factor λ multiplies the third cumulant by λ^3. A χ^2_1 has variance 2 and third cumulant 8 (Section 8.4.1), so

$\lambda_k v_k^2$ contributes $2\lambda_k^2$ and $8\lambda_k^3$. The skewness, third cumulant over variance to the power 3/2, is

$$\frac{8\sum_k \lambda_k^3}{\left(2\sum_k \lambda_k^2\right)^{3/2}} = \frac{\sqrt{8}\sum_k \lambda_k^3}{\left(\sum_k \lambda_k^2\right)^{3/2}} \qquad \text{against} \qquad \sqrt{\frac{8}{\nu_{\text{eff}}}} \quad \text{for the model.}$$

With n equal weights the left side is $\sqrt{8}\,n\lambda^3/(n\lambda^2)^{3/2} = \sqrt{8/n}$ and the two agree exactly; they stay close as long as the weights are not too unequal.

The mode count for a smooth spectrum. Hivon et al.'s count can be recovered from (8.71) with one picture of the modes. At high ℓ a mode of multipole ℓ is a short wave, and the window acts on it locally: a mode concentrated near direction $\hat{\boldsymbol{n}}$ is multiplied by $W(\hat{\boldsymbol{n}})$, and its variance by $W(\hat{\boldsymbol{n}})^2$. If the spectrum is smooth, so that it does not matter which neighbouring multipole the mode came from, the $2\ell+1$ eigenvalues are $\lambda_k \approx C_\ell W(\hat{\boldsymbol{n}}_k)^2$, with the $2\ell+1$ directions $\hat{\boldsymbol{n}}_k$ spread evenly over the sphere. Sums over k then become sky averages, which are the moments of the workdef in Section 8.9.2:

$$\sum_k \lambda_k \overset{(1)}{=} (2\ell+1)\,C_\ell \frac{1}{4\pi}\int \mathrm{d}\Omega\, W^2 = (2\ell+1) C_\ell f_{\text{sky}} w_2, \qquad \sum_k \lambda_k^2 \overset{(2)}{=} (2\ell+1)\,C_\ell^2 f_{\text{sky}} w_4,$$
$$\nu_{\text{eff}} \overset{(3)}{=} \frac{(2\ell+1)^2 C_\ell^2 f_{\text{sky}}^2 w_2^2}{(2\ell+1) C_\ell^2 f_{\text{sky}} w_4} = (2\ell+1)\, f_{\text{sky}} \frac{w_2^2}{w_4}. \tag{8.72}$$

(1) An average over evenly spread directions is the sky average; $\frac{1}{4\pi}\int W^2 = f_{\text{sky}} w_2$ by the definition of the moments.
(2) The same with $\lambda_k^2 \propto W^4$.
(3) (8.71).

For a bin of $\Delta\ell$ multipoles, treated as independent, the count is multiplied by $\Delta\ell$. A binary mask gives $(2\ell+1)f_{\text{sky}}$: the fraction of the modes that land on the patch. A taper gives the modes near the edge less weight than the others, which makes the λ_k unequal and lowers the count by $w_2^2/w_4 \le 1$. The picture assumes that each mode belongs to one multipole and sees the window at one place; that is the assumption tested below.

Working definition 8.7 (effective number of modes)

The **effective number of modes** of an estimate X that is a positive quadratic form in Gaussian variables is $\nu_{\text{eff}} = 2\langle X\rangle^2/\mathrm{Var}\,X$. *Operationally:* from N simulations, divide twice the squared mean by the variance; then model X as $\langle X\rangle \chi^2_{\nu_{\text{eff}}}/\nu_{\text{eff}}$. On the full sky $\nu_{\text{eff}} = 2\ell+1$ for $\widehat{C}_\ell$; for a bin of $\Delta\ell$ multipoles on a patch [Hivon (17)] predicts $\nu_\ell = (2\ell+1)\Delta\ell\, f_{\text{sky}} w_2^2/w_4$, which is (8.72) times $\Delta\ell$.

The numbers. On the full sky the simulations give $\nu_{\text{eff}} = 4.7$ at $\ell = 2$ against the exact 5 (the difference is the Monte Carlo error of a variance, ±3.2%, together with the noise and the beam, which are not quite negligible), 20.7 at $\ell = 10$ and 102.3 at $\ell = 50$. Through the 30% cap, 3.8, 17.6 and 93.4; through the 10% cap, 2.1, 9.7 and 57.1. The scaled χ^2 fits where the Gaussian does not. To test the two models fairly, the mean and ν_{eff} are fitted on half of the skies and the test is run on the other half, so that no model is tuned to the very skies it is judged on. The fitted mean and ν_{eff} carry noise of their own, which makes the distance a little larger than the textbook rule expects, so its 5% critical value is found the same way the test is run: draw as many values as there are skies from the model itself, fit on one half, measure the distance on the other, and repeat 300 times. That gives 0.060. The Kolmogorov–Smirnov distance between the test skies and the model is 0.034 for the χ^2 and 0.115 for the Gaussian on the full sky at

$\ell = 2$; 0.022 against 0.117 through the 30% cap; and 0.066 against 0.167 through the 10% cap. At $\ell = 2$ the Gaussian is far beyond the critical value every time, and the χ^2 is much closer. Through the 10% cap even the χ^2 distance grows, to around the critical value or past it: with $\nu_{\rm eff} \approx 2$ the weights are very unequal, and matching two moments is no longer quite enough. By $\ell = 50$ both models pass: their distances are similar and below the critical value (0.050 and 0.044 through the 30% cap, 0.033 and 0.020 through the 10% cap). With this many modes the distribution is close to Gaussian, and both models describe it. The ratio of the simulated skewness to the χ^2 value $\sqrt{8/\nu_{\rm eff}}$ is one to within the Monte Carlo noise at every ℓ and for every mask (Figure 8.32b).

The hidden assumption in the mode count. The plain claim is [Hivon (17)]: a patch supplies $(2\ell + 1) f_{\rm sky} w_2^2 / w_4$ modes per multipole. The everyday question behind it is: how many independent numbers does one pseudo-multipole of the patch contain? The claim answers it as if each pseudo-multipole held only modes of its own ℓ, fewer of them, and nothing from its neighbours. Figure 8.32a puts that to the test, plotting the simulated $\nu_{\rm eff}$ divided by the count. On the full sky the ratio is one (1.00 at high ℓ), and for the holes, which barely couple anything, 1.02. For the caps it is *larger* than one and grows as the patch shrinks: 1.40 for 70%, 2.90 for 30%, 5.21 for 10% and 6.42 for the tapered 10% cap. The band, which couples each multipole to its same-parity neighbours, has 1.36. A single pseudo-multipole of a small cap contains *more* effective modes than the count. Two steps explain the size of the excess and where it goes.

Step one: whose modes are inside $\widetilde{C}_\ell$? The count (8.72) gives each pseudo-multipole the modes of its own ℓ only. But the coupling kernel feeds $\widetilde{C}_\ell$ the power of every true multipole within its width, and each of those brings its own $\approx (2\ell + 1) f_{\rm sky}$ independent modes. For a patch of diameter Θ the kernel has a half-width of about π/Θ (Section 8.10.4). For the 10% cap, $\Theta = 2\theta_c = 1.29$ rad and $\pi/\Theta \approx 2.4$, so the kernel spans about $2 \times 2.4 + 1 \approx 6$ multipoles, and $\widetilde{C}_\ell$ holds roughly five to six times the counted modes, which is the measured 5.21. For the 30% cap, $\Theta = 2.32$ rad gives $2 \times 1.35 + 1 \approx 3.7$, against the measured 2.90: the estimate is rough, since the kernel's edges carry little weight, but it follows the trend.

Step two: the same modes are counted again next door. $\widetilde{C}_{\ell+1}$ draws on a window of about six multipoles shifted by one, so the two windows share about five of their six multipoles, and the two pseudo-multipoles share that fraction of their modes. Two sums that share a fraction r of their independent terms are correlated at about r: here $5/6 \approx 0.8$, close to the measured 0.88 of Section 8.12.2.

So a single pseudo-multipole of a small patch has about five times the counted modes, but they are not its own. Taken at face value, its $\nu_{\rm eff}$ would make an average over neighbouring multipoles look five times less noisy than it is, because the same modes would be counted once in each neighbour. For a bin wide compared with the kernel the sharing averages out and the count is right again. One-line lesson: counting modes is a statement about bins wider than the kernel, never about single multipoles.

What this means for the likelihood is the subject of Chapter 9, but the lesson can be said now, starting from the observation. *The event:* one sky, and one number from it, the quadrupole $\widetilde{C}_2 = x$ through the 10% cap. *What could have produced it?* Any true spectrum: every candidate C_2 would have produced skies with a spread of values, and the observed x is one draw. *How probable is x under each candidate?* The distribution of the estimate, computed for each C_2 and then read at the fixed observed x as a function of C_2, answers that; it is the likelihood (Sections 3.8 and 8.4.1). It measures how readily each candidate spectrum, plus the fluctuation of one sky, turns into the number we actually saw.

Use the scaled χ^2 just found, with mean C_2 and $\nu_{\rm eff} \approx 2$ (we measured 2.1). A χ^2_2 has density

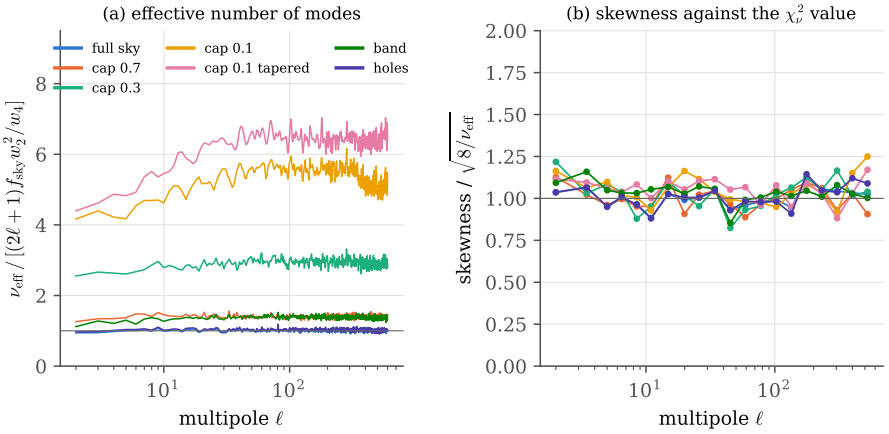


Figure 8.32: (a) The effective number of modes of a single pseudo-multipole from 2000 skies, divided by the count $(2\ell+1) f_{\rm sky} w_2^2/w_4$ of [Hivon (17)]. The smaller the cap, the more each multipole borrows from its neighbours, and the larger the ratio. (b) The simulated skewness divided by the $\chi^2_{\nu_{\rm eff}}$ value $\sqrt{8/\nu_{\rm eff}}$: one for every mask, so the moment-matched χ^2 gets the third moment right too.

$\frac{1}{2} e^{-y/2}$, so $x = C_2\, y/2$ has density, and likelihood,

$$\mathcal{L}(C_2) = \frac{1}{C_2}\, e^{-x/C_2},$$

against the Gaussian of the same mean and variance,

$$\mathcal{L}_{\rm G}(C_2) = \frac{1}{\sqrt{2\pi}\, C_2}\, e^{-(x-C_2)^2/2C_2^2}$$

(a scaled χ^2_ν has variance $2C_2^2/\nu = C_2^2$). Compare a candidate twice the observed value with one equal to it. The exact likelihood gives $\mathcal{L}(2x)/\mathcal{L}(x) = \frac{1}{2} e^{-1/2}/e^{-1} = 0.82$: a spectrum twice as large as what we saw is almost as plausible, because a χ^2_2 often falls well below its mean. The Gaussian gives $\mathcal{L}_{\rm G}(2x)/\mathcal{L}_{\rm G}(x) = \frac{1}{2} e^{-1/8} = 0.44$, half as much. Where does each peak? The exact likelihood at $C_2 = x$. The Gaussian at $C_2 = 0.62\, x$ (setting the derivative of $-\ln C_2 - (x - C_2)^2/2C_2^2$ to zero gives $u^2 - u - 1 = 0$ for $u = x/C_2$, so $u = 1.62$). A Gaussian likelihood of the same width therefore favours spectra that are too low, because it does not know that most skies fall below their mean. Hivon et al. made the same observation for their Boomerang bins and concluded that at small sky coverage the likelihood must be modelled as a χ^2 at low ℓ [Hivon §4.2]; we repeat their test in Section 8.14.1.

8.12.2 Correlations between multipoles

The exact covariance. For a Gaussian sky the covariance of the pseudo-spectrum follows from the covariance (8.57) of the masked coefficients by Wick's theorem. For jointly Gaussian complex variables with zero mean, $\langle x x^* y y^* \rangle = \langle x x^* \rangle \langle y y^* \rangle + \langle x y \rangle \langle x^* y^* \rangle + \langle x y^* \rangle \langle x^* y \rangle$ (the sum over the ways of pairing the four factors, Section 3.5.4). Then

$$\begin{aligned}
\operatorname{Cov}\big(\widetilde{C}_\ell, \widetilde{C}_{\ell'}\big) &\overset{(1)}{=} \frac{1}{(2\ell+1)(2\ell'+1)} \sum_{mm'} \Big(\langle |\tilde{a}_{\ell m}|^2 |\tilde{a}_{\ell' m'}|^2 \rangle - \langle |\tilde{a}_{\ell m}|^2 \rangle \langle |\tilde{a}_{\ell' m'}|^2 \rangle \Big) \\
&\overset{(2)}{=} \frac{1}{(2\ell+1)(2\ell'+1)} \sum_{mm'} \Big(|\langle \tilde{a}_{\ell m} \tilde{a}^*_{\ell' m'} \rangle|^2 + |\langle \tilde{a}_{\ell m} \tilde{a}_{\ell' m'} \rangle|^2 \Big) \\
&\overset{(3)}{=} \frac{2}{(2\ell+1)(2\ell'+1)} \sum_{mm'} |\langle \tilde{a}_{\ell m} \tilde{a}^*_{\ell' m'} \rangle|^2 .
\end{aligned} \tag{8.73}$$

(1) The definition (8.55) of $\widetilde{C}_\ell$ in both factors, and bilinearity of the covariance.
(2) Wick's theorem with $x = \tilde{a}_{\ell m}$, $y = \tilde{a}_{\ell' m'}$: the first pairing cancels the product of means, and $\langle xy^*\rangle\langle x^*y\rangle = |\langle xy^*\rangle|^2$, $\langle xy\rangle\langle x^*y^*\rangle = |\langle xy\rangle|^2$.
(3) The masked map is real, so $\tilde{a}_{\ell' m'} = (-1)^{m'}\tilde{a}^*_{\ell',-m'}$ and $|\langle \tilde{a}_{\ell m}\tilde{a}_{\ell' m'}\rangle| = |\langle \tilde{a}_{\ell m}\tilde{a}^*_{\ell',-m'}\rangle|$; summing over all m', the second sum is the first one relabelled.

This is [134, eqs 14 and 18]. On the full sky, $\langle a_{\ell m}a^*_{\ell' m'}\rangle = C_\ell\delta_{\ell\ell'}\delta_{mm'}$, the sum has $2\ell+1$ equal terms, and the covariance is $2C_\ell^2/(2\ell+1)$ on the diagonal and zero elsewhere, as it must be. Through a mask, the covariance (8.57) connects (ℓ, m) to every (ℓ', m') the window couples, and the spectrum inherits those correlations, squared.

An exact calculation for an azimuthal mask. For a mask that is the same along every ring of constant latitude, like the band, the exact covariance can be computed without any $3j$ symbol, because the window cannot change m. Write $Y_{\ell m} = \lambda_{\ell m}(\theta)e^{im\phi}$ with $\lambda_{\ell m}$ real, and do the integral (8.56) as the HEALPix pixel sum: rings r at colatitude θ_r, each with $n_{\phi,r}$ pixels at $\phi_j = \phi_{0,r} + 2\pi j/n_{\phi,r}$, all of area $\Omega_{\rm pix}$ and window value W_r. Then

$$K_{\ell m,\ell' m'} \overset{(1)}{=} \sum_r \Omega_{\rm pix} W_r\, \lambda_{\ell m}(\theta_r)\lambda_{\ell' m'}(\theta_r)\, e^{i(m'-m)\phi_{0,r}} \sum_{j=0}^{n_{\phi,r}-1} e^{2\pi i(m'-m)j/n_{\phi,r}}$$
$$\overset{(2)}{=} \delta_{mm'}\, K^{(m)}_{\ell\ell'}, \qquad K^{(m)}_{\ell\ell'} \equiv \sum_{\text{rings } r} n_{\phi,r}\, \Omega_{\rm pix}\, W_r\, \lambda_{\ell m}(\theta_r)\lambda_{\ell' m}(\theta_r). \tag{8.74}$$

(1) The pixel sum of $Y^*_{\ell m} W Y_{\ell' m'}$; W and θ are constant along a ring, so only the phase $e^{i(m'-m)\phi_j}$ changes from pixel to pixel.
(2) The sum over j is a geometric series in $q = e^{2\pi i(m'-m)/n_{\phi,r}}$, equal to $(1-q^{n_{\phi,r}})/(1-q) = 0$ when $q \neq 1$, because $q^{n_{\phi,r}} = e^{2\pi i(m'-m)} = 1$; and equal to $n_{\phi,r}$ when $q = 1$, that is $m' = m$ (this needs $|m'-m| < n_{\phi,r}$; it can fail only on the few short rings next to the poles, where $\lambda_{\ell m} \propto \sin^{|m|}\theta$ is negligible for the large $|m|$ concerned).

The window couples ℓ to ℓ' but never m to $m' \neq m$. Two covariances follow. The signal enters through (8.57), with the beam and pixel window on the true spectrum, and K real and diagonal in m: $\sum_L K^{(m)}_{\ell L} T_L^2 C_L K^{(m)}_{\ell' L}$. The noise, white with variance σ^2 per pixel and independent between pixels, is masked pixel by pixel, so the step of (8.14) picks up W_r^2 in each pixel and the same ring sum applies: $N\sum_r n_{\phi,r}\Omega_{\rm pix}W_r^2\lambda_{\ell m}\lambda_{\ell' m}$, with $N = \sigma^2\Omega_{\rm pix}$. Together, $\langle \tilde{a}_{\ell m}\tilde{a}^*_{\ell' m}\rangle = \sum_L K^{(m)}_{\ell L}T_L^2C_LK^{(m)}_{\ell' L} + N\sum_r n_{\phi,r}\Omega_{\rm pix}W_r^2\lambda_{\ell m}\lambda_{\ell' m}$. The functions $\lambda_{\ell m}$ are computed with the standard three-term recursion in ℓ at fixed m, which needs no factorials and is stable to ℓ of several thousand (`lambda_lm` in `lib_masks.py`). Inserting into (8.73) gives the exact covariance of the band's pseudo-spectrum at $\ell, \ell' \le 40$ in seconds. The ratio of the simulated variance to the exact one has mean 1.005 and scatter 0.035 over those multipoles, which is the ±3.2% expected of a variance from 2000 skies: the simulations and the exact formula agree.

What the correlation matrices look like. Figure 8.33 shows the correlation matrix $\mathrm{Corr}(\widetilde{C}_\ell, \widetilde{C}_{\ell'}) = \mathrm{Cov}/\sqrt{\mathrm{Var}\ \mathrm{Var}}$ for $2 \le \ell, \ell' \le 100$ from the simulations, and Figure 8.34 the correlation of each multipole with its first and second neighbours. Three dependences stand out. *On the sky fraction:* the correlation with the next multipole is 0.13 for the 70% cap, 0.65 for the 30% cap and 0.88 for the 10% cap; with the second neighbour, 0.03, 0.14 and 0.60. The width of the correlated band in ℓ grows as the patch shrinks, as π/Θ for a patch of size Θ. *On the geometry:* the band has no correlation with its immediate neighbour (−0.00) but 0.11 with the next-but-one, the parity selection of Section 8.10.3 seen in the covariance; the holes, which remove 3% of the sky

in small pieces, correlate nothing (−0.00). *On apodisation:* the tapered 10% cap is even more correlated than the sharp one, 0.92 and 0.72, because its effective region is smaller and its kernel wider. *On ℓ:* hardly at all; the curves of Figure 8.34 are flat from $\ell = 10$ to 600, because the kernel width is set by the patch, not by the multipole. A correlation estimated from 2000 skies has a noise of $1/\sqrt{2000} = 0.022$ (Section 3.5.2), so every value quoted here is far above noise except those of the holes and the band's first neighbour.

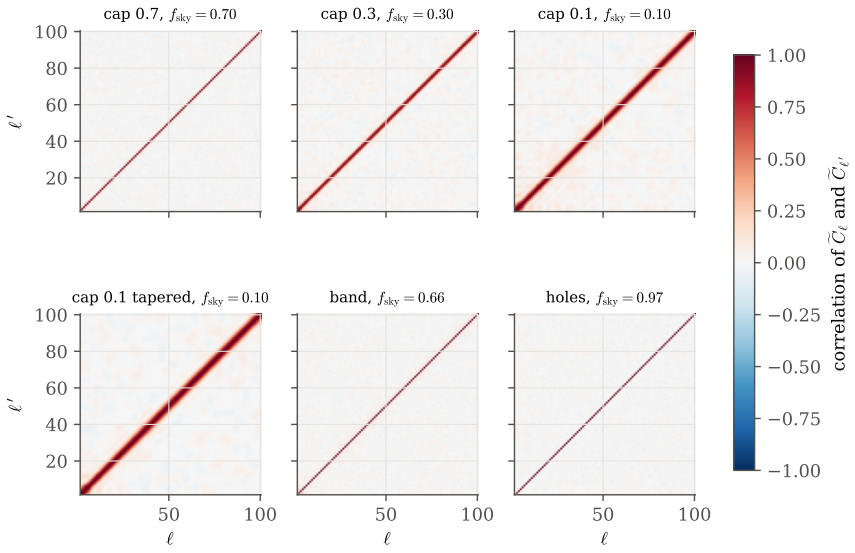


Figure 8.33: Correlation matrices of $\widetilde{C}_\ell$ for $2 \le \ell, \ell' \le 100$ from 2000 skies. The diagonal band widens as the cap shrinks from 70% to 10% and widens further when the 10% cap is tapered. The band cut and the holes leave the matrix nearly diagonal (the band's correlation with $\ell \pm 2$ is too faint to see at this scale).

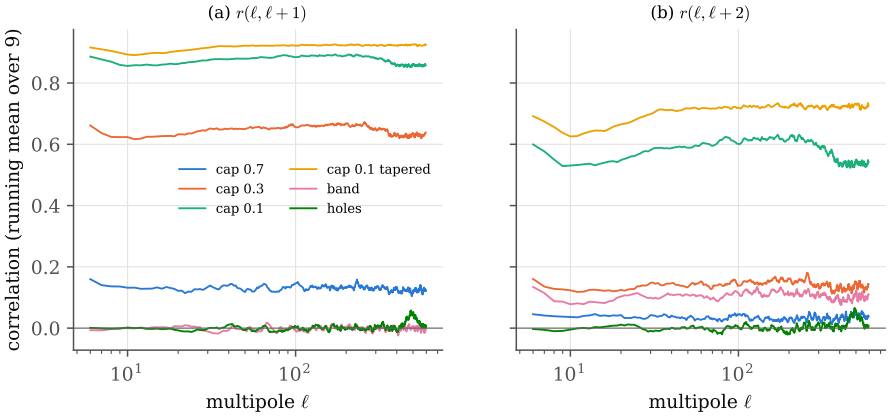


Figure 8.34: Correlation of $\widetilde{C}_\ell$ with $\widetilde{C}_{\ell+1}$ (a) and $\widetilde{C}_{\ell+2}$ (b), smoothed over nine multipoles. The level is set by the mask and is almost independent of ℓ. The band correlates only multipoles of the same parity: zero in (a), nonzero in (b).

8.12.3 Quick formulas against the simulations

Efstathiou's approximation. The exact covariance (8.73) needs every element of the $\tilde{a}$ covariance. There are about $\ell_{\max}^2$ pairs (ℓ, m), so about $\ell_{\max}^4$ elements, and each is a sum (8.57) over another $\ell_{\max}^2$ pairs (ℓ_1, m_1): a cost of order $\ell_{\max}^6$ for a general mask, 2×10^{17} operations at $\ell_{\max} = 767$. Efstathiou [134] found a shortcut for multipoles above the width of the kernel.

The approximation, and its error. In the sum (8.57) for $\langle \tilde{a}_{\ell m} \tilde{a}^*_{\ell' m'} \rangle$, the product $K_{\ell m, \ell_1 m_1} K^*_{\ell' m', \ell_1 m_1}$ is appreciable only for ℓ_1 within the kernel width of both ℓ and ℓ'. Write $C_{\ell_1} = \sqrt{C_\ell C_{\ell'}}\,(1 + \epsilon_{\ell_1})$ there. If the spectrum changes by less than a fraction ϵ across the kernel, every $|\epsilon_{\ell_1}| \lesssim \epsilon$, and dropping them changes the sum by a relative $O(\epsilon)$. For $\ell = \ell'$ the replacement is $C_{\ell_1} \to C_\ell$, exact for a spectrum that is constant across the kernel. Then

$$\langle \tilde{a}_{\ell m} \tilde{a}^*_{\ell' m'} \rangle \overset{(1)}{=} \sum_{\ell_1 m_1} C_{\ell_1} \langle \ell m | \hat{W} | \ell_1 m_1 \rangle \langle \ell_1 m_1 | \hat{W} | \ell' m' \rangle \overset{(2)}{\approx} \sqrt{C_\ell C_{\ell'}}\, \langle \ell m | \hat{W}^2 | \ell' m' \rangle,$$

$$\mathrm{Cov}\big(\widetilde{C}_\ell, \widetilde{C}_{\ell'}\big) \overset{(3)}{\approx} \frac{2 C_\ell C_{\ell'}}{(2\ell+1)(2\ell'+1)} \sum_{m m'} |\langle \ell m | \hat{W}^2 | \ell' m' \rangle|^2 \overset{(4)}{=} 2 C_\ell C_{\ell'}\, \Xi\big(\ell, \ell', \mathcal{W}^{(2)}\big), \tag{8.75}$$

(1) (8.57) with $K_{\ell m, \ell_1 m_1} = \langle \ell m | \hat{W} | \ell_1 m_1 \rangle$ and $K^*_{\ell' m', \ell_1 m_1} = \langle \ell_1 m_1 | \hat{W} | \ell' m' \rangle$ ($\hat{W}$ is real, hence Hermitian).

(2) The approximation, then completeness, $\sum_{\ell_1 m_1} |\ell_1 m_1\rangle\langle \ell_1 m_1| = \mathbb{1}$: two multiplications by W are one multiplication by W^2.

(3) Insert into the exact result (8.73).

(4) The sum over m, m', worked out below.

For step 4, notice that $\langle \ell m | \hat{W}^2 | \ell' m' \rangle$ is the coupling K of (8.56) for a different window, W^2 in place of W. Expand $W^2 = \sum_{LM} w^{(2)}_{LM} Y_{LM}$, with spectrum $\mathcal{W}^{(2)}_L$. Then

$$\sum_{m m'} |\langle \ell m | \hat{W}^2 | \ell' m' \rangle|^2 \overset{(1)}{=} \frac{(2\ell+1)(2\ell'+1)}{4\pi} \sum_{m m'} \Big| \sum_{LM} w^{(2)}_{LM} \sqrt{2L+1} \begin{pmatrix} \ell & L & \ell' \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \ell & L & \ell' \\ -m & M & m' \end{pmatrix} \Big|^2$$

$$\overset{(2)}{=} \frac{(2\ell+1)(2\ell'+1)}{4\pi} \sum_{LM} |w^{(2)}_{LM}|^2 \begin{pmatrix} \ell & \ell' & L \\ 0 & 0 & 0 \end{pmatrix}^2 \overset{(3)}{=} (2\ell+1)(2\ell'+1)\, \Xi\big(\ell, \ell', \mathcal{W}^{(2)}\big),$$

(1) The Gaunt integral (8.58), exactly as in line 2 of (8.60), with $w^{(2)}$ for w; the sign $(-1)^m$ has modulus one and has been dropped.

(2) The collapse (8.59) with $c_{LM} = w^{(2)}_{LM} \sqrt{2L+1}\,(\ell\ L\ \ell'; 0\ 0\ 0)$, and the squared symbol is symmetric in its columns.

(3) $\sum_M |w^{(2)}_{LM}|^2 = (2L+1)\mathcal{W}^{(2)}_L$, and the definition $\Xi(\ell, \ell', \mathcal{W}) \equiv \sum_L \frac{2L+1}{4\pi} \mathcal{W}_L (\ell\ \ell'\ L; 0\ 0\ 0)^2$.

Each element of the approximate covariance is now one sum over L, about $\ell_{\max}$ terms, for each of the $\ell^2_{\max}$ pairs (ℓ, ℓ'): a cost of order $\ell^3_{\max}$, 4.5×10^8 operations at $\ell_{\max} = 767$, against 2×10^{17}. This is [134, eqs 15a–15b]. For a binary mask $W^2 = W$, and the covariance is $2 C_\ell C_{\ell'} G_{\ell\ell'}$ with the symmetric coupling matrix $G = M/(2\ell'+1)$ [134, eq. 16]; for the deconvolved estimate it is $M^{-1} V (M^{-1})^{\mathsf{T}}$ [134, eq. 17], and for MASTER bandpowers the same with $K^{-1}P$. With beam and noise, C_ℓ is the spectrum of the field that was masked, $T^2_\ell C_\ell + N$. A common refinement replaces it by the coupled spectrum $\sum_{\ell'} M_{\ell\ell'} (T^2_{\ell'} C_{\ell'} + N)/f_{\mathrm{sky}} w_2$, which already contains the power that leaked in from elsewhere.

Key idea: Covariance of the cut-sky spectrum

For a Gaussian sky, $\mathrm{Cov}(\widetilde{C}_\ell, \widetilde{C}_{\ell'}) = \frac{2}{(2\ell+1)(2\ell'+1)} \sum_{m m'} |\langle \tilde{a}_{\ell m} \tilde{a}^*_{\ell' m'} \rangle|^2$ exactly. Above the width of the coupling kernel it is approximately $2 C_\ell C_{\ell'}\, \Xi(\ell, \ell', \mathcal{W}^{(2)})$: the full-sky cosmic variance spread over neighbouring multipoles by the spectrum of the *squared* window. Below it, only the exact form or simulations will do.

How good is it? Figure 8.35a divides the approximation's variance of $\widetilde{C}_\ell$ by the simulated one. At high ℓ (100–600) the plain approximation gives 0.976 (70% cap), 0.930 (30%), 0.825 (10%),

0.929 (band) and 0.843 (holes); with the coupled spectrum the same ratios become 1.004, 0.995, 0.970, 0.992 and 0.996. The refinement matters exactly where leakage matters: sharp, small or holey masks at high ℓ, where the power that leaked in from the acoustic peaks carries its own variance. At low ℓ the approximation fails, as it was designed to. At $\ell = 2$ it is off by factors of 3.34 and 6.33 for the 30% and 10% caps, at $\ell = 10$ by 0.82 and 0.62. Against the exact band covariance (Figure 8.35b) the approximation is 1.43 times the truth at $\ell = 2$, 0.83 at $\ell = 4$, 0.90 at $\ell = 20$ and 0.95 at $\ell = 40$. Efstathiou found the same pattern for a $\pm 10^\circ$ cut, accurate at $\ell \gtrsim 20$ and wrong by up to 25% at $\ell = 2$ [134, §2.2]; our $\pm 20^\circ$ cut removes twice as much and the low-ℓ error is correspondingly larger.

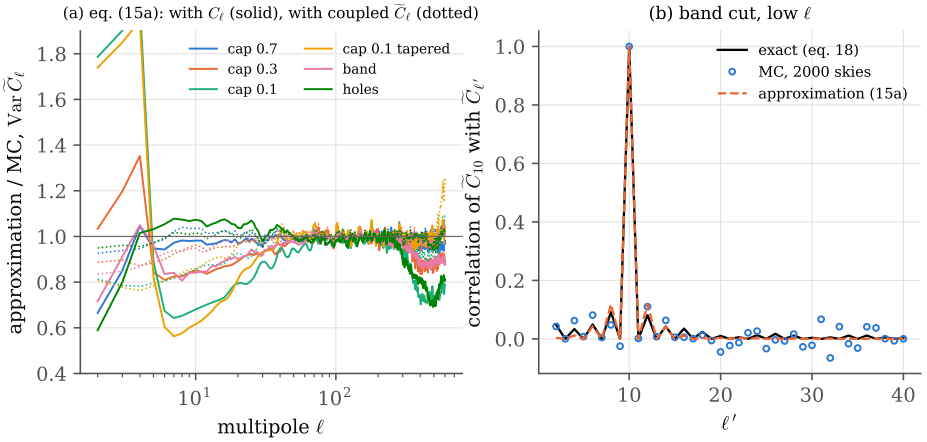


Figure 8.35: (a) Efstathiou's approximation (8.75) for the variance of $\widetilde{C}_\ell$, divided by the simulated variance, for six masks: with the masked spectrum (solid) and with the coupled spectrum (dotted). Above $\ell \approx 30$ it is good to a few per cent with the refinement; at low ℓ it fails by large factors. (b) The band cut: correlation of $\widetilde{C}_{10}$ with $\widetilde{C}_{\ell'}$, exact from (8.74) (black), from 2000 skies (blue circles) and from the approximation (orange dashed). Only multipoles of the same parity as $\ell = 10$ are correlated, and the three agree on the pattern.

Knox's formula, for the MASTER bins. The simplest error bar of all is the mode-counting rule of Section 8.4.2, $\operatorname{Var}\widehat{D}_b = 2(D_b + N_b)^2/[(2\ell_b + 1)\Delta\ell\, f_{\rm sky}]$, with N_b the binned noise on the sky. Figure 8.36 divides the simulated variance of the MASTER bandpowers (bins of 20) by it. On the full sky the ratio is 1.00 at high ℓ and 1.05 in the first bin, the full-sky verification of Section 8.6.2 in another guise. Through the masks it is larger than one almost everywhere and grows as the patch shrinks: at high ℓ, 1.08 for the 70% cap, 1.18 for 30%, 1.40 for 10% and 1.18 for the band. The holes are fine at low ℓ (1.00) and worst at the highest multipoles (1.33), where the leaked power of the peaks adds variance that mode counting knows nothing about. For the tapered cap the version with Hivon's w_2^2/w_4 gives 1.15, while $f_{\rm sky}$ alone gives 1.54, the pitfall of Section 8.9.3 in numbers. The Monte Carlo error of each ratio is 3.2%.

Why is the excess there at all, in a bin of 20 multipoles, many times wider than the kernel? The counting formula is the variance of an *optimal* estimate from the modes the patch contains. MASTER is not optimal: it weights every pixel of the window equally whatever the signal-to-noise, and the deconvolution K^{-1} undoes the smoothing of the coupling. Undoing a smoothing amplifies the scatter, and a two-bin toy shows by how much. The MASTER bandpowers are $\widehat{\boldsymbol{D}} = K^{-1}\boldsymbol{y}$ with $\boldsymbol{y} = P(\widetilde{\boldsymbol{C}} - \langle\widetilde{\boldsymbol{N}}\rangle)$, so their covariance is $K^{-1}\operatorname{Cov}(\boldsymbol{y})(K^{-1})^{\mathsf{T}}$. Take two bins with unit diagonal and coupling c between them, and give the two binned pseudo-spectra equal variance σ^2 and no correlation:

$$K = \begin{pmatrix} 1 & c \\ c & 1 \end{pmatrix}, \quad K^{-1} \overset{(1)}{=} \frac{1}{1-c^2}\begin{pmatrix} 1 & -c \\ -c & 1 \end{pmatrix}, \quad \operatorname{Var}\widehat{D}_1 \overset{(2)}{=} \sigma^2\,\frac{1+c^2}{(1-c^2)^2}.$$

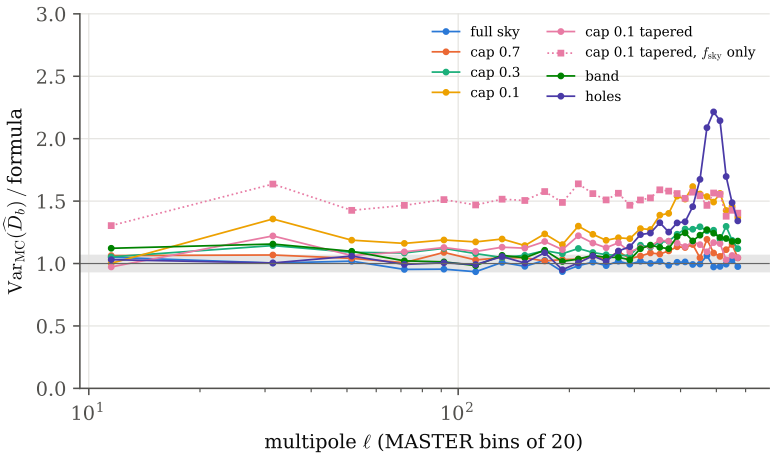


Figure 8.36: Simulated variance of the MASTER bandpowers divided by the mode-counting formula, bins of 20 multipoles. The full sky sits on one; the smaller the patch, the larger the excess. For the tapered cap the formula with $f_{\mathrm{sky}} w_2^2/w_4$ (solid pink) does much better than with f_{sky} alone (dotted pink). The holes rise at the highest multipoles, where leakage dominates.

(1) The inverse of a 2×2 matrix: swap the diagonal, negate the off-diagonal, divide by the determinant $1 - c^2$.
(2) $\widehat{D}_1 = (y_1 - c y_2)/(1 - c^2)$, and the variances of independent terms add: $\sigma^2(1 + c^2)/(1 - c^2)^2$.

The factor exceeds one for every $c \neq 0$ and grows with c: 1.13 for $c = 0.2$, 1.32 for $c = 0.3$. The inverse has a negative off-diagonal entry, the "wing" that subtracts the neighbour's leaked power, and that subtraction brings in the neighbour's scatter too (the long negative wings of K^{-1} in Figure 8.40 are the same thing for a real patch). A wider kernel means larger couplings between bins, so the amplification grows as the patch shrinks. A 40% excess in the variance is an excess of $\sqrt{1.40} = 1.18$, or 18%, in the error bar, the price of an estimator that is fast and unbiased. Efstathiou showed that at high ℓ the pseudo-C_ℓ estimator with suitable weights is nonetheless nearly as good as the optimal quadratic one [134, §§3–5], and that at low ℓ the optimal one should be used instead, the hybrid strategy of real analyses (Section 8.15).

Watch out: Quick covariances at low ℓ

Both the mode-counting formula and Efstathiou's approximation assume that the spectrum is constant across the coupling kernel. At low ℓ, and for any patch smaller than a hemisphere, they are wrong by tens of per cent to factors of several, and the distribution is not Gaussian anyway. Use the exact expression (8.73) (cheap at low resolution), or simulations, for the multipoles below a few times π/Θ.

The laptop version of these tests uses 2000 Gaussian skies at $N_{\mathrm{side}} = 256$, which measure each variance to 3.2% and each correlation to ± 0.022; a research covariance is built from a few hundred end-to-end simulations at full resolution together with an analytic model of the Efstathiou type that includes the noise and the beam for every pair of maps, and the simulations serve to validate and correct the model rather than to replace it [75]. What the reduction costs is resolution ($\ell \leq 767$ instead of 2500) and realism (no foreground residuals, no correlated noise), not the logic of the test.

8.13 How many simulations?

Everything in Section 8.12 came from 2000 simulated skies, and the next step of a real analysis puts the simulated covariance into a likelihood: the probability of the bandpowers $\widehat{\boldsymbol{D}}$ we actually observed, read as a function of the parameters $\boldsymbol{\theta}$. For each $\boldsymbol{\theta}$ the model predicts $\boldsymbol{D}(\boldsymbol{\theta})$, and the question is how probable the fluctuation $\widehat{\boldsymbol{D}} - \boldsymbol{D}(\boldsymbol{\theta})$ is that would turn that prediction into our data (Section 3.8). For Gaussian bandpowers that probability is proportional to $\exp[-\frac{1}{2}(\widehat{\boldsymbol{D}} - \boldsymbol{D})^{\mathsf{T}}\mathbf{C}^{-1}(\widehat{\boldsymbol{D}} - \boldsymbol{D})]$, so $-2\ln\mathcal{L} = (\widehat{\boldsymbol{D}} - \boldsymbol{D})^{\mathsf{T}}\mathbf{C}^{-1}(\widehat{\boldsymbol{D}} - \boldsymbol{D})$ plus a constant, and it needs the inverse of the covariance, the **precision matrix**. Take the MASTER bandpowers of the 30% cap of Section 8.11.3: $p = 29$ bins of 20 multipoles below $\ell = 600$, with neighbouring bins correlated through the mask. With 2000 simulations the covariance is good to a few per cent. A full-resolution end-to-end simulation of a real experiment costs hours, not a second, and a few hundred of them is a typical budget. The question is how few we can afford, and what goes wrong when there are too few. Two different things go wrong, and they need to be kept apart: the elements of the estimated covariance are noisy, and the inverse of a noisy matrix is biased.

8.13.1 Noise in the elements

The sample covariance from N independent Gaussian draws has, element by element, the variance $(\Sigma_{aa}\Sigma_{bb} + \Sigma_{ab}^2)/(N-1)$ of (3.42). On the diagonal that is a relative error $\sqrt{2/(N-1)}$, and for a correlation coefficient near zero an absolute error of about $1/\sqrt{N}$. To see this on the CMB, script `28_nsims.py` draws subsets of N skies from the 2000 and compares each subset's covariance with the one from all of them. Because the reference is itself built from a finite set, the subset fluctuates around it with a variance smaller by the finite-population factor $1 - N/N_{\mathrm{all}}$: a subset that is the whole set does not fluctuate at all. At $N = 100$ the measured rms error is 0.139 on the diagonal, and $\sqrt{2/99} \times \sqrt{1 - 100/2000} = 0.139$; for the correlations it is 0.098, and $\sqrt{1/100} \times \sqrt{0.95} = 0.098$ (Figure 8.37).

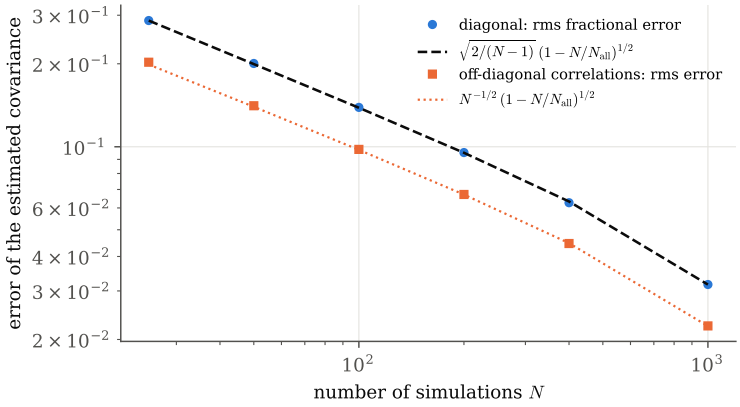


Figure 8.37: Noise in a covariance estimated from N of the 2000 simulated skies, for the 29 MASTER bins of the 30% cap. Blue: rms relative error of the diagonal elements; orange: rms error of the off-diagonal correlation coefficients. The lines are $\sqrt{2/(N-1)}$ and $1/\sqrt{N}$, each with the finite-population factor $(1 - N/N_{\mathrm{all}})^{1/2}$.

Example: can we see the correlations at all? The 70% cap correlates neighbouring multipoles at $r = 0.13$ (Section 8.12.2). To detect that at three standard deviations from a covariance estimated from N skies we need $3/\sqrt{N} < 0.13$, so $N > (3/0.13)^2 \approx 530$. With 100 skies such a correlation is indistinguishable from the spurious ones that every pair of bins shows at the level

of ±0.1. The correlations that matter for a likelihood are often exactly this small, and many of them add up.

8.13.2 Why the inverse is biased

Inverting a matrix is a nonlinear operation, and the average of a nonlinear function of a noisy quantity is not the function of its average. Inversion is not just nonlinear but convex: a fluctuation of the variance towards zero raises its inverse by more than a fluctuation of the same size away from zero lowers it. Jensen's inequality, $\mathbb{E}[1/X] \geq 1/\mathbb{E}[X]$ for a positive random X, makes the inverse of an unbiased covariance biased high. How much is a calculation that starts with one bin.

One bin. With $p = 1$ and $N = n$ draws, the sample variance (mean subtracted) is $s^2 = \sigma^2 \chi^2_k / k$ with $k = n - 1$ degrees of freedom (Chapter 3). Then

$$\mathbb{E}\Big[\frac{1}{s^2}\Big] \overset{(1)}{=} \frac{k}{\sigma^2}\,\mathbb{E}\Big[\frac{1}{\chi^2_k}\Big] \overset{(2)}{=} \frac{k}{\sigma^2}\int_0^\infty \frac{1}{x}\,\frac{x^{k/2-1}e^{-x/2}}{2^{k/2}\Gamma(k/2)}\,\mathrm{d}x \overset{(3)}{=} \frac{k}{\sigma^2}\,\frac{2^{k/2-1}\Gamma(k/2-1)}{2^{k/2}\Gamma(k/2)} \overset{(4)}{=} \frac{k}{k-2}\,\frac{1}{\sigma^2}. \tag{8.76}$$

(1) $1/s^2 = (k/\sigma^2)/\chi^2_k$, and constants come out of the expectation.
(2) The χ^2_k density.
(3) The integrand is $x^{(k/2-1)-1}e^{-x/2}$, a gamma integral equal to $\Gamma(k/2-1)\,2^{k/2-1}$; it converges for $k > 2$.
(4) $\Gamma(k/2) = (k/2-1)\Gamma(k/2-1)$, so the ratio is $\frac{1}{2}/(k/2-1) = 1/(k-2)$.

With $k = n-1$ the bias factor is $(n-1)/(n-3)$: 11% for 20 simulations, harmless for a thousand.

Many bins. For p bins the same mechanism acts in each direction, and the number of directions makes it worse. Collect the $k = n - 1$ mean-subtracted draws (after the rotation of Section 3.5.4, k independent vectors $\boldsymbol{z}_j \sim \mathcal{N}(\boldsymbol{0}, \Sigma)$) into $A = \sum_{j=1}^k \boldsymbol{z}_j \boldsymbol{z}_j^\mathsf{T} = k\widehat{\mathbf{C}}$, a Wishart matrix. Look at one diagonal element of its inverse, say the first. Block inversion of a symmetric matrix gives

$$\frac{1}{(A^{-1})_{11}} \overset{(1)}{=} A_{11} - A_{1r}A_{rr}^{-1}A_{r1} \overset{(2)}{=} \min_{\boldsymbol{\beta}} \sum_{j=1}^{k} \big(z_{1j} - \boldsymbol{\beta}^\mathsf{T}\boldsymbol{z}_{r,j}\big)^2 \;\sim\; \sigma^2_{1\,|\,r}\,\chi^2_{k-(p-1)}, \tag{8.77}$$

(1) The Schur complement: for a matrix split into the first index and the rest (r), the inverse's $(1,1)$ element is one over $A_{11} - A_{1r}A_{rr}^{-1}A_{r1}$.
(2) That expression is the residual sum of squares of the least-squares fit of the first component on the other $p-1$ components across the k draws ($\hat{\boldsymbol{\beta}} = A_{rr}^{-1}A_{r1}$). For Gaussian draws the residual of z_1 given the others is Gaussian with the conditional variance $\sigma^2_{1\,|\,r}$, and a fit with $p-1$ coefficients leaves a χ^2 with $k-(p-1)$ degrees of freedom (Chapter 3); this holds for any values of the regressors, hence also on average over them.

The conditional variance is again a Schur complement, of the true covariance: $\sigma^2_{1\,|\,r} = \Sigma_{11} - \Sigma_{1r}\Sigma_{rr}^{-1}\Sigma_{r1} = 1/(\Sigma^{-1})_{11}$. So $(A^{-1})_{11}$ is $(\Sigma^{-1})_{11}$ divided by a χ^2_{n-p} variable, and (8.76) with $k \to n-p$ gives its mean:

$$\mathbb{E}\big[(\widehat{\mathbf{C}}^{-1})_{11}\big] \overset{(1)}{=} k\;\mathbb{E}\big[(A^{-1})_{11}\big] \overset{(2)}{=} \frac{n-1}{n-p-2}\,(\Sigma^{-1})_{11}. \tag{8.78}$$

(1) $\widehat{\mathbf{C}} = A/k$, so $\widehat{\mathbf{C}}^{-1} = kA^{-1}$, with $k = n-1$.

(2) $\mathbb{E}[1/\chi^2_{n-p}] = 1/(n-p-2)$ from (8.76).

Nothing was special about the first index: in a rotated basis any unit vector $\boldsymbol{u}$ can be the first axis, so $\boldsymbol{u}^\mathsf{T}\,\mathbb{E}[\widehat{\mathbf{C}}^{-1}]\boldsymbol{u} = \frac{n-1}{n-p-2}\boldsymbol{u}^\mathsf{T}\Sigma^{-1}\boldsymbol{u}$ for every $\boldsymbol{u}$, and two symmetric matrices with equal quadratic forms are equal. The whole precision matrix is biased by the same factor [6, eq. 16].

Key idea: The Hartlap factor, and why it is there

For a covariance estimated from n Gaussian simulations of p bins ($p < n-2$),

$$\mathbb{E}\big[\widehat{\mathbf{C}}^{-1}\big] = \frac{n-1}{n-p-2}\,\mathbf{C}^{-1}, \qquad \widehat{\mathbf{C}^{-1}} = \frac{n-p-2}{n-1}\,\widehat{\mathbf{C}}^{-1} \text{ is unbiased}$$

[6, eq. 17]. Each direction of the inverse is one over a residual variance, and fitting the other $p-1$ directions costs $p-1$ degrees of freedom: the noisy matrix "explains away" part of every fluctuation, its small eigenvalues come out too small, and the inverse too large.

This is the factor quoted in (3.43); there it was stated, here it has been derived. Two more facts complete the picture [6, §§2.2, 4]. For $p > n-1$ the sample covariance is singular, because $n-1$ mean-subtracted vectors span at most $n-1$ dimensions, and no factor can rescue it. And the sample mean and the sample covariance of Gaussian draws are independent, so a χ^2 built from both is unbiased once the Hartlap factor is applied; a marginalised likelihood and a Fisher matrix are still biased by smaller, model-dependent amounts (about 8% in the likelihood, and up to 30% in $\sqrt{\det \mathbf{F}^{-1}}$ as $p/n \to 1$, in their examples).

Example: what forty bins and sixty simulations do. For $n = 60$ and $p = 40$ the factor is $(60-40-2)/(60-1) = 0.305$: without the correction every χ^2 is three times too large, and every error bar $\sqrt{3.3} = 1.8$ times too small.

8.13.3 Testing the Hartlap factor

Hartlap et al.'s experiment, repeated. Their test [6, §3.2, Fig. 1] draws $n = 60$ Gaussian vectors of $p_1 = 240$ bins, from three population covariances (constant diagonal, linearly falling diagonal, and a non-diagonal $\sigma^2/(1+\epsilon|i-j|)$ with $\epsilon = 0.05$, their eqs 18–20), rebins them to $p = 240/j$ bins by joining j neighbours, and compares the trace of the true precision matrix with the trace of the estimated one. We repeat it with 2000 repetitions instead of their 10^4 (Figure 8.38a). At $p = 40$ the raw ratio is 0.307 against the predicted $(n-p-2)/(n-1) = 0.305$, and after the correction every model and every p lies within 0.005 of one.

The same on the CMB. The sky is not a Gaussian vector of bandpowers: the bandpowers are quadratic in a Gaussian field and, at low ℓ, skewed. To see whether the factor still holds we take subsets of n skies from the 2000, invert the covariance of their $p = 29$ MASTER bins and compare with the precision matrix from all of them (Figure 8.38b). At $n = 60$ the trace ratio is 0.49, and the Hartlap factor is 0.49. The reference itself carries the factor $(2000-29-2)/(2000-1) = 0.985$, a 1.5% correction even with two thousand simulations.

8.13.4 From an unbiased precision matrix to a trustworthy error bar

An unbiased precision matrix is still a noisy one, and its noise reaches the parameters. Taylor, Joachimi, and Kitching [165] worked out how. Their result for the fractional scatter of a parameter's variance, estimated with a Hartlap-corrected precision matrix from N_S simulations

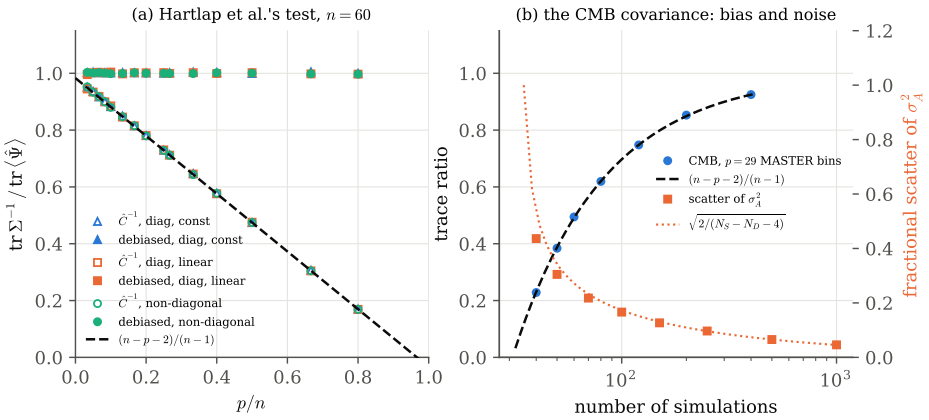


Figure 8.38: (a) Hartlap et al.'s test: the trace of the true precision matrix divided by that of the estimated one, for $n = 60$ and p up to 60, three covariance models (shapes), raw (open) and corrected (filled). The raw points follow $(n - p - 2)/(n - 1)$ (dashed) down to zero; the corrected ones sit at one. (b) The CMB: the same ratio for the MASTER bins of the 30% cap from subsets of n skies (blue, left axis), on the Hartlap curve; and the scatter of the fitted amplitude's variance (orange, right axis) against Taylor et al.'s $\sqrt{2/(N_S - N_D - 4)}$ (dotted).

of N_D data points, is $\varepsilon = \sqrt{2/(N_S - N_D - 4)}$ [165, eq. 55], and it leads to a planning rule:

$$N_S > \frac{2}{\varepsilon^2} + N_D + 4 \qquad \text{[165, eq. 58].} \tag{8.79}$$

The first term is the familiar $1/\sqrt{N}$ of any average; the second says that each data point costs one more simulation, the same degrees of freedom that appeared in the Hartlap factor. For a parameter error good to 5% the variance must be good to 10%, $\varepsilon = 0.1$, and (8.79) asks for $200 + N_D + 4$ simulations: 233 for our 29 bins.

The simplest case, exactly. For one parameter the calculation of (8.77) gives the whole distribution. Fit an amplitude A of the fiducial spectrum, $\boldsymbol{D} = A\boldsymbol{t}$ with $\boldsymbol{t}$ the fiducial bandpowers.

Its error bar. With a precision matrix Ψ the likelihood is quadratic in A: $-2\ln\mathcal{L}(A) = A^2\,\boldsymbol{t}^\mathsf{T}\Psi\boldsymbol{t} - 2A\,\boldsymbol{t}^\mathsf{T}\Psi\widehat{\boldsymbol{D}} + \text{const}$. A Gaussian in A of width σ_A has $-2\ln\mathcal{L} = (A - \hat{A})^2/\sigma_A^2 + \text{const}$, and matching the coefficients of A^2 gives $\sigma_A^2 = 1/(\boldsymbol{t}^\mathsf{T}\Psi\boldsymbol{t})$, one over the Fisher information (Chapter 10; the full chain is in Section 8.15.2). With the true covariance, $\sigma_A^2 = 1/(\boldsymbol{t}^\mathsf{T}\mathbf{C}^{-1}\boldsymbol{t})$.

Its distribution. Whiten the data: $\mathbf{C}^{-1/2}$ turns the simulated vectors into draws with identity covariance, and their sample covariance into $\widehat{S} = \mathbf{C}^{-1/2}\widehat{\mathbf{C}}\,\mathbf{C}^{-1/2}$. With $\boldsymbol{s} = \mathbf{C}^{-1/2}\boldsymbol{t}$ and the unit vector $\boldsymbol{u} = \boldsymbol{s}/|\boldsymbol{s}|$,

$$\boldsymbol{t}^\mathsf{T}\widehat{\mathbf{C}}^{-1}\boldsymbol{t} \overset{(1)}{=} \boldsymbol{s}^\mathsf{T}\widehat{S}^{-1}\boldsymbol{s} \overset{(2)}{=} |\boldsymbol{s}|^2\,(\widehat{S}'^{-1})_{11} \overset{(3)}{=} \frac{1}{\sigma_A^2}\,\frac{n-1}{\chi^2_{n-p}}.$$

(1) $\widehat{\mathbf{C}}^{-1} = \mathbf{C}^{-1/2}\widehat{S}^{-1}\mathbf{C}^{-1/2}$.

(2) Rotate the basis so that $\boldsymbol{u}$ is the first axis; whitened Gaussian draws stay independent standard Gaussians under a rotation, and $\widehat{S}'$ is their sample covariance in the new basis.

(3) $|\boldsymbol{s}|^2 = \boldsymbol{t}^\mathsf{T}\mathbf{C}^{-1}\boldsymbol{t} = 1/\sigma_A^2$. For $\widehat{S}' = A/(n-1)$, (8.77) with $\Sigma = \mathbb{1}$ (so $\sigma^2_{1\,|\,r} = 1$) says $1/(A^{-1})_{11} \sim \chi^2_{n-p}$, hence $(\widehat{S}'^{-1})_{11} = (n-1)/\chi^2_{n-p}$.

Multiplying the inverse by the Hartlap factor, $\Psi = \frac{n-p-2}{n-1}\widehat{\mathbf{C}}^{-1}$, gives

$$\widehat{\sigma_A^2} = \frac{n-1}{n-p-2}\,\frac{1}{\boldsymbol{t}^{\mathsf{T}}\widehat{\mathbf{C}}^{-1}\boldsymbol{t}} = \sigma_A^2\,\frac{\chi^2_{n-p}}{n-p-2}, \qquad \frac{\mathbb{E}\,\widehat{\sigma_A^2}}{\sigma_A^2} = \frac{n-p}{n-p-2}, \qquad \frac{\sigma\big(\widehat{\sigma_A^2}\big)}{\mathbb{E}\,\widehat{\sigma_A^2}} = \sqrt{\frac{2}{n-p}}. \tag{8.80}$$

Even with the unbiased precision matrix, the parameter's variance is biased *high*, for the same convexity reason as before: σ_A^2 is one over a linear function of Ψ. The test of `28_nsims.py` draws n Gaussian bandpower vectors from the CMB covariance ($\sigma_A = 0.0069$ with the true precision matrix), 2000 times for each n. At $n = 40$ the mean of $\widehat{\sigma_A^2}$ is high by 22%, and (8.80) gives $(40-29)/(40-31) = 1.22$; its fractional scatter is 0.44, against $\sqrt{2/11} = 0.43$ from (8.80) and 0.53 from Taylor et al.'s general formula. At $n = 100$ the scatter is 0.166, against $\sqrt{2/71} = 0.168$ and 0.173. For a single parameter (8.80) is exact.

What the "−4" covers. The same calculation shows where Taylor et al.'s formula sits. Their expression is the relative scatter of the *information* $\boldsymbol{t}^{\mathsf{T}}\Psi\boldsymbol{t} = 1/\widehat{\sigma_A^2}$, which is proportional to $1/\chi^2_k$ with $k = n - p$. From the gamma integral of (8.76), $\mathbb{E}[1/\chi^2_k] = 1/(k-2)$ and $\mathbb{E}[1/(\chi^2_k)^2] = 1/[(k-2)(k-4)]$, so

$$\mathrm{Var}\,\frac{1}{\chi^2_k} = \frac{1}{(k-2)(k-4)} - \frac{1}{(k-2)^2} = \frac{2}{(k-2)^2(k-4)}, \qquad \frac{\sigma(1/\chi^2_k)}{\mathbb{E}(1/\chi^2_k)} = \sqrt{\frac{2}{k-4}} = \sqrt{\frac{2}{N_S - N_D - 4}}.$$

The variance itself is proportional to χ^2_k, whose relative scatter is $\sqrt{2/k}$. The two agree when k is large, because a small relative error of a quantity is also the relative error of its inverse; at $k = 11$ the inverse χ^2, with its longer tail, scatters more. The "−4" is that tail, and a planning rule built on it errs on the safe side.

Watch out: Too few simulations cannot be repaired

When $p \geq n-1$ the sample covariance is singular. A pseudo-inverse obtained by discarding small singular values produces a finite matrix, but its bias depends on the covariance model and is not removed by any factor [6, §3.2]. Plan the simulations so that $n - p$ is comfortably larger than $2/\varepsilon^2$, compress the data into fewer bins, or model the covariance analytically and use simulations only to test the model.

What our numbers are worth. Our covariance uses 2000 skies for 29 bins. The Hartlap factor is 0.985, and by (8.80) the variance of a single parameter is uncertain by $\sqrt{2/1971} = 3.2\%$, its error bar by 1.6%. A research analysis has larger data vectors and dearer simulations. Planck's 300 end-to-end simulations [75] would give, for a vector of a few hundred bandpowers, an $n - p$ of order one hundred and a 14% uncertainty on each parameter variance. That is why the high-ℓ likelihoods of Planck and of the ground-based experiments use analytic covariances of the Efstathiou type, with the simulations used to test and correct them: the cost of a reduction in simulations is paid in the precision of the error bars, and the cure is to need fewer simulations, not to pretend that a small number is enough.

8.14 Reproducing the cut-sky literature

The full-sky content of Knox (1995) and Hivon et al. (2002) was reproduced in Section 8.7. The papers behind this part of the chapter make claims about masked skies, and each of them can be put to the same test: rebuild the experiment with our own code, at laptop scale, and set our numbers beside theirs. Hivon et al. tested MASTER on a simulated balloon survey of a

small patch. Efstathiou derived and tested an approximation for the covariance. Hartlap et al. derived and tested the bias of an inverted covariance, and Taylor et al. its consequences for error bars. The last three were reproduced along the way in Sections 8.12.3, 8.13.3 and 8.13.4; the first needs its own experiment.

8.14.1 Hivon et al. (2002): the Boomerang test

Their question. A balloon such as Boomerang maps a few per cent of the sky in one flight. Its detectors drift slowly, with a noise that rises as $1/f$ at low frequencies, so the timestreams are high-pass filtered before the map is made, and the filter removes large-scale power in a way that depends on how the sky was scanned. Neither the filter's effect on the spectrum nor the detector noise, once projected onto a map, has a formula. Hivon et al. asked whether the pseudo-spectrum recipe of Section 8.11, with every unknown measured on simulations, still delivers what a likelihood needs: an unbiased spectrum, an error bar that can be trusted, and the right shape of distribution. A small patch with a filter is the hard case on every count. The coupling kernel is widest for a small patch (Section 8.10.4), the filter acts on the large scales that a small patch barely resolves, and few modes per multipole make the distribution far from Gaussian (Section 8.12.1).

Their setup, one item at a time.

- *The patch and its weighting.* They analysed an ellipse inside the scanned region, under three weightings: a top-hat, a cosine taper and a Gaussian. A softer weighting lowers the leakage through the edge and pays for it in statistical weight, the trade of Section 8.10.4 and of the factor w_2^2/w_4 of Section 8.9.2. We keep the ellipse and the three weightings.

- *The filter.* Detector drifts are slow in time. Along a scan, slow in time means large on the sky, so the high-pass filter removes large-scale structure along the scan direction. We keep a sharp filter but idealise the scans.

- *The noise.* One detector's noise has a white level, a $1/f$ rise below a knee frequency, and lines at the harmonics of the scan. In the map the white part becomes pixel noise, set by how long each pixel was observed. We keep only the white part, at the same average depth.

- *The beam and the pixels.* Both kept.

- *The simulations.* Three sets with three jobs: noise-only skies measure the noise pseudo-spectrum $\langle \widetilde{N}_\ell \rangle$, and from it the noise on the sky; noise-free skies measure the transfer function F; signal-plus-noise skies test the whole recipe. We keep their sizes.

The values are in Table 8.7. The full version simulates timestreams with the real pointing and coloured noise and makes maps from them, which cost them half a CPU-day per 300 simulations [Hivon (33)]. What our reduction costs is the realism of the filter and of the noise: our transfer function is that of an ideal scan, and our noise has no stripes.

Two conversions behind our setup. *From scan frequency to azimuthal wavenumber.* Why does a filter in time remove a set of m? Put the patch on the equator of our coordinates and let the telescope sweep along a ring of constant colatitude θ at a speed v on the sky (radians of arc per second). The ring has radius $\sin\theta$, so the azimuth advances as $\phi = vt/\sin\theta$. A

	Hivon et al. (2002)	ours
data	one detector, about 5×10^7 samples, 4.4% of the sky scanned	maps simulated directly
analysis region	ellipse of semi-axes $20^\circ \times 12^\circ$, $f_{\rm sky} = 1.82\%$	the same, $f_{\rm sky} = 1.82\%$
weightings	top-hat, cosine, Gaussian	the same
beam, pixels	10′; $N_{\rm side} = 512$ (7′)	10′; $N_{\rm side} = 512$
filter	sharp high-pass at 100 mHz on the timestream, scans at $1^\circ/$s in azimuth	modes with $\lvert m\rvert < 50$ removed (ideal constant-latitude scans)
noise	$130\,\mu$K s$^{1/2}$ white, $1/f$ below a knee near 100 mHz, scan lines; about 500 s deg^{-2}	white, uniform, $350\,\mu$K arcmin
simulations	450 noise-only, 250 noise-free, 1350 signal plus noise	450, 252, 1350
cost	half a CPU-day per 300 simulations	31 minutes for all

Table 8.7: The Boomerang test [Hivon §4.1, Table 1] and our laptop version (`29_boomerang_mc.py`).

harmonic carries the factor $e^{im\phi}$, which along the scan becomes $e^{imvt/\sin\theta}$, an oscillation in time of frequency

$$\nu = \frac{m\,v}{2\pi \sin\theta}.$$

A high-pass filter at ν_c therefore removes every harmonic with $|m| < 2\pi\nu_c \sin\theta/v$, whatever its ℓ; near the equator $\sin\theta \approx 1$. Boomerang swung in azimuth at a rate $v_{\rm az}$ at an elevation $\theta_{\rm el}$. The azimuth circle at that elevation has radius $\cos\theta_{\rm el}$, so the speed on the sky is $v = v_{\rm az}\cos\theta_{\rm el}$, and the cut is $\ell_c = 2\pi\nu_c/(v_{\rm az}\cos\theta_{\rm el})$ [Hivon App. B]. With $\nu_c = 0.1$ Hz and $v_{\rm az} = 1^\circ/\text{s} = 0.0175$ rad/s, $2\pi\nu_c/v_{\rm az} = 0.628/0.0175 = 36$, and an elevation near 45° ($\cos\theta_{\rm el} \approx 0.7$) raises it to about 50. We take $\ell_c = 50$.

From detector noise to map depth. A white level of $130\,\mu$K s$^{1/2}$ means that averaging the timestream for t seconds leaves an error of $130/\sqrt{t}\,\mu$K. A pixel of area A (in deg^2) is observed for $t = 500\,\text{s deg}^{-2} \times A$ on average, so its noise is $\sigma_{\rm pix} = 130/\sqrt{500A}\,\mu$K. The product $\sigma_{\rm pix}\sqrt{A}$ no longer depends on the pixel size: $130/\sqrt{500} = 5.8\,\mu$K deg, or, with 60 arcmin to the degree, $350\,\mu$K arcmin, our depth of $350\,\mu$K arcmin.

Their key equations, and our symbols. Their estimator is MASTER, derived in Section 8.11.3 from the master equation of Section 8.10; their eqs. (15) and (24)–(26) are our master equation and the workdef of Section 8.11.3. Their flat-band power C_b is our D_b, their beam window B_ℓ our T_ℓ (beam times pixel window), and their $M_{\ell\ell'}$, F_ℓ, $P_{b\ell}$, $Q_{\ell b}$, $K_{bb'}$, $\langle\widetilde{N}_\ell\rangle_{\rm MC}$ and noise on the sky $\widehat{\mathcal{N}}_b$ are ours. Two more equations are compared below: their error bar (35)–(36), derived next, and the transfer function of a parallel scan (B7) and (B15), derived with the transfer test.

Their error bar, derived. How many independent modes does one bin of a filtered, weighted patch hold? Start from the full sky and remove what the observation removes, one factor at a

time:

$$\operatorname{Var}\widehat{D}_b \overset{(1)}{=} \frac{2(D_b+N_b)^2}{(2\ell_b+1)\Delta\ell} \overset{(2)}{=} \frac{2(D_b+N_b)^2}{(2\ell_b+1)\Delta\ell\, f_{\rm sky}\, w_2^2/w_4} \overset{(3)}{=} \frac{2(D_b+N_b)^2}{\nu_b^f}, \qquad \nu_b^f = (2\ell_b+1)\,\Delta\ell\, f_{\rm sky}\,\frac{w_2^2}{w_4}\,F_b. \tag{8.81}$$

(1) The full-sky variance of a bin of $\Delta\ell$ multipoles centred on ℓ_b, (8.41): $2/(2\ell+1)$ per multipole, $\Delta\ell$ independent multipoles. It is written for D instead of C, since the factor $\ell(\ell+1)/2\pi$ is nearly constant across a bin; N_b is the binned noise on the sky.
(2) A weighted patch keeps $(2\ell+1)f_{\rm sky}w_2^2/w_4$ effective modes per multipole, (8.72).
(3) A sharp filter removes the modes whose along-scan frequency is below the cut and leaves the others untouched, and F_b is the fraction of modes it keeps ((8.82) below makes this explicit). Deconvolving by F_b restores the amplitude of the spectrum, but it is now estimated from fewer modes.

The square root is their analytic error bar, $\Delta\widehat{C}_b = (C_b + \widehat{\mathcal{N}}_b)\sqrt{2/\nu_b^f}$ [Hivon (36)], with ν_b^f their eq. (35). By the moment matching of (8.71), the same ν_b^f is the number of degrees of freedom of the χ^2 they compare with the histograms. In our language ν_b^f is the $\nu_{\rm eff}$ of a bin, as predicted by counting modes; $f_{\rm sky}$, w_i, F_b and $\Delta\ell$ are the same symbols. Step 2 carries the assumption tested in Section 8.12.1, that the bin is much wider than the coupling kernel. Here it is: the ellipse is at least $24^\circ = 0.42$ rad across, so the kernel is about $\pi/0.42 \approx 7.5$ multipoles wide, against bins of 50.

What we compute. `29_boomerang_mc.py` makes the three sets of skies: Gaussian skies of the fiducial spectrum, smoothed by the beam and the pixel window, filtered by removing the harmonics with $|m| < \ell_c$, plus white noise of the depth above, generated in harmonic space and filtered like the sky; it stores their pseudo-spectra through the three weightings. `30_boomerang.py` is MASTER exactly as in Section 8.11.3: bins of 50 centred on 50, 100, . . . , 1250; F_b solved in bins from the noise-free skies, with the binned system of Section 8.11.4; $\langle\widetilde{N}_\ell\rangle$ from the noise-only skies; then $K_{bb'}$, its inverse, and $\widehat{D}_b$ for every signal-plus-noise sky. The mean, scatter, correlations and histograms of those estimates are the comparisons below, and the numbers side by side are collected in Table 8.8.

The window factors (their Table 1). These depend only on the geometry. Ours, $w_2^2/w_4 = 1.000$, 0.514 and 0.223 for the top-hat, cosine and Gaussian weightings, agree with their 1, 0.514 and 0.223, and our sky fraction is 1.82% like theirs.

The transfer function (their Fig. 2 and App. B). What fraction of the power at multipole ℓ survives a sharp high-pass along parallel scans? *One plane wave first.* On a patch this small the sky is nearly flat, and a mode of multipole ℓ is a plane wave $e^{i\boldsymbol{k}\cdot\boldsymbol{x}}$ with $|\boldsymbol{k}| = \ell$. Let the scans run along x, and let the wave vector make an angle θ with them. Along a scan the wave is $e^{i\ell\cos\theta\, x}$: the scan sees only the along-scan wavenumber $\ell\cos\theta$, and since x grows in proportion to time, a time frequency proportional to $\ell\cos\theta$ (the flat-sky version of the m of the conversion above). The filter multiplies the wave's amplitude by its response $f(\ell\cos\theta)$, and its power by $f(\ell\cos\theta)^2$. *Then all directions.* An isotropic sky has waves of every direction with equal power, so the fraction of the power kept at ℓ is the average over θ, $F_\ell^{(0)} = \frac{1}{2\pi}\int_0^{2\pi} \mathrm{d}\theta\, f(\ell\cos\theta)^2$ [Hivon (B7)]. For a sharp cut, $f = 0$ below ℓ_c and 1 above,

$$F_\ell^{(0)} \overset{(1)}{=} 1 - \frac{1}{2\pi}\left|\{\theta : |\cos\theta| < \ell_c/\ell\}\right| \overset{(2)}{=} 1 - \frac{4\arcsin(\ell_c/\ell)}{2\pi} = 1 - \frac{2}{\pi}\arcsin\frac{\ell_c}{\ell} \qquad (\ell \geq \ell_c), \tag{8.82}$$

(1) The filter keeps a wave unless its along-scan frequency is below the cut; F is the fraction of directions kept.
(2) $|\cos\theta| < x$ holds within $\arcsin x$ of $\theta = \pi/2$ and of $3\pi/2$, on both sides: four arcs of length $\arcsin x$.

and $F^{(0)} = 0$ below ℓ_c [Hivon (B14)–(B15)]. The paper does not tabulate F, but it follows from their published degrees of freedom by solving (8.81) for F_b with the top-hat values. Our simulated F_b follows (8.82) closely; theirs agrees with both at $\ell = 50$ and lies about 10% lower above it (Table 8.8, Figure 8.39b). *Why they differ.* Their own Fig. 2 shows the same gap between the measured curves and the dotted toy model: real scans are not parallel lines, they cross the patch at varying angles and speeds, and the filter removes more. Our simulation implements the toy filter exactly, so it reproduces the toy curve and not their measurement.

How well F is known. Their eq. (38), $\delta F/F \approx a\,10^{-2}\sqrt{100/N}\sqrt{400/\ell}$, is the counting rule of Section 8.11.4 in their units. With $2\ell + 1 \approx 2\ell$,

$$\frac{\delta F}{F} = \sqrt{\frac{2}{2\ell\,\Delta\ell\,f_{\rm sky}N}} = \sqrt{\frac{400}{\ell}}\sqrt{\frac{100}{N}}\,\frac{1}{\sqrt{400\times100\times\Delta\ell\,f_{\rm sky}}},$$

and with $\Delta\ell = 50$ and $f_{\rm sky} = 0.0182$ the last factor is $1/\sqrt{36\,400} = 0.52\times10^{-2}$: counting predicts $a = 0.52$. Our 252 noise-free skies, split into their six chunks, give $a = 0.63 \pm 0.04$: a little above the counting value, and at the low end of their 0.6–0.9. Counting assumes each bin is measured on its own; the correction for the mask couples neighbouring bins, and that adds scatter, as the ring-filter test of Section 8.11.4 also found.

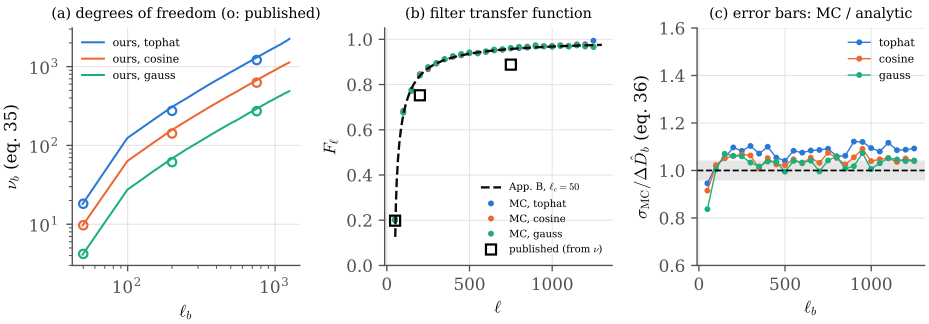


Figure 8.39: Our Boomerang test against the published numbers. (a) Degrees of freedom ν_b of [Hivon (35)] for the three weightings (lines) and the nine published values of their Fig. 6 (circles). (b) The transfer function: our simulated F_b (dots), the parallel-scan formula (8.82) (dashed), and the values implied by the published ν (squares). (c) The simulated error bars divided by the analytic ones of [Hivon (36)]; the grey band is the Monte Carlo error.

The degrees of freedom (their Fig. 6). Inserting our F into (8.81) gives our ν for the three weightings at $\ell = 50$, 200 and 750 (Figure 8.39a, Table 8.8). Above $\ell = 50$ ours are larger, and the difference is the ratio of the two transfer functions and nothing else: $307.2/274.6 = 1.12 = 0.846/0.75$. A sharper check of their table is internal. In (8.81) only w_2^2/w_4 changes between weightings, so multiplying their top-hat ν by our ratios of w_2^2/w_4 should give their cosine and Gaussian values. It does, to within 1%.

The binned kernel (their Fig. 3). Figure 8.40 shows the diagonal and two rows of $K_{bb'}$ and of its inverse for the top-hat window, in the format of their figure. The diagonal is 0.012 at $\ell_b = 200$ and falls to 1.0×10^{-3} at 1200, the beam, pixel window and $f_{\rm sky}w_2$ in one number; the

rows fall by four to six decades within a few hundred multipoles, and the inverse has the long negative wings of a deconvolution, as in their figure.

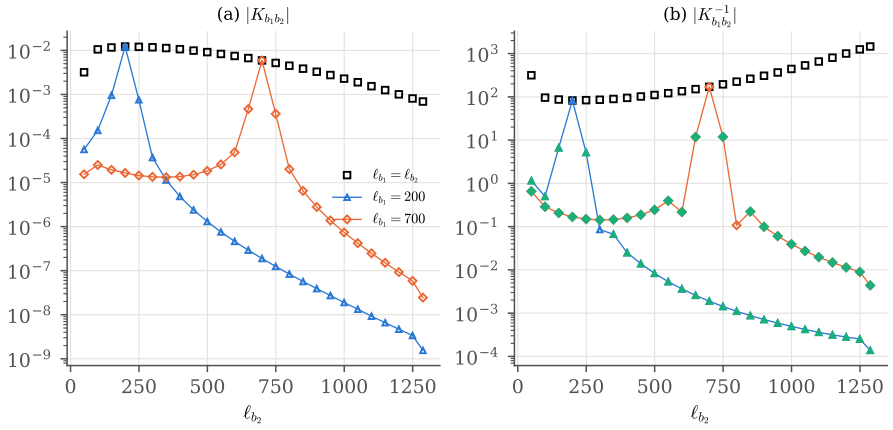


Figure 8.40: The binned coupling kernel of the top-hat ellipse ($f_{\rm sky} = 1.8\%$, $\Delta\ell = 50$), in the format of Hivon et al.'s Fig. 3. (a) $|K_{b_1 b_2}|$: the diagonal (squares) and the rows $\ell_{b_1} = 200$ and 700. (b) $|K^{-1}_{b_1 b_2}|$; green marks negative elements.

The recovered spectrum and its error bars (their Fig. 5): what they claimed. They found the mean of 1352 MASTER estimates "in perfect agreement (to an error on the mean)" with the bin-averaged input, and the simulated error bars "almost identical" to the analytic error bar (8.81) [Hivon §4.2]; off-diagonal correlations were "at most 10%".

How we test it. For the error bar we divide the simulated scatter of $\widehat{D}_b$ by the square root of (8.81). For the mean we use *pulls*, as in Section 8.10.3: the mean of our estimates in a bin, minus its prediction, divided by the Monte Carlo error of that mean (the scatter of single estimates over $\sqrt{1350}$). If the prediction is right, the pulls are draws of a standard Gaussian, and the largest of 25 should be about two, rarely above three. Two predictions are possible: the plain bin average of the input, as in the paper, or the bandpower-window prediction (8.68). The numbers for the three weightings are in Table 8.8 and Figure 8.39c, Figure 8.41. In short: the analytic error bar is a few per cent optimistic in most bins and pessimistic in the first; the mean agrees with the window prediction, with pulls larger than chance until the errors shared by all skies are counted; against the plain bin average it is off by up to a third of a single sky's error bar; and the top-hat's largest correlation exceeds 10%.

Why they differ. *The error bar is a few per cent too small.* Equation (8.81) counts modes, which gives the variance of an optimal estimate. MASTER is not optimal, and its deconvolution amplifies the scatter (Section 8.12.3); a 5% excess, against a Monte Carlo error of 2%, is invisible at the precision of "almost identical".

In the first bin it is too large. The bin $\ell = 25$–74 contains a steep rise of both D_ℓ and F_ℓ, and the rule evaluates them at the bin's centre. That is step 3 of (8.41), which failed in the full sky's first bin for the same reason (Section 8.7.2).

The pulls against the windows are larger than chance. All 1350 estimates share two things measured from smaller sets. They are deconvolved with the same F, measured from only 252 skies. Since $\widehat{D}_b \propto 1/F_b$, a relative error δ of F_b shifts every estimate by the same $-\delta D_b$. And the same noise power, measured from only 450 noise-only skies, is subtracted from every one of them, so its error shifts them all alike too. Averaging over 1350 skies shrinks their independent

scatter by $\sqrt{1350}$ but leaves the common shifts untouched, so in units of the Monte Carlo error of the mean they are magnified. For the top-hat, the scatter of F between the six chunks of the noise-free set gives common shifts of up to 3.1 such errors, and the noise power up to 1.7, largest in the noise-dominated bins at high ℓ. Adding both to the error of the mean, the largest pulls become 1.6, 1.5 and 2.0 for the three weightings: about what chance gives, since the largest of 25 standard normal numbers is typically about 2.

The plain bin average is the wrong prediction. The offsets from it sit in the first bin, where D_ℓ and F_ℓ rise steeply and the bandpower window differs from a flat average by 7%, and at the peaks and troughs; the bandpower windows account for them (the pitfall of Section 8.11.3). At the resolution of their Fig. 5 such offsets are invisible; with 1350 simulations they are measurable.

The top-hat correlates more. Beyond the first three bins the correlation matrix (Figure 8.41b) is diagonal-dominated, but the top-hat's largest off-diagonal correlation slightly exceeds their 10%. It has the sharpest edge, and so the longest leakage tail.

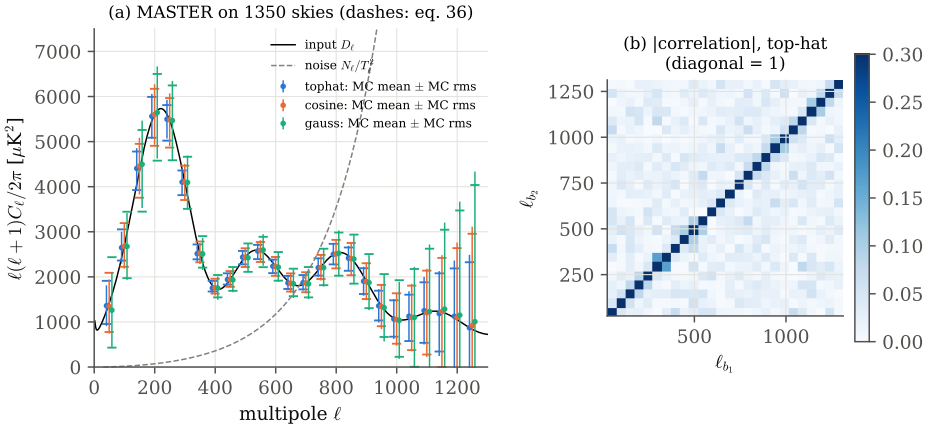


Figure 8.41: MASTER on 1350 simulated Boomerang-like skies, in the format of Hivon et al.'s Fig. 5. (a) Mean and scatter of the estimates for the three weightings (slightly offset in ℓ), the input spectrum (black), the analytic error bars of [Hivon (36)] (dashes) and the noise on the sky (grey dashed). (b) Absolute correlation matrix of the top-hat bandpowers.

The shape of the distribution (their Fig. 6). Their last test compared the histogram of the estimates in the bins at $\ell = 50$, 200 and 750 with a χ^2_ν with the ν of (8.81), with no free parameter, and with a Gaussian of the same mean and variance. Figure 8.42 is the same figure from our simulations. At $\ell = 50$, where ν is 18, 10 and 4, the histograms are skewed and the χ^2_ν follows them where the Gaussian does not, most clearly for the Gaussian weighting with $\nu \approx 4$; at $\ell = 200$ and 750 the two models coincide. Their conclusion carries over unchanged: for small sky coverage the likelihood of a low-ℓ bandpower must be modelled as a χ^2, not as a Gaussian [Hivon (37)].

Lesson. At the resolution of the paper's figures MASTER passes every test; with the same number of simulations, read as numbers, each few-per-cent residual becomes measurable and has a cause we can name: a non-optimal estimator, a transfer function measured once and shared by every estimate, and a theory that must be passed through the bandpower windows.

8.14.2 Efstathiou (2004), Hartlap et al. (2007), Taylor et al. (2013), NaMaster

Efstathiou. *His question:* through a Galactic cut, can a cheap pseudo-C_ℓ estimator be nearly as good as exact maximum likelihood, and can its covariance be computed without simulations?

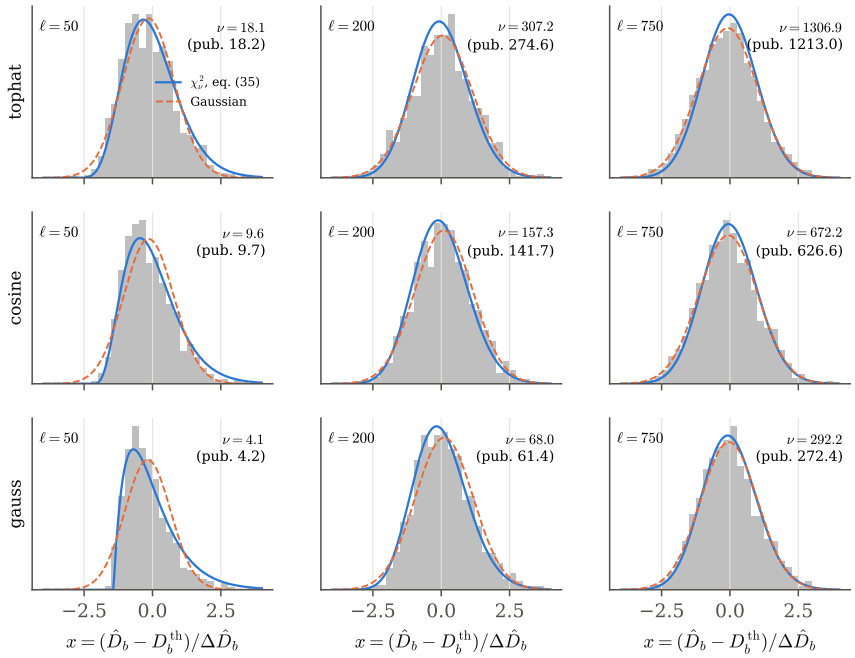


Figure 8.42: Distributions of the MASTER estimates, centred on the input and divided by the analytic error, at $\ell = 50$, 200 and 750 (columns) for the three weightings (rows), as in Hivon et al.'s Fig. 6. Blue: the χ^2_ν of eq. (35) with our ν (their published ν in brackets); orange dashed: a Gaussian of the same mean and variance.

His equations, here: the coupling matrix (his eqs. 9–11) is (8.61), with his symmetric $G_{\ell\ell'} = M_{\ell\ell'}/(2\ell'+1)$ and his $\Xi(\ell, \ell', \mathcal{W})$ ours; the exact covariance (his 14, 18) is (8.73); the approximation (15a) is (8.75), with his $W^{(2)}$ the spectrum $\mathcal{W}^{(2)}$ of the squared window. *What we ran:* the 2000 skies through a ±20° band, the exact band covariance from the ring sum (8.74), and the approximation against both. His Fig. 1 (the mean pseudo-spectrum is the input convolved with M, and the deconvolved estimate is unbiased) is our mean-square pull of 1.06 (Section 8.10.3) and the unbiased MASTER of Section 8.11.3; his Fig. 4 (exact covariance against simulations) is our ratio 1.005 ± 0.035, the Monte Carlo precision. *The difference:* his approximation is accurate at $\ell \gtrsim 20$ and wrong by up to 25% at $\ell = 2$; ours is wrong by 43% at $\ell = 2$ and 10% at $\ell = 20$. His cut of ±10° removes $\sin 10° = 17\%$ of the sky, ours $\sin 20° = 34\%$, twice as much; the wider cut widens the coupling kernel, and the approximation, which assumes the spectrum constant across the kernel, fails further up in ℓ. *Lesson:* the approximation is a high-ℓ tool, and the multipole where it becomes good moves up with the cut.

Hartlap et al. *Their question:* when a covariance is estimated from n simulations and inverted, is the inverse unbiased, and if not, by how much? *Their equation, here:* the factor $(n-p-2)/(n-1)$ of their eq. 17 is derived in (8.78); their n and p (simulations and data points) are ours, and their debiased inverse is our Ψ. *What we ran:* their Fig. 1 test, $n = 60$, $p_1 = 240$, three covariance models, with 2000 repetitions instead of 10^4 (Figure 8.38a). *Side by side:* the raw trace ratio follows $(n - p - 2)/(n - 1)$ and the corrected one is one to within 0.005. There is no difference to explain beyond the Monte Carlo noise. *Lesson:* the factor is exact for Gaussian vectors, and it holds for skewed CMB bandpowers too (Section 8.13.3).

Taylor et al. *Their question:* once the precision matrix is unbiased, how noisy is the error bar it gives, and how many simulations does a target precision need? *Their equations:* $\varepsilon = \sqrt{2/(N_S - N_D - 4)}$ (their eq. 55) and the planning rule (8.79) (their eq. 58), with their N_S, N_D

our n, p. *What we ran:* the one-parameter amplitude fit of Section 8.13.4, 2000 times for each n. *Side by side:* at $N_S = 40$ the scatter of $\widehat{\sigma_A^2}$ is 0.44, against 0.53 from their eq. (55) and 0.43 from the exact (8.80). *The difference:* their formula is the relative scatter of the Fisher information, an inverse χ^2_{n-p}, and the "−4" is that distribution's long tail; the variance itself is a χ^2_{n-p} and scatters as $\sqrt{2/(n-p)}$, smaller at small $n-p$ (Section 8.13.4). *Lesson:* their rule errs on the safe side; it asks for 233 simulations for 5% error bars on our 29 bins.

NaMaster, in words. NaMaster is the standard pseudo-C_ℓ code of current surveys [169], and the question for us is whether it computes what we derived. Its conventions map onto ours one to one. Its mode-coupling matrix for two spin-0 fields with masks v and w is $M^{00}_{\ell\ell'} = \frac{2\ell'+1}{4\pi}\sum_{\ell''} P^{vw}_{\ell''}(\ell\ \ell'\ \ell'';0\ 0\ 0)^2$ with $P^{vw}_{\ell''} = (2\ell''+1)\,\mathrm{PCL}_{\ell''}(v,w)$ [169, eq. 12], which for $v = w = W$ is our (8.61). Its bandpowers are weighted sums of the coupled pseudo-C_ℓ over the multipoles of each band, with weights normalised to one; the binned coupling matrix is inverted, and theory is compared with data only through the bandpower windows $\mathcal{F}_{q\ell}$ [169, eqs 15–19], our (8.68). The default weights are flat in C_ℓ; Hivon's flat-in-D_ℓ binning corresponds to weights proportional to $\ell(\ell+1)$. Beyond what we use, it handles spin-2 fields (polarisation, shear) with E/B purification, and deprojects foreground templates with the corresponding bias correction. NaMaster is not installed in our environment, so the comparison is with its published equations, not with its output.

8.15 How it is done in research

A real temperature analysis has one sky and a stack of simulations. Everything it publishes about the spectrum, from the error bar of a single bandpower to the claim that no systematic effect is left, is a statement about how the estimator behaves on simulated skies that went through the same mask and the same code as the data. The order in which a full-sky analysis runs (maps, spectra, deconvolution, covariance, likelihood, checks) was laid out in Section 8.8. The mask changes what each of those steps has to do, and a real analysis makes a definite choice at each one. Seeing those choices once, and then running a small analysis that makes the same ones on a laptop, ties together the pieces of this chapter.

8.15.1 The cut-sky choices of a real analysis

The reference here is the Planck 2018 high-ℓ temperature likelihood [75], which is a pseudo-C_ℓ analysis of exactly the kind built in Sections 8.9 to 8.11.

The mask. A Galactic mask is made by thresholding a map of foreground emission, so that it follows the dust and not a straight line of latitude, and it is multiplied by masks of point sources, of carbon monoxide line emission and of extended objects, each specific to the frequency of the map. Planck keeps 66%, 57% and 47% of the sky at 100, 143 and 217 GHz (the masks T66, T57 and T47), because dust is brighter at the higher frequency. The Galactic part is apodised with a Gaussian of 4.71° full width at half maximum and the point-source holes with one of 30′ [75, §3.2]. The reason given is the one of Section 8.10.4: a soft edge confines the coupling to neighbouring multipoles and makes the bandpowers less correlated. The price is the one we computed there in modes: the apodisation removes about 10% of the effective sky fraction [75, §3.2].

quantity	published	ours	why they differ
Hivon et al. (2002), Boomerang test			
$f_{\rm sky}$ of the ellipse	1.82%	1.82%	—
w_2^2/w_4: top-hat, cosine, Gaussian	1, 0.514, 0.223	1.000, 0.514, 0.223	—
F at $\ell = 50, 200, 750$	0.20, 0.75, 0.89 (from ν)	0.199, 0.846, 0.960 (eq. B15: 0.185, 0.838, 0.957)	real scans vs ideal parallel scans
ν, top-hat, $\ell = 50, 200, 750$	18.2, 274.6, 1213	18.1, 307.2, 1306.9	ratio of the F's
ν, cosine	9.7, 141.7, 626.6	9.6, 157.3, 672.2	ratio of the F's
ν, Gaussian	4.2, 61.4, 272.4	4.1, 68.0, 292.2	ratio of the F's
their cosine ν from their top-hat ν	9.7, 141.7, 626.6	9.4, 141.2, 623.8	—
their Gaussian ν from their top-hat ν	4.2, 61.4, 272.4	4.1, 61.3, 270.7	—
$\delta F/F$ coefficient a, eq. (38)	0.6–0.9	0.63 ± 0.04 (counting: 0.52)	mask couples the bins
MC / analytic error, eq. (36), median	"almost identical"	1.08, 1.05, 1.04	MASTER not optimal
the same, first bin	—	0.95, 0.92, 0.84	D_ℓ, F_ℓ steep in the bin
largest pull of the mean vs windows	none	3.6, 5.4, 4.9	common errors of F and of the noise; with them, 1.6, 1.5, 2.0
largest offset vs plain bin average, in single-sky σ	none	0.32, 0.20, 0.21	bandpower windows
largest off-diagonal correlation	$\leq 10\%$	17, 9, 8%	sharp top-hat edge
low-ℓ shape	χ^2_ν, not Gaussian	same	—
simulations	450 + 250 + 1350	450 + 252 + 1350	maps, not timestreams
Efstathiou (2004)			
approximation (15a), $\ell = 2$	$\leq 25\%$ off ($\pm 10°$)	1.43 ($\pm 20°$)	wider cut
approximation (15a), $\ell \gtrsim 20$	accurate	0.90 at 20, 0.95 at 40	—
exact (18) vs simulations	"almost perfect"	1.005 ± 0.035	—
M invertible	cuts $\lesssim 30°$	band yes, caps $< 50\%$ no	pair-separation rule
Hartlap et al. (2007); Taylor et al. (2013)			
raw trace ratio, $n = 60$, $p = 40$	$(n - p - 2)/(n - 1) = 0.305$	0.307	—
corrected ratio	1	1 ± 0.005	—
scatter of σ_A^2, $N_S = 40$	0.53 (eq. 55)	0.44	"−4": scatter of the information, not of the variance
N_S for 5% errors, $N_D = 29$	$200 + N_D + 4$	233	—

Table 8.8: Published cut-sky results and the same quantities from our code. Where three values are given for the Boomerang test they are for the top-hat, cosine and Gaussian weightings.

The coupling and the bandpowers. The coupling matrix of each mask is computed exactly from the mask's own spectrum, as in (8.61), and inverted in bins. The bins are $\Delta\ell = 5$ for $30 \le \ell \le 99$, 9 up to $\ell = 1503$, 17 up to 2013 and 33 up to 2508 [75, §3.2.5]. They are narrow because a 60% mask couples multipoles over only a few ℓ (Figure 8.27); they widen at high ℓ because there the noise, not the coupling, sets the useful resolution. The spectra are cross-spectra between the two half-missions and between detector sets, so that no noise bias enters (Section 8.3.3).

The covariance. Planck does not estimate its covariance from simulations. With hundreds of bandpowers for each of several cross-spectra, (8.79) would demand thousands of end-to-end simulations, and only 300 exist. The covariance is computed analytically for a fiducial model, with the large Galactic mask treated by the MASTER-type expressions of Section 8.12.3 (the approximation that Efstathiou derived), the point-source masks by a correction measured on Monte Carlo simulations, and the anisotropy of the noise by a heuristic correction [75, §3.1]. The 300 FFP10 end-to-end simulations are used to test the result [75].

The likelihood and the checks. Above $\ell = 30$ the bandpowers enter a Gaussian likelihood, $-2\ln\mathcal{L} = (\hat{C} - C(\theta))^{\mathsf{T}}\Sigma^{-1}(\hat{C} - C(\theta))$, with the covariance Σ fixed at the fiducial model. The Gaussian shape is a choice justified by the central limit theorem, and its symmetry causes a small bias for $30 \le \ell < 100$ [75, §3.1], which is the skewness of Section 8.12.1 seen from the likelihood side. Its size follows from the mode count (8.72). A bin of five multipoles at $\ell = 50$ on about 60% of the sky holds $\nu \approx (2 \times 50 + 1) \times 5 \times 0.6 \approx 300$ modes, so the scaled χ^2 has skewness $\sqrt{8/\nu} \approx 0.16$; at $\ell = 30$, $\nu \approx 180$ and the skewness is 0.21. Small, but not zero, and it is the multipole range where a symmetric likelihood errs most; above $\ell = 100$ the bins are wider and the skewness falls below 0.1.

Null tests difference half-missions, detector sets, and odd and even rings of the scan, so that the sky cancels and only noise and systematics remain. The logic is that of Section 4.1.1: we assume the difference is noise alone, because for that case the simulations give us its distribution; and we ask whether a result like the observed one, or one further in the same direction, would be unremarkable if that were true. The probability of such a result is the probability to exceed (PTE) under the simulations [75]; the null-test example of Section 8.15.2 works one through. Ground-based experiments such as ACT and SPT map patches of a few to a few tens of percent of the sky, filter their timestreams, and so add the transfer function of Section 8.11.4, measured on noise-free simulations; in every other respect their spectra are built the same way.

8.15.2 Our pipeline, end to end

The setting. The script `31_pipeline.py` runs the whole chain of Figure 8.43 once, on a sky whose random seed is kept apart from those of the simulations, so that it plays the part of the data. The instrument is the toy experiment of Section 8.2: $N_{\mathrm{side}} = 256$, a 30′ Gaussian beam and white noise of 200 μK arcmin. The data are two half-mission maps of the same sky, each with noise $\sqrt{2}$ larger, so that their average has the noise of the full experiment. The window is the product of a Galactic band, $|b| < 20°$ with a 5° cosine taper, and 400 randomly placed point-source holes of radius 1°, now themselves given a 1° taper (the band and the holes of Section 8.9.2, multiplied together). It keeps $f_{\mathrm{sky}} = 0.638$ of the sky with $w_2 = 0.878$. The questions are the ones a real analysis answers: what is the spectrum, how uncertain is it, is anything in the data that is not sky, and what amplitude of the fiducial spectrum does the data prefer?

Listing 8.7: One observation and its bandpowers (`code/ch08/31_pipeline.py`)

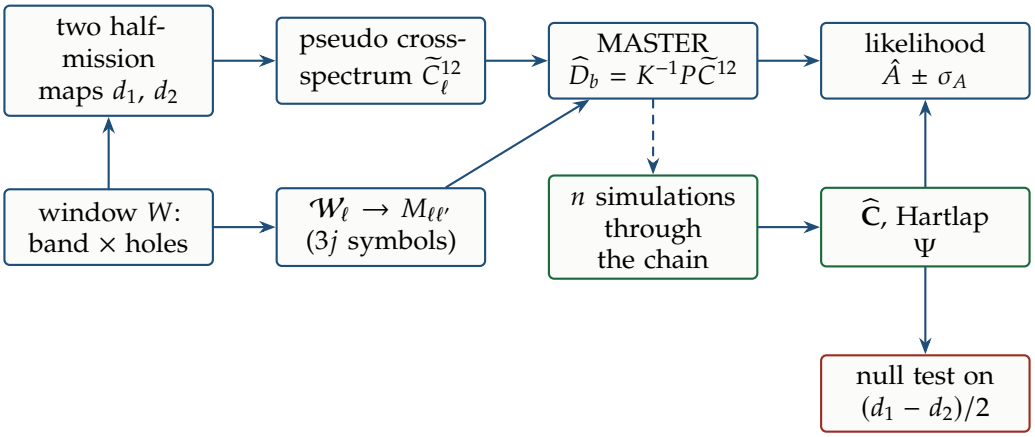


Figure 8.43: The cut-sky analysis as it runs. The window enters twice: it multiplies the maps and, through its spectrum alone, fixes the coupling matrix. The simulations (green) pass through the same chain as the data (the dashed arrow) and supply what no formula gives here: the covariance, its inverse and the distribution of the null test (red).

```python
def observe(rng, systematic=False):
    """Two half-mission maps of one sky: same signal, independent noise of variance 2 sigma^2."""
    a = cs.synalm(cl, L, rng)
    s = hp.alm2map(hp.almxfl(a, exp.transfer(L)), NSIDE, lmax=L)
    d1 = s + cs.noise_map(exp, rng, scale=np.sqrt(2.0))
    d2 = s + cs.noise_map(exp, rng, scale=np.sqrt(2.0))
    if systematic:
        d2 = d2 + STRIPE
    return d1, d2


def bandpowers(d1, d2):
    """MASTER cross-spectrum of the halves, and the MASTER spectrum of the half-difference."""
    a1 = hp.map2alm(W * d1, lmax=L, iter=0)
    a2 = hp.map2alm(W * d2, lmax=L, iter=0)
    cross = hp.alm2cl(a1, a2)
    null = hp.alm2cl(0.5 * (a1 - a2))
    return Kinv @ (P @ cross), Kinv @ (P @ null)
```

The code. `observe` draws one Gaussian sky from the fiducial spectrum, smooths it with the beam and pixel window, and adds two independent noise maps; the optional stripes are the systematic used below. `bandpowers` multiplies both maps by the window, transforms them, and forms two spectra: the cross spectrum of the halves and the spectrum of their half-difference, in which the sky cancels. Both are binned with P (bins of 20) and deconvolved with K^{-1}, the binned coupling of (8.68)'s definition with $F_\ell = 1$ (no filter) and T_ℓ^2 the beam and pixel window. The coupling matrix is computed once from the window's spectrum, and the same two functions run on the data and on each of the 400 simulations. The bandpowers used are the $p = 29$ bins with $22 \le \ell \le 601$.

The covariance and its inverse. The covariance of the $p = 29$ bandpowers is the sample covariance of the $n = 400$ simulations, and its inverse is corrected by the Hartlap factor of

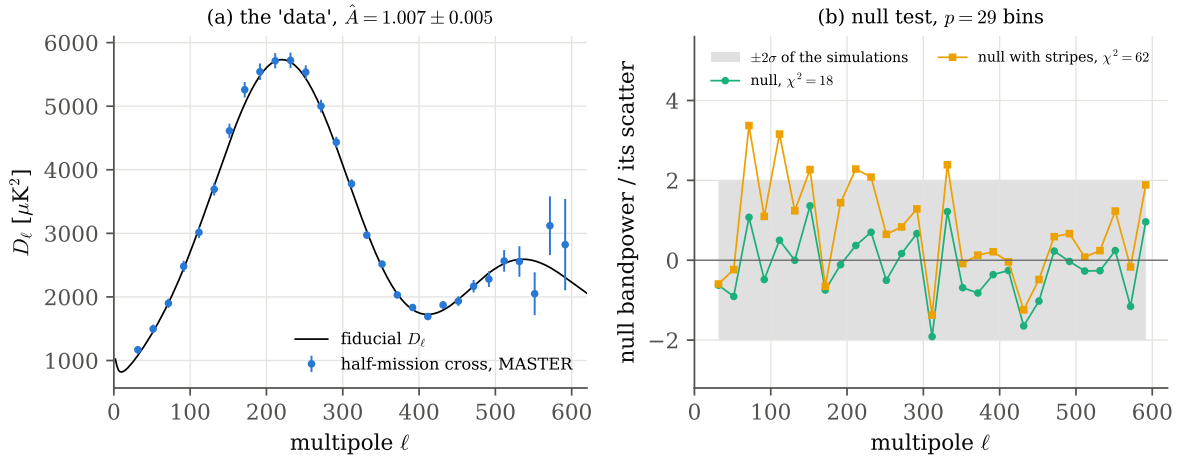


Figure 8.44: (a) The "data": MASTER bandpowers of the half-mission cross spectrum through the band-and-holes window, with error bars from the simulations, on the fiducial spectrum. (b) The null spectrum of the half-difference map, minus its simulated mean and divided by its simulated scatter, without (green) and with (orange) a 3 μK ring-striping systematic in the second half. The stripes add power, so almost every orange point sits above its green partner, most of all below $\ell \approx 350$; which orange points leave the grey band, and by how much, changes from one draw to the next.

Section 8.13.2:

$$\Psi = \frac{n-p-2}{n-1}\,\widehat{\mathbf{C}}^{-1} \overset{(1)}{=} \frac{400-29-2}{400-1}\,\widehat{\mathbf{C}}^{-1} \overset{(2)}{=} 0.925\,\widehat{\mathbf{C}}^{-1}.$$

(1) $n = 400$ simulations and $p = 29$ bins.
(2) $369/399 = 0.9248$, the script's 0.925.

Without the factor every χ^2 would be 8% too large and every error bar 4% too small. With it, the variance of a single parameter is still uncertain by $\sqrt{2/(n-p)} = \sqrt{2/371} = 7.3\%$ by (8.80), its error bar by 3.7%, and biased high by $(n-p)/(n-p-2) - 1 = 0.5\%$. Those are the cost of 400 simulations.

Example: the amplitude and its error bar. The simplest physics question is how much CMB power there is relative to the fiducial model. Write the model for the bandpowers as $\boldsymbol{D} = A\boldsymbol{t}$, with $\boldsymbol{t}$ the fiducial bandpowers $P_{b\ell}C_\ell^{\rm fid}$ and A a free amplitude, and the Gaussian likelihood with the precision matrix Ψ:

$$\begin{aligned} -2\ln\mathcal{L}(A) &\overset{(1)}{=} (\widehat{\boldsymbol{D}} - A\boldsymbol{t})^{\mathsf{T}}\Psi(\widehat{\boldsymbol{D}} - A\boldsymbol{t}) + \text{const} \\ &\overset{(2)}{=} A^2\,\boldsymbol{t}^{\mathsf{T}}\Psi\boldsymbol{t} - 2A\,\boldsymbol{t}^{\mathsf{T}}\Psi\widehat{\boldsymbol{D}} + \widehat{\boldsymbol{D}}^{\mathsf{T}}\Psi\widehat{\boldsymbol{D}} + \text{const} \\ &\overset{(3)}{=} \frac{(A-\hat{A})^2}{\sigma_A^2} + \text{const}', \qquad \hat{A} = \frac{\boldsymbol{t}^{\mathsf{T}}\Psi\widehat{\boldsymbol{D}}}{\boldsymbol{t}^{\mathsf{T}}\Psi\boldsymbol{t}}, \quad \sigma_A^2 = \frac{1}{\boldsymbol{t}^{\mathsf{T}}\Psi\boldsymbol{t}}. \end{aligned}$$

(1) The Gaussian likelihood of the bandpowers, with the covariance fixed (it does not depend on A).
(2) Expand the quadratic form; Ψ is symmetric, so the two cross terms are equal.
(3) Complete the square in A; everything without A goes into the constant.

The likelihood in A is exactly Gaussian, with its peak at a weighted average of the bin-by-bin ratios $\widehat{D}_b/t_b$ and its width given by the Fisher information $\boldsymbol{t}^{\mathsf{T}}\Psi\boldsymbol{t}$ (Chapter 10). Because $\hat{A}$ is

linear in $\widehat{\boldsymbol{D}}$ and MASTER is unbiased, $\hat{A}$ is unbiased whatever Ψ is; the precision matrix only sets the weights, and a noisy one gives slightly suboptimal weights, not a bias. On the data $\hat{A} = 1.007 \pm 0.005$. Run on each simulation, the same estimator has mean 1.0006 and standard deviation 0.0046, which matches the Fisher width, and the interval $\hat{A} \pm \sigma_A$ covers the input $A = 1$ in 68% of the simulations, as a 68% interval should. The data sit within two standard deviations of the input, and the fit has $\chi^2 = 18.0$ for $p - 1 = 28$ degrees of freedom (PTE 0.93): one amplitude describes all 29 bins.

Example: a null test, and what could have produced its value. *The event.* The half-difference map $(d_1 - d_2)/2$ contains the noise of the two halves and no sky. Its bandpowers, minus their mean over the simulations, give $\chi^2 = 17.8$ with the precision matrix Ψ_{null} of the simulated null spectra, for $p = 29$ bins. *What could have produced it?* Two things: noise behaving as the simulations say, or noise plus something the simulations do not contain (a drift, a stripe, a calibration change between halves). *How probable is the value under the first?* The simulations answer directly: the fraction of simulated null spectra with a larger χ^2 is the PTE, 0.92 in this run. That is not small enough to doubt the noise model, and the test passes. It is a test of the first hypothesis alone (Chapter 4): a small PTE would say that the noise model fails, not what replaced it.

To see the test fail, the script adds to the second half a pattern of the kind a scan can leave: a random offset of 3 μK rms on every ring of constant latitude. Such a pattern $f(\theta)$ depends only on the colatitude, so its coefficients are $a_{\ell m} = \int f(\theta) \lambda_{\ell m}(\theta) \sin\theta \, \mathrm{d}\theta \int_0^{2\pi} e^{-im\phi} \mathrm{d}\phi$, and the ϕ integral vanishes for every $m \neq 0$. On the full sky all its power sits in the $m = 0$ modes, a little at every ℓ; the mask spreads it to other m. The null χ^2 becomes 62, with PTE = 0.003. Figure 8.44b shows why a χ^2 is the right summary. In this run the largest single excess is 3.4 standard deviations, which a reader scanning 29 bins might put down to bad luck. The χ^2 does better: it adds up the excess of every bin, the small ones as well as the large. Under noise alone it is a χ^2_{29}, with mean 29, variance $2 \times 29 = 58$ and standard deviation 7.6; the observed value lies $(62 - 29)/7.6 = 4.3$ standard deviations above the mean. A second, cruder count looks only at signs: under noise alone each bin is as likely to sit above zero as below, so the number of high bins is binomial with mean 14.5 and standard deviation $\sqrt{29/4} = 2.7$. Here 20 of the 29 bins are high, and 20 or more would happen by chance with probability $p = 0.031$. The count throws away how far each bin is from zero, so it is weaker than the χ^2: stripes that push a few bins far up can leave it unremarkable. Is this one data set a lucky catch? The script repeats the test on 20 fresh data sets, each with its own sky, noise and stripes. The null χ^2 has median 63, so stripes of this size usually fail the test. But not always: the lowest of the 20, 41, lies only 1.5 standard deviations above the mean, a value noise alone reaches often enough that the test could pass. A null test that passes does not prove the data are clean; it says the contamination is not much larger than the noise.

Example: why the simulated χ^2 values come out low. There is a subtlety in using the same simulations for the covariance and for the null distribution. For the simulations themselves, with residuals $\boldsymbol{r}_i = \boldsymbol{v}_i - \bar{\boldsymbol{v}}$,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{r}_i^{\mathsf{T}} \Psi \boldsymbol{r}_i &\overset{(1)}{=} \frac{n-p-2}{n(n-1)} \sum_i \operatorname{tr}\!\left(\widehat{\mathbf{C}}^{-1} \boldsymbol{r}_i \boldsymbol{r}_i^{\mathsf{T}}\right) \\ &\overset{(2)}{=} \frac{n-p-2}{n(n-1)} \operatorname{tr}\!\left(\widehat{\mathbf{C}}^{-1} (n-1) \widehat{\mathbf{C}}\right) \\ &\overset{(3)}{=} \frac{(n-p-2)\,p}{n} = \frac{369 \times 29}{400} = 26.75.\end{aligned}$$

(1) $\Psi = \frac{n-p-2}{n-1}\widehat{\mathbf{C}}^{-1}$, and a number equals its own trace, $\boldsymbol{r}^{\mathsf{T}}\mathsf{A}\boldsymbol{r} = \mathrm{tr}(\mathsf{A}\boldsymbol{r}\boldsymbol{r}^{\mathsf{T}})$.

(2) The sum of the outer products of the residuals is $(n-1)\widehat{\mathbf{C}}$, by the definition of the sample covariance.

(3) $\mathrm{tr}\,\mathbb{1}_p = p$.

This is an identity, true for every set of simulations, and the script finds 26.8 (scatter 7.2). An independent data vector $\boldsymbol{d}$ behaves differently, and three facts give its value. *The residual's covariance.* $\boldsymbol{r} = \boldsymbol{d} - \bar{v}$, with $\boldsymbol{d}$ independent of the simulations; $\boldsymbol{d}$ has covariance $\mathbf{C}$ and the mean of n simulations has $\mathbf{C}/n$, so $\mathbb{E}[\boldsymbol{r}\boldsymbol{r}^{\mathsf{T}}] = \mathbf{C}(1+1/n)$. *Independence.* Ψ is built from the simulations, so it is independent of $\boldsymbol{d}$; for Gaussian draws the sample covariance is also independent of the sample mean $\bar{v}$ (Section 8.13.2), so Ψ is independent of $\boldsymbol{r}$ and the average of the product factorises. *Unbiasedness.* $\mathbb{E}[\Psi] = \mathbf{C}^{-1}$, by (8.78) with the Hartlap factor. Then

$$\mathbb{E}[\boldsymbol{r}^{\mathsf{T}}\Psi\boldsymbol{r}] = \mathbb{E}\,\mathrm{tr}(\Psi\boldsymbol{r}\boldsymbol{r}^{\mathsf{T}}) = \mathrm{tr}\big(\mathbb{E}[\Psi]\;\mathbb{E}[\boldsymbol{r}\boldsymbol{r}^{\mathsf{T}}]\big) = \mathrm{tr}\big(\mathbf{C}^{-1}\mathbf{C}(1+1/n)\big) = p\,(1+1/n) = 29.07.$$

The simulations, by contrast, were used to build Ψ: each $\boldsymbol{r}_i$ helped make the covariance it is measured against, and the identity above gives 26.75. They fit their own covariance a little too well, so the reference distribution sits about two units of χ^2 low and the PTE of real data comes out slightly too small. The test errs towards failing, which is the safe side; a cleaner test keeps a separate set of simulations for the null distribution.

What the laptop version leaves out. Each step of Figure 8.43 was reduced to fit on a laptop. Each reduction removes something whose effect is either known or not the subject here, and each has a price.

Taxonomy: The cut-sky pipeline: research version and laptop version

step	research version	here	cost of the reduction
maps	several frequencies, $N_{\mathrm{side}} = 2048$, $\ell \le 2508$, foregrounds	one channel, $N_{\mathrm{side}} = 256$, CMB only, $\ell \le 601$	no damping tail, no foreground model
mask	thresholded foreground maps, thousands of sources, T66–T47 per frequency	geometric band, 400 random holes	same mathematics, not tied to real foregrounds
coupling	exact, from $\mathcal{W}_\ell$ to $2\ell_{\mathrm{max}}$	the same	none
noise	anisotropic, correlated, from the scan	white, uniform, two halves	no anisotropic-noise correction
covariance	analytic for a fiducial model, Monte Carlo corrections, tested on 300 end-to-end simulations	400 Gaussian simulations, Hartlap factor	7% uncertainty on a parameter variance; no systematics in $\mathbf{C}$
null tests	half-missions, detector sets, odd/even rings, every frequency	one half-mission null, one injected systematic	tests only what we put in
likelihood	multi-frequency, foreground and instrument nuisance parameters	one amplitude	no degeneracies to marginalise

The reduction that matters most for the statistics is in the covariance. At $N_{\mathrm{side}} = 2048$ and $\ell_{\mathrm{max}} = 2500$ each simulation costs about $(2048/256)^3 \approx 500$ times more than ours (the transform cost grows as $N_{\mathrm{pix}}^{1/2}\ell_{\mathrm{max}}^2 \propto N_{\mathrm{side}}^3$, [Hivon (30)]), and the data vector has hundreds of bins per spectrum. That is why a research analysis models its covariance and uses the simulations

to test the model, while ours can afford to use the simulations directly. The structure is the same in both: the same code on data and on simulations, and every statement about the data read off the simulations.

How this is used in real research: The simulations as the instrument

In Chapter 7 the simulations reproduced the $\chi^2_{2\ell+1}$ law and the cosmic variance, which were known in advance: they *verified* a formula, and agreement tested the code. In a cut-sky analysis the same loop is the measuring *instrument*. The transfer function of a filtered scan, the covariance that enters the likelihood, the distribution of a null-test χ^2 and the coverage of the final error bars exist only as properties of the simulations. Planck, ACT and SPT all compute their spectra with pseudo-C_ℓ estimators of this kind, and all of them obtain those four things, or test the analytic models of them, by running simulated skies through the pipeline.

8.15.3 A mask also changes the shapes in a map

The spectrum is not the only statistic a mask disturbs. Take an **excursion set** of a map [the region where the temperature is above a chosen threshold]. Its shape can be summarised by counting: the number of separate hot regions, and the number of holes inside them (cold islands surrounded by hot sky). The difference of the two counts is the Euler characteristic, a number that does not change under any continuous deformation of the regions (see the topology notes), and for a Gaussian field its average at each threshold is known exactly (Chapter 12).

A mask changes both counts, and not by blurring. A Galactic band that crosses a hot region cuts it in two, and the count of regions goes up by one. A point-source hole that falls inside a hot region punches a hole in it, and the count of holes goes up by one. Regions that touch the edge of the mask are cut open, and it is not clear whether they should be counted at all. Figure 8.45 shows the three cases. Schmalzing and Górski found that the shape statistics of a map stay unbiased if the field is smoothed first and cut afterwards, but that smoothing a map which has already been cut biases them by up to one standard deviation, and that the remedy is to use only pixels far from the cut [110]. The cure is the one of this chapter. The same mask is applied to every simulated sky, the shape statistics are computed on data and simulations alike, and their sampling distribution, their covariance and their inverse are read off the simulations. Chapter 12 builds those statistics, and the simulations of this chapter will provide their sampling distributions.

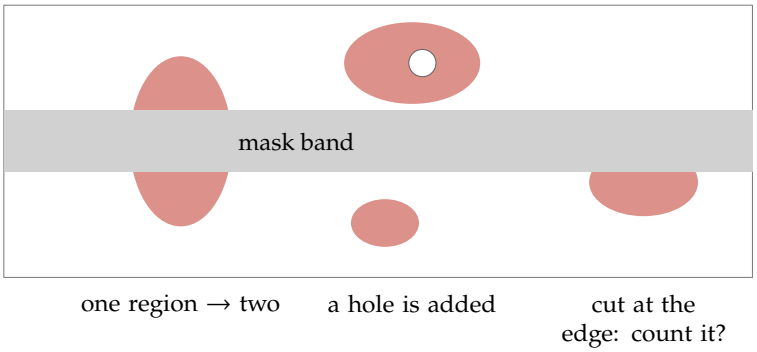


Figure 8.45: An excursion set (red) seen through a mask. Left: a hot region that crosses the band is counted as two. Centre: a point-source hole (white disc) inside a hot region adds a hole, while the small region below it is untouched. Right: a region cut open by the edge of the band; whether it counts, and as what, is a choice the analysis must make and the simulations must share.

Recap

- A window W multiplies the map; in harmonic space it convolves. The pseudo-coefficients $\tilde{a}_{\ell m}$ mix true coefficients at many (ℓ', m') through the window's own $w_{\ell m}$, and a sharp edge gives a mask spectrum $\mathcal{W}_\ell \propto \ell^{-3}$ that moves power far in ℓ. The moments $f_{\rm sky} w_i = \frac{1}{4\pi}\int W^i$ set the simple scalings.
- On average $\langle \widetilde{C}_\ell \rangle = \sum_{\ell'} M_{\ell\ell'} F_{\ell'} T^2_{\ell'} C_{\ell'} + \langle \widetilde{N}_\ell \rangle$, with $M_{\ell\ell'} = \frac{2\ell'+1}{4\pi}\sum_L (2L+1)\mathcal{W}_L (\ell\,\ell'\,L;0\,0\,0)^2$ computed from the mask alone; the matrix passes a direct mode-by-mode check and the mean of the simulated skies.
- Dividing by $f_{\rm sky} w_2$ is exact only for a flat spectrum and biases a curved one. M itself is invertible only if the window contains pairs of directions at every separation; small patches have null modes. MASTER inverts the coupling in bins, $\widehat{D}_b = K^{-1}_{bb'} P_{b'\ell}(\widetilde{C}_\ell - \langle \widetilde{N}_\ell \rangle)$, is unbiased for the binned spectrum, and is compared with theory through the bandpower windows $\mathcal{F}_{b\ell}$. A filter adds a transfer function, measured on noise-free simulations.
- Apodisation trades sky (modes) for locality of the coupling and less leakage.
- The cut-sky estimate is a weighted sum of χ^2_1 variables: approximately $\chi^2_{\nu_{\rm eff}}$ with $\nu_{\rm eff} = 2\langle X\rangle^2/\mathrm{Var}\,X < 2\ell+1$, skewed at low ℓ and Gaussian at high ℓ. Hivon's $\nu = (2\ell+1)\Delta\ell f_{\rm sky} w_2^2/w_4$ counts modes and fails where the mask's correlations matter.
- Neighbouring multipoles are correlated. The exact covariance is a sum over $|\langle \tilde{a}\tilde{a}^*\rangle|^2$; above the width of the kernel it is close to $2C_\ell C_{\ell'}\Xi(\ell,\ell',\mathcal{W}^{(2)})$ (Efstathiou), and Knox's $2C_\ell^2/[(2\ell+1)f_{\rm sky}]$ is a mode count that underestimates the scatter of a sharp mask.
- A covariance from n simulations of p bins is noisy element by element, and its inverse is biased by $(n-1)/(n-p-2)$; the Hartlap factor removes the bias, not the noise. A parameter variance is then uncertain by $\sqrt{2/(n-p)}$, and $n > 2/\varepsilon^2 + p + 4$ simulations give fractional precision ε.
- Our code reproduces the cut-sky results of Hivon et al., Efstathiou, Hartlap et al. and Taylor et al. (Table 8.8), and a half-mission analysis through a band-and-holes mask returns an unbiased amplitude with correct coverage, passes its null test and fails it when a 3 μK striping is added.
- A mask also cuts and pierces the regions of an excursion set: shape statistics must be computed through the same mask on data and simulations.

Reading the literature

- **The pseudo-C_ℓ method.** Hivon et al., the MASTER paper: the window and its moments, the coupling kernel through $3j$ symbols, the transfer function, binning, the Monte Carlo covariance and the Boomerang test [168]. Alonso, Sanchez and Slosar on NaMaster, the same framework written for spin-0 and spin-2 fields with bandpower windows and contaminant deprojection [169]. Hobson et al. (eds.), *Bayesian Methods in Cosmology*, Chapter 10, for the CMB data model with beam, noise and window and for where pseudo-C_ℓ estimators sit next to likelihood methods [HobsonBMC pp. 231–239].
- **Errors on a power spectrum.** Knox on noise, beam, sky fraction and cosmic variance [166]; Efstathiou on pseudo-C_ℓ against maximum likelihood, on the invertibility of the coupling and on the analytic covariance of a cut-sky spectrum [134]; Hamimeche and Lewis on the exact full-sky likelihood and its approximations [16].

- **Covariances from simulations.** Hartlap, Simon and Schneider on the bias of an inverted sample covariance [6]; Taylor, Joachimi and Kitching on how its noise propagates into parameter errors and how many simulations that requires [165].
- **A real analysis.** Planck 2018 results V, on power spectra and likelihoods: masks, apodisation, cross-spectra, binning, analytic covariances, end-to-end simulations and null tests [75]. Górski et al. on HEALPix [7].
- **Shapes through a mask.** Schmalzing and Górski on Minkowski functionals of CMB maps on an incomplete sky [110].

Chapter summary

An instrument never delivers the sky. It smooths it with a beam, adds noise, and sees only part of it. The beam and the noise act on each multipole separately and can be undone exactly, which is why the full-sky analysis was a verification: every property of the estimator had a formula, and the simulations confirmed it. A mask acts on each direction separately, which mixes the multipoles. The average can still be undone, with a coupling matrix computed from the mask alone, but the scatter, the correlations and the shape of the distribution of the estimates have no general formula. They are measured on simulated skies sent through the same mask and the same code as the data, and the number of simulations then sets the precision of everything that follows. In the pipeline

$$\text{physical model} \to \text{field} \to \text{summary } S \to \boxed{P(S \mid \theta)} \to \text{infer physics},$$

this chapter supplies the sampling distribution of the observed power spectrum: its mean (through $M_{\ell\ell'}$, T_ℓ and F_ℓ), its covariance (from simulations or a tested approximation), its shape (a χ^2 with an effective number of modes), and the correction for having measured its inverse from a finite set of skies. Chapters 9 and 10 turn that distribution into parameter constraints, Chapter 12 applies the same machinery to the shapes of the map, and Chapter 13 runs it on Planck data.

The instrument on the full sky

concept	operational meaning	used later in
observation model	sky $\to$ beam and pixel window $T_\ell = B_\ell p_\ell \to$ noise; each $a_{\ell m}$ multiplied by T_ℓ, noise added after	Chapters 9 and 13
noise power	uniform white noise: $N_\ell = \sigma^2_{\rm pix}\Omega_{\rm pix}$ at every ℓ; any pattern: $\Omega_{\rm pix}\langle\sigma^2\rangle$ on average	Chapters 10 and 13
debiased estimator	$(\widehat{C}_\ell^{\rm obs} - N_\ell)/T_\ell^2$: unbiased, the full-sky maximum-likelihood estimator	Chapters 9 and 10
cross-spectrum	spectrum between maps with independent noise: no noise bias, at most $\sqrt{2}$ larger error	Chapter 13
Knox formula	$\mathrm{Var}\,\widehat{C}_\ell = 2(C_\ell + N_\ell/T_\ell^2)^2/[(2\ell+1)f_{\rm sky}]$: cosmic variance plus noise, a mode count	Chapter 10
shape of $\widehat{C}_\ell$	shifted, scaled $\chi^2_{2\ell+1}$; skewed at low ℓ, can be negative where noise dominates	Chapter 9
cumulative S/N	$[\sum_\ell \mathrm{SNR}_\ell^2]^{1/2}$ is one over the error of an overall amplitude; saturates where noise equals signal	Chapter 10
Monte Carlo of an estimator	same code on many simulated skies; bias needs $(\sigma/\text{target})^2$ skies, a 10% error bar about 50	Chapters 9 and 12
verification vs research	full sky: simulations confirm formulas; cut sky: simulations are the measurement	Chapters 12 and 13

The cut sky

concept	operational meaning	used later in
window, moments	$W \in [0,1]$ per pixel; $f_{\rm sky}$ = fraction with $W > 0$, w_i = mean of W^i there; $\mathcal{W}_\ell$ = `anafast` of the mask	Chapters 12 and 13
pseudo-C_ℓ	spectrum of the masked map; mixes multipoles; mean $\sum M_{\ell\ell'} F T^2 C + \langle \widetilde{N} \rangle$	Chapter 13
coupling matrix $M_{\ell\ell'}$	from $\mathcal{W}_L$ and $3j$ symbols; depends on the mask only; rows sum to $f_{\rm sky} w_2$	Chapter 13
naive $f_{\rm sky}$ rescaling	divide by $f_{\rm sky} w_2$: exact for flat spectra, biased at peaks and in the damping tail	Chapter 13
null modes	M singular when the patch lacks pairs at some separations; small patches lose information	Chapter 13
MASTER	bin, then invert $K = PMFT^2Q$; unbiased bandpowers; compare theory through bandpower windows $\mathcal{F}_{b\ell}$	Chapters 9 and 13
transfer function F_ℓ	effect of a filter, measured on noise-free simulations; solve in bins	Chapter 13
apodisation	smooth mask edge: less leakage and correlation, fewer modes	Chapters 12 and 13
effective modes $\nu_{\rm eff}$	$2\langle X \rangle^2 / \mathrm{Var}\, X$; model X as $\langle X \rangle \chi^2_{\nu_{\rm eff}} / \nu_{\rm eff}$; non-Gaussian at low ℓ	Chapter 9
ℓ–ℓ' covariance	exact from $\lvert\langle \tilde{a}\tilde{a}^* \rangle\rvert^2$; Efstathiou's $2CC'\Xi(\mathcal{W}^{(2)})$ above the kernel width; Knox underestimates for sharp masks	Chapters 9 and 10
covariance from simulations	sample covariance of n skies; diagonal noisy by $\sqrt{2/(n-1)}$, correlations by $\pm 1/\sqrt{n}$; singular for $p \geq n$	Chapters 9 and 12
Hartlap factor	$(n-p-2)/(n-1)$ times the inverse sample covariance is unbiased	Chapters 9, 10 and 12
simulation budget	parameter variance uncertain by $\sqrt{2/(n-p)}$; $n > 2/\varepsilon^2 + p + 4$ (Taylor et al.)	Chapters 12 and 13
null test	spectrum of a difference of splits; χ^2 against simulations; PTE; catches coherent excesses	Chapter 13
masks and shapes	cuts and holes change counts of hot regions and holes; compute through the same mask on simulations	Chapter 12

Chapter 9

Markov Chains and MCMC: The Biased Random Walker

Where this chapter is going

By the end of Chapter 5 every inference problem had the same shape. The data $\boldsymbol{d}$ have been observed, the parameters $\boldsymbol{\theta}$ are the possible explanations, and Bayes' theorem gives the posterior $p(\boldsymbol{\theta} \mid \boldsymbol{d}) \propto \mathcal{L}(\boldsymbol{\theta})\, p(\boldsymbol{\theta})$, the likelihood times the prior. We can evaluate the right-hand side for any specified $\boldsymbol{\theta}$, but the evidence $Z = p(\boldsymbol{d})$ requires an integral over parameter space. Posterior means, marginal distributions and credible intervals also require integration. Even a modest grid of 100 points along each axis would require 10^{12} evaluations for six cosmological parameters.

A representative sample from the posterior gives another way to approximate these integrals. A histogram estimates a density, an average of a function estimates its expectation, and retaining just the columns of interest gives a sample from their marginal distribution [Kruschke §7.1]. The problem is how to obtain those draws when we can evaluate the density only up to an unknown normalisation, one point at a time.

Sometimes nature answers it for us. An oscillator of fixed energy, looked at at random instants, hands us a sample of its phase-space orbit with no effort at all, because its motion carries it over the orbit at exactly the right rate. That trick fails as soon as the system has more than a few degrees of freedom, and it does not exist for a posterior, which has no motion. So we build the motion ourselves: a random walker that hops from point to point, remembering only where it is now, with hopping rules chosen so that the fraction of time it spends anywhere equals the probability we want. Such a walker is a Markov chain. Three facts make it work: the target distribution is *stationary* (a walker distributed according to it stays so), the walker *forgets where it started* (every starting distribution converges to the same stationary one, at a rate we can compute), and *time averages equal ensemble averages* (the ergodic theorem), so one long walk is as good as many independent draws, only noisier. Metropolis–Hastings then gives a recipe for the hopping rule for *any* target, which is the biased random walker; diagnostics tell us when a finite walk can be trusted, and the same machine produces the posterior of cosmological parameters from a CMB power spectrum (Figure 9.1).

In the pipeline of the book (physical model → field → what matters → summary → sampling distribution → inference) this chapter is the last arrow. Chapters 7 and 8 produced the sampling distribution of a summary, $p(\widehat{C}_\ell \mid \boldsymbol{\theta})$; here that likelihood becomes a cloud of posterior samples from which every constraint is read off.

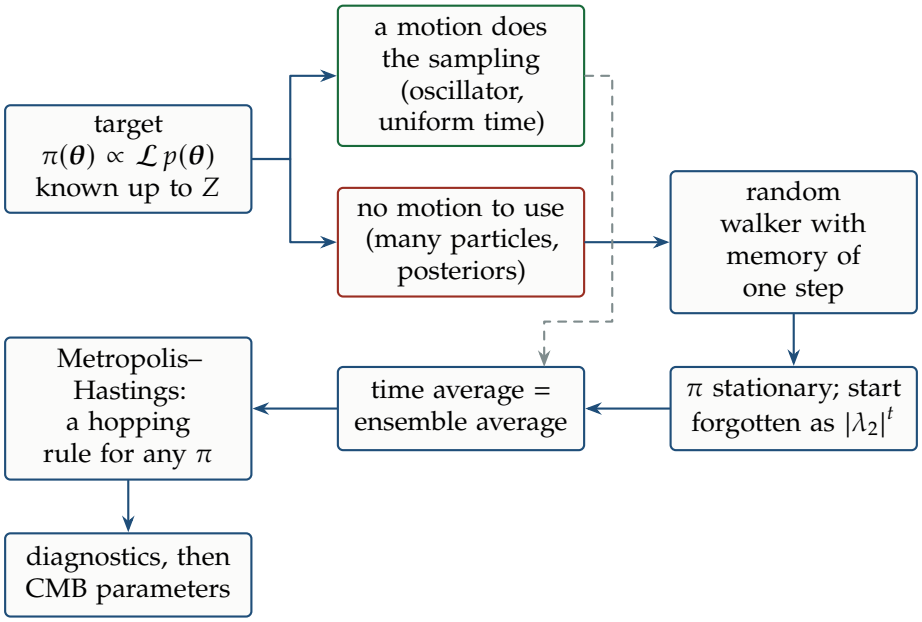


Figure 9.1: The route of the chapter. A distribution we can evaluate but not integrate is sampled directly when some motion already walks over it at the right rate (green), which is the oscillator. When there is no such motion (red) we construct a random walk whose long-run time fractions are the target; the dashed line marks that the oscillator and the Markov chain rest on the same property, equality of time and ensemble averages.

9.1 When the motion samples for us: the oscillator

The easiest sampling problem in physics has already been solved in Chapter 1, and it is worth seeing again why it was so easy, because the reason is exactly what a Markov chain will have to reproduce by force.

The question. A mass on a spring with fixed energy moves as $x(t) = A\cos\omega t$, $p(t) = -m\omega A\sin\omega t$. We look at it at a randomly chosen instant. Where do we find it? At this energy the possible states are the points (x, p) of the energy ellipse in phase space, so the ellipse is the sample space. The motion passes each of its points once per period, so a random instant picks a random point of it. To find the distribution of position, then, we histogram x for a large set of random instants. Before doing so, recall the chain of Section 1.11.3 that predicts what the histogram must look like.

Uniform time. "A randomly chosen instant" is a sampling prescription: every instant of one period $T = 2\pi/\omega$ is equally likely, $P(t)\,\mathrm{d}t = \mathrm{d}t/T$, which is (1.34).

From time to position. The position $x = A\cos\omega t$ changes at the speed $|\mathrm{d}x/\mathrm{d}t| = |v| = \omega\sqrt{A^2 - x^2}$, so the mass spends a time $\mathrm{d}t = \mathrm{d}x/|v|$ in the interval $[x, x + \mathrm{d}x]$, and it passes through that interval twice per period (once going out, once coming back). Adding the two visits,

$$P(x)\,\mathrm{d}x = 2 \times \frac{1}{T}\,\frac{\mathrm{d}x}{|v|} = \frac{2}{T}\,\frac{\mathrm{d}x}{\omega\sqrt{A^2 - x^2}}, \qquad \text{so} \qquad P(x) = \frac{1}{\pi\sqrt{A^2 - x^2}}, \qquad |x| < A, \tag{9.1}$$

the arcsine law, using $T\omega = 2\pi$ in the last step. It is large near the turning points because the mass is slow there.

The same law with no clock. Suppose we did not know that the instant is uniform, and asked instead which states of the energy shell $E < H < E + \Delta E$ are equally likely (the

microcanonical ensemble). The vertical strip of phase space over $[x, x + \mathrm{d}x]$ cuts the shell in two thin patches, at p_+ and p_-. Each patch has width $\mathrm{d}x$ and height Δp, and the height follows from $H = p^2/2m + \frac{1}{2}m\omega^2x^2$: at fixed x, $\mathrm{d}H = (p/m)\,\mathrm{d}p$, so $\Delta p = \Delta E/|v|$. The shell area inside the strip is therefore $2\,\mathrm{d}x\,\Delta E/|v|$, again proportional to $\mathrm{d}x/|v|$, and normalising gives (9.1) a second time.

Why the two agree. The first route averages over the instants of one trajectory; the second averages over all the states of the shell. They give the same answer because the time the motion spends in a patch of the shell, $\mathrm{d}x/|v|$, is proportional to the patch's area. Time average equals ensemble average, and that equality is what *ergodic* will mean below.

The sampler is one line. In units $m = \omega = 1$ the script draws a uniform phase $\theta = \omega t$ and lets the motion do the rest.

Listing 9.1: The oscillator: uniform phases, then the deterministic map to the orbit (`code/ch09/01_oscillator_phase_space.py`)

```
30 # --- the whole sampler: a random instant, then the deterministic motion -------------------
31 theta = rng.uniform(0.0, 2.0 * np.pi, size=N)   # random phase = omega * (random instant)
32 x = A * np.cos(theta)                            # where the motion has carried the
       oscillator
33 p = -A * np.sin(theta)                           # its momentum at that instant
34
35 # --- compare with the exact law, bin by bin (bin probabilities from CDF differences)
       -------
36 edges = np.linspace(-A, A, NBINS + 1)
37 counts, _ = np.histogram(x, bins=edges)
38 expected = N * np.diff(arcsine_cdf(edges, A))
39 chi2 = float(((counts - expected) ** 2 / expected).sum())
```

Line 31 is the only random line, and it chooses *when* we look. Lines 32–33 are the equation of motion: they carry the chosen instant to its point (x, p) of the orbit, the sampled element of the sample space. Lines 36–39 compare the histogram with (9.1) bin by bin, using the exact bin probabilities $F(x_{k+1}) - F(x_k)$ with $F(x) = \frac{1}{2} + \frac{1}{\pi}\arcsin(x/A)$ (Method 1 of Section 1.11.3). With $N = 50\,000$ phases the mismatch is $\chi^2 = 75.9$ over 80 bins, as it should be for a correct law (a χ^2 variable with ν degrees of freedom has mean ν and standard deviation $\sqrt{2\nu} \approx 13$ here), and the average of x^2 over the sampled instants is $0.501\,A^2$ against the exact $A^2/2$ (Figure 9.2).

Why it was so easy. Two things happened together. The equation of motion visits every state of the sample space, and it spends in each part of it a time proportional to that part's probability. So drawing from the distribution reduces to drawing a uniform time, which any random-number generator does. In the language of statistical mechanics the agreement of the two derivations of (9.1), uniform time and microcanonical shell area, says that the average of any quantity over one long trajectory equals its average over the ensemble of all states of the energy shell. That property has a name.

Key idea: Ergodicity: time average = ensemble average

A dynamics is **ergodic** on a set of states if, for every function g of the state, the average of g along one long trajectory equals the average of g over the set with the ensemble's probabilities. Then one trajectory, looked at at random times, is a sample of the ensemble. A Markov chain is useful for exactly this reason.

The hidden assumption, and a counterexample. For one oscillator the property is automatic: the energy shell is a single closed curve and the orbit is all of it. Nothing guarantees it in general,

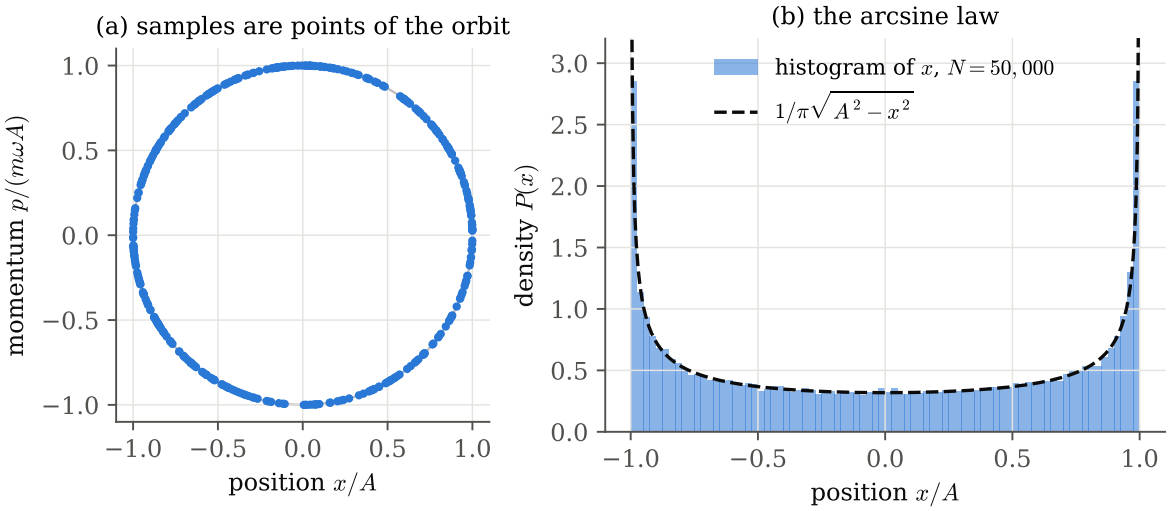


Figure 9.2: (a) Four hundred random instants, drawn as the phase-space points the oscillator occupies at those instants: every sample lies on the energy circle (scaled units), which is the whole sample space. (b) The positions of 5×10^4 such instants follow the arcsine law, piling up at the turning points where the mass moves slowly.

and a slightly bigger system already breaks it. Take two identical oscillators ($m = \omega = 1$) joined by a weak spring of stiffness $\kappa = 0.05$,

$$H = \tfrac{1}{2}(p_1^2 + p_2^2 + x_1^2 + x_2^2) + \tfrac{1}{2}\kappa(x_1 - x_2)^2, \tag{9.2}$$

start them with all the energy E in oscillator 1 ($x_1 = A$, everything else zero), and ask what fraction $u = E_1/E$ of the energy oscillator 1 holds, where $E_1 = \tfrac{1}{2}(p_1^2 + x_1^2)$. The energy shell $H = E$ is now a three-dimensional surface in a four-dimensional phase space. The microcanonical ensemble spreads probability evenly over all of it; the trajectory, started in one particular state, goes where its equations send it. Do the two agree?

The ensemble. Neglect the small spring term for the moment, so that $2H = z_1^2 + z_2^2 + z_3^2 + z_4^2$ with $z = (x_1, p_1, x_2, p_2)$, and the shell is a sphere of radius $\sqrt{2E}$. Uniform points on a sphere in four dimensions are Gaussian vectors rescaled to the right length (a standard Gaussian vector has a density that depends only on $|z|$, so its direction is uniform). Then

$$\begin{aligned} u = \frac{z_1^2 + z_2^2}{z_1^2 + z_2^2 + z_3^2 + z_4^2} &\overset{(1)}{=} \frac{X}{X+Y}, \qquad X, Y \overset{\text{iid}}{\sim} \operatorname{Exp}(\tfrac{1}{2}), \\ \mathrm{P}(u \le c) &\overset{(2)}{=} \int_0^\infty \tfrac{1}{2} e^{-y/2}\, \mathrm{P}\Big(X \le \tfrac{c}{1-c}\, y\Big)\, \mathrm{d}y \\ &\overset{(3)}{=} \int_0^\infty \tfrac{1}{2} e^{-y/2}\big(1 - e^{-cy/2(1-c)}\big)\, \mathrm{d}y \overset{(4)}{=} 1 - (1-c) = c. \end{aligned} \tag{9.3}$$

(1) The length rescaling cancels in the ratio, so we may use the Gaussian vector itself. $z_1^2 + z_2^2$ is a χ^2 variable with 2 degrees of freedom, which is the exponential distribution with rate $\frac{1}{2}$ (Chapter 2), and the two halves are independent.

(2) $X/(X+Y) \le c$ is the same as $X \le cY/(1-c)$; condition on $Y = y$ and average over its density $\frac{1}{2}e^{-y/2}$.

(3) The exponential CDF $\mathrm{P}(X \le a) = 1 - e^{-a/2}$.

(4) $\int_0^\infty \frac{1}{2} e^{-y/2} e^{-cy/2(1-c)} \mathrm{d}y = \frac{1}{2} \big/ \big(\frac{1}{2} + \frac{c}{2(1-c)}\big) = 1 - c$.

So over the energy shell the share of oscillator 1 is *uniform* on $[0,1]$: every split of the energy is equally likely.

The trajectory: two independent oscillators in disguise. The spring couples x_1 and x_2, but it only feels their difference. So use the sum and the difference as coordinates, the **normal modes** $q_\pm = (x_1 \pm x_2)/\sqrt{2}$, with momenta $P_\pm = (p_1 \pm p_2)/\sqrt{2}$. Inverting, $x_1 = (q_+ + q_-)/\sqrt{2}$ and $x_2 = (q_+ - q_-)/\sqrt{2}$, and the same for the momenta. Substituting into (9.2),

$$H \overset{(1)}{=} \tfrac{1}{2}\big(P_+^2 + P_-^2 + q_+^2 + q_-^2\big) + \tfrac{1}{2}\kappa(x_1 - x_2)^2 \overset{(2)}{=} \tfrac{1}{2}\big(P_+^2 + P_-^2 + q_+^2 + q_-^2\big) + \kappa q_-^2$$
$$\overset{(3)}{=} \tfrac{1}{2}\big(P_+^2 + \omega_+^2 q_+^2\big) + \tfrac{1}{2}\big(P_-^2 + \omega_-^2 q_-^2\big), \qquad \omega_+ = 1, \quad \omega_- = \sqrt{1+2\kappa}. \tag{9.4}$$

(1) $x_1^2 + x_2^2 = \tfrac{1}{2}\big[(q_+ + q_-)^2 + (q_+ - q_-)^2\big] = q_+^2 + q_-^2$, the cross terms cancelling; the same for $p_1^2 + p_2^2$.
(2) $x_1 - x_2 = \sqrt{2}\,q_-$, so $\tfrac{1}{2}\kappa(x_1 - x_2)^2 = \kappa q_-^2$.
(3) Collect the q_-^2 terms, $\tfrac{1}{2}q_-^2 + \kappa q_-^2 = \tfrac{1}{2}(1+2\kappa)q_-^2$.

The Hamiltonian is a sum of two uncoupled oscillators, of frequencies ω_+ and ω_-, and each keeps its own energy forever.

Solve each mode. The start $x_1 = A$, $x_2 = 0$, all momenta zero, gives $q_+(0) = q_-(0) = A/\sqrt{2}$ and zero initial momenta, so $q_\pm(t) = (A/\sqrt{2})\cos\omega_\pm t$. Back in the original coordinate,

$$x_1(t) \overset{(1)}{=} \tfrac{A}{2}\big(\cos\omega_+ t + \cos\omega_- t\big) \overset{(2)}{=} A\cos\bar\omega t\,\cos\tfrac{1}{2}\Delta\omega\, t, \qquad \bar\omega = \tfrac{1}{2}(\omega_+ + \omega_-), \quad \Delta\omega = \omega_- - \omega_+. \tag{9.5}$$

(1) $x_1 = (q_+ + q_-)/\sqrt{2}$.
(2) The sum-to-product identity $\cos a + \cos b = 2\cos\frac{a+b}{2}\cos\frac{a-b}{2}$, with $a = \omega_- t$, $b = \omega_+ t$.

Name the two factors. $\cos\bar\omega t$ is the fast *carrier*, the ordinary oscillation at nearly the natural frequency. $\cos\frac{1}{2}\Delta\omega\, t$ is the slow *beat*: for a weak spring $\Delta\omega \approx \kappa$ is small, so it changes little during one carrier cycle and acts as a slowly varying amplitude $a(t) = A\cos\frac{1}{2}\Delta\omega\, t$.

The energy share. Over one carrier cycle oscillator 1 is a plain oscillator of amplitude $a(t)$ and frequency $\bar\omega \approx 1$: $x_1 \simeq a\cos\bar\omega t$ and $p_1 \simeq -a\sin\bar\omega t$, the derivative of the slow factor being smaller by $\Delta\omega$. Then $E_1 = \frac{1}{2}(p_1^2 + x_1^2) \simeq \frac{1}{2}a^2(\sin^2\bar\omega t + \cos^2\bar\omega t) = \frac{1}{2}a^2$, the carrier dropping out entirely. With $E \simeq \frac{1}{2}A^2$ (the spring's share is of order κ),

$$u(t) = \frac{E_1}{E} \simeq \cos^2\phi, \qquad \phi = \tfrac{1}{2}\Delta\omega\, t. \tag{9.6}$$

During each beat the energy sloshes completely from one oscillator to the other and back (Figure 9.3a). At a random instant the beat phase ϕ is uniform, and the density of $u = \cos^2\phi$ follows from the change-of-variables rule of Section 1.11:

$$P(u) \overset{(1)}{=} 4 \times \frac{1}{2\pi}\left|\frac{\mathrm{d}\phi}{\mathrm{d}u}\right| \overset{(2)}{=} \frac{4}{2\pi}\,\frac{1}{2\big|\cos\phi\sin\phi\big|} \overset{(3)}{=} \frac{1}{\pi\sqrt{u(1-u)}}, \qquad 0 < u < 1. \tag{9.7}$$

(1) ϕ is uniform on $[0, 2\pi)$ and $\cos^2\phi$ takes each value in $(0,1)$ at four phases per period, which contribute equally.
(2) $\mathrm{d}u/\mathrm{d}\phi = -2\cos\phi\sin\phi$.
(3) $\big|\cos\phi\sin\phi\big| = \sqrt{\cos^2\phi\,(1-\cos^2\phi)} = \sqrt{u(1-u)}$.

It is the arcsine law again, now for the energy share: the trajectory lingers where the energy is almost all in one oscillator, because that is where the beat turns round.

The verdict. The script `02_two_oscillators.py` keeps the spring term, draws 2×10^5 exactly uniform points of the true shell (in scaled normal-mode coordinates H is again a sum of four squares) and 2×10^5 random instants of the trajectory. Both give the same mean share, 0.489 and 0.489 (slightly below $\frac{1}{2}$ because the spring holds a little energy). But the trajectory spends a fraction 0.205 of its time with $u<0.1$, while the shell has only 0.102 of its states there, close to the values $\frac{2}{\pi}\arcsin\sqrt{0.1}=0.205$ and 0.1 of (9.7) and (9.3); the second moments $\langle u^2\rangle$ are 0.357 and 0.318 (Figure 9.3b). Time average and ensemble average disagree.

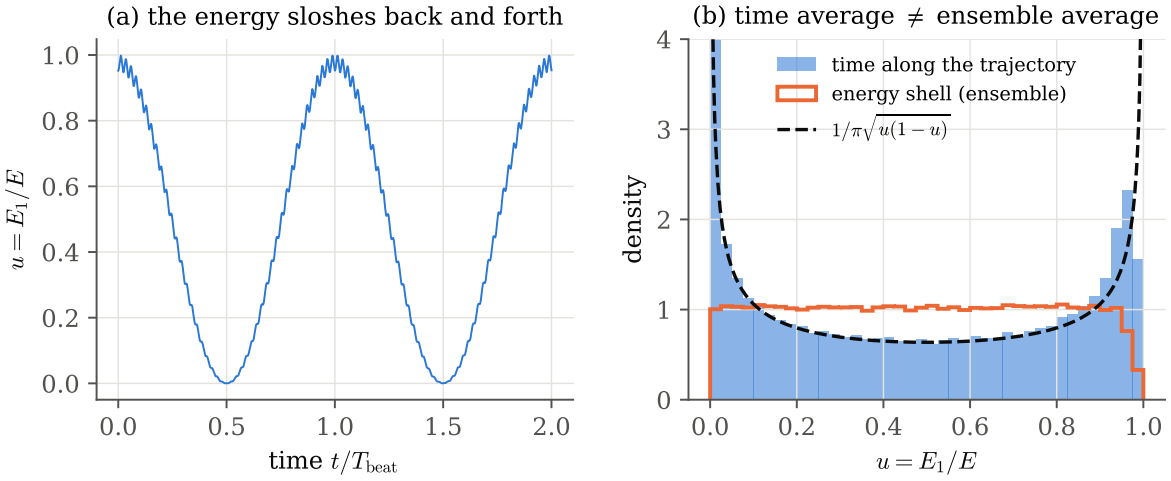


Figure 9.3: Two weakly coupled oscillators, started with all the energy in the first. (a) The first oscillator's share of the energy beats between 1 and 0 over two beat periods. (b) The time the trajectory spends at each share (blue) follows the arcsine curve and crowds at the ends; the energy shell itself (orange outline) gives every share the same weight. One trajectory is not a fair sample of the shell.

What went wrong is that the motion has more conserved quantities than the energy. Each normal-mode energy is constant, so the trajectory is trapped on a two-dimensional torus inside the three-dimensional shell and can never reach the states that the ensemble counts. Every linear system has this disease, one conserved energy per normal mode, which is why harmonic crystals do not thermalise on their own and why it took nonlinear couplings, collisions, or contact with a heat bath to justify statistical mechanics. The lesson is a sharp version of Chapter 1's: a motion samples a distribution only if it is ergodic for that distribution, and that has to be shown, not assumed.

Why this cannot be the general method. Even an ergodic motion is rarely usable as a sampler.

1. *Many degrees of freedom.* For N particles the energy shell is a surface of dimension $6N-1$. We cannot parametrise it to draw uniform points, and integrating the equations of motion for long enough to cover it, if the dynamics covers it at all, is as expensive as the problem we started with.
2. *The canonical ensemble.* A system in contact with a heat bath at temperature T is described by $\rho(x,p)\propto e^{-H(x,p)/kT}$. For one oscillator this is a product of Gaussians and trivial to sample. For N interacting particles the normalisation $\int e^{-H/kT}\,\mathrm{d}^{3N}x\,\mathrm{d}^{3N}p$ has no closed form, there is no orbit to follow (the bath changes the energy), and there is no direct sampler. Metropolis, the Rosenbluths and the Tellers met exactly this problem for a fluid of hard discs, and solved it with a random walk through configurations [36].
3. *Posteriors have no dynamics.* $p(\boldsymbol{\theta}\mid\boldsymbol{d})\propto\mathcal{L}(\boldsymbol{\theta})p(\boldsymbol{\theta})$ is a function on parameter space, known up to Z. Nothing moves.

Direct, independent sampling is hard for a second reason too: the generic methods, rejection sampling and the inverse CDF, need either an envelope that wastes exponentially many draws as the dimension grows or a CDF we cannot write down [Lambert §12.7].

So we keep the idea and change the motion. Instead of a deterministic orbit we use a *random* walker, and we choose its hopping rule so that the fraction of time it spends anywhere equals the target probability there. The walker will need four properties. Its next step should depend only on where it is now, so that a step is cheap to compute. The target should be an equilibrium of its motion. It should forget where it started. And one long walk should be a fair sample, time averages equal to ensemble averages. The first property defines a Markov chain; the other three are theorems about Markov chains.

What kind of random walker? The simplest one, which steps left or right with equal chance wherever it stands, is an *unbiased* walker; it has no preference for any place, so in the long run it spreads evenly and visits every region equally often (Section 9.7). That samples only a flat distribution. What we want instead is a *biased* walker, whose hopping rule depends on the target so that it stays longer where the target is larger and hurries through where it is small, in exactly the proportion that makes its fractions of time equal to the target probabilities.

9.2 A walker with a memory of one step

An elected politician lives on a chain of seven islands, numbered $\theta = 1, \ldots, 7$ from west to east [Kruschke §7.2.1]. He wants to visit each island in proportion to its population, but he does not know the total population, nor even how many islands there are. His advisers can obtain just two population counts at a time, one from the mayor of his current island and the other from the mayor of a neighbouring island under consideration.

Kruschke's question. The story is Kruschke's way of asking the question every posterior sampler faces. Can a walker who never sees a distribution as a whole, only the ratio of its values at two neighbouring points, still end up visiting every point in proportion to the distribution? For a posterior the values are $\mathcal{L}(\boldsymbol{\theta})p(\boldsymbol{\theta})$, which we can compute at any point, and the total is the evidence Z, which we cannot (Chapter 9). If the politician can do without his total, a sampler can do without Z.

His setup. Every evening the politician flips a fair coin to propose the island to the east or to the west. If the proposed island has more people than the current one, he goes. If it has fewer, he goes only with probability $P_{\text{proposed}}/P_{\text{current}}$, decided by a spinner marked from 0 to 1; otherwise he stays another day. Beyond the ends lie two imaginary islands, 0 and 8, where nobody lives, so a proposal to go there is always refused [Kruschke §7.2.2]. In one table:

ingredient	Kruschke	meaning
islands	$\theta = 1, \ldots, 7$	the states the walker can be in
populations	$P(\theta) = \theta$	the target, known only up to its total 28
proposal	fair coin: east or west	where to try next, chosen blindly
acceptance	spinner: $\min\bigl(1, P_{\text{proposed}}/P_{\text{current}}\bigr)$	the only place the target enters
edges	islands 0 and 8 have $P = 0$	proposals off the ends are refused

The questions. Suppose he starts on island 4.

1. On day t, what is the probability that he is on island θ?

2. Over a long term of office, what fraction of his days does he spend on each island?

Kruschke's claim is that the second fraction is $\theta/28$, the island's share of the total population of $1+2+\cdots+7=28$, so that the politician reaches his goal without ever learning the total. His rule is the simplest case of the Metropolis algorithm. To check the claim we first need to know what kind of object the politician's itinerary is, and how its distribution moves from day to day; the reason the claim holds is his Eq. (7.2), which needs only that description and is derived after the transition matrix.

9.2.1 Stochastic processes and the Markov property

The itinerary $X_1, X_2, X_3, \ldots$ is a sequence of random variables, one per day, and the days are not independent: where he sleeps tomorrow depends on where he slept tonight. A collection of random variables $\{X_t : t \in \mathcal{T}\}$ indexed by time is a **stochastic process**. The values X_t lie in a set $\mathcal{X}$, the **state space**, and $\mathcal{T}$ is the **index set** [Wasserman p. 381]. Both may be discrete or continuous.

- An iid sequence $X_1, X_2, \ldots$ is a stochastic process, the simplest one, with no dependence at all.
- The daily weather, $\mathcal{X} = \{\text{sunny}, \text{cloudy}\}$, $\mathcal{T} = \{1, 2, \ldots\}$: discrete states, discrete time.
- A stock price or a detector's baseline voltage recorded continuously: a continuous state (for all practical purposes) and continuous time.
- The empirical CDF $\widehat{F}_n(x) = \frac{1}{n}\sum_i I(X_i \le x)$ of Chapter 3, viewed as a function of $x \in [0,1]$: for each x it is a random variable, so the whole curve is a process with continuous state and continuous "time" x.

Any joint distribution can be built up one variable at a time by the product rule of Chapter 1,

$$f(x_1,\ldots,x_n) = f(x_1)\,f(x_2\mid x_1)\,f(x_3\mid x_1,x_2)\cdots f(x_n\mid x_1,\ldots,x_{n-1}) = \prod_{i=1}^{n} f(x_i \mid \text{past}_i), \tag{9.8}$$

where $\text{past}_i = (x_1,\ldots,x_{i-1})$. In general the ith factor needs the whole past. The politician's rule needs much less: his decision tonight uses the population of the island he is on and of the island he proposes, and nothing about how he got there. The past influences the future only through the present.

Working definition 9.1 (Markov chain)

A discrete-time process $X_0, X_1, X_2, \ldots$ with values in a countable state space $\mathcal{X}$ is a **Markov chain** if, for every t and every state x,

$$\mathrm{P}(X_{t+1} = x \mid X_0, X_1, \ldots, X_t) = \mathrm{P}(X_{t+1} = x \mid X_t). \tag{9.9}$$

In words: given the present state [the value of X_t, one number], the next state is independent of the earlier history [the values $X_0,\ldots,X_{t-1}$]. The chain has a **memory of one step**. *Operational test:* if you can write a simulation whose step $t \to t+1$ reads only the current state (plus fresh random numbers), the process is a Markov chain.

For a Markov chain every conditional in (9.8) keeps only its last argument, and the joint law collapses to a product of one-step factors,

$$f(x_0,x_1,\ldots,x_n) = f(x_0)\,f(x_1\mid x_0)\,f(x_2\mid x_1)\cdots f(x_n\mid x_{n-1}). \tag{9.10}$$

As a graph, each variable has a single parent, the one before it (Figure 9.4a) [Wasserman p. 383]. Two remarks place the definition. First, independence is the special case in which the factor $f(x_{t+1} \mid x_t)$ does not depend on x_t at all, so the Markov property is the mildest generalisation of independence: dependence, but only on the immediate past [CasellaBerger §5.8.5]. Second, the name records the *amnesia* of the stepping rule. A chain whose step depends on the last m states is called a Markov chain of order m [Lambert §12.9]; it becomes an ordinary (first-order) chain if we take as the state the block of the last m values, so nothing new is needed for it.

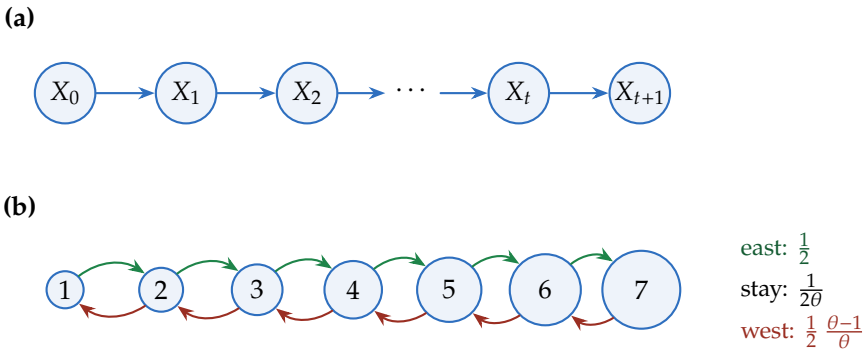


Figure 9.4: (a) A Markov chain as a graph: each variable depends directly only on the one before it. (b) The politician's chain. Circles are islands, drawn larger for more populous ones; upper (green) arrows are moves east, always accepted when proposed, and lower (red) arrows are moves west, accepted with the population ratio. Island 7's eastward proposal and island 1's westward one are always refused, which adds to the probability of staying.

9.2.2 The transition matrix

A Markov chain is **homogeneous** if the one-step probabilities $\mathrm{P}(X_{t+1} = j \mid X_t = i)$ do not change with t. All the chains of this chapter are homogeneous: the politician uses the same rule every night. A homogeneous chain on a finite state space is then described completely by one square table of numbers.

Working definition 9.2 (transition matrix)

For a homogeneous chain on states $1, \ldots, K$ the **transition probabilities** are $p_{ij} = \mathrm{P}(X_{t+1} = j \mid X_t = i)$, and the $K \times K$ **transition matrix** P has p_{ij} in row i [the state we are in] and column j [the state we may move to]. Every entry is non-negative and every row sums to one, $\sum_j p_{ij} = 1$, because from state i the chain must go somewhere. Each row is a probability mass function: the forecast for tomorrow given today. A matrix with these two properties is called a **stochastic matrix**.

Example: Wasserman's weather. Suppose sunny and cloudy days follow each other as a Markov chain, the weather tomorrow depending only on today's [Wasserman p. 385]. A typical transition matrix is

$$P = \begin{pmatrix} 0.4 & 0.6 \\ 0.8 & 0.2 \end{pmatrix} \qquad \text{(rows and columns: sunny, cloudy)}, \tag{9.11}$$

so after a sunny day there is a 60% chance of cloud, and after a cloudy day an 80% chance of sun. Each row sums to one; the columns need not.

Example: a random walk with absorbing barriers. On the states $1, \ldots, N$, flip a coin with $\mathrm{P}(\text{heads}) = p$, $q = 1 - p$, and step right on heads and left on tails, except that a walker who

reaches 1 or N stays there forever [Wasserman p. 384]. The transition matrix has $p_{i,i+1} = p$ and $p_{i,i-1} = q$ in the interior rows and $p_{11} = p_{NN} = 1$:

$$P = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ q & 0 & p & \cdots & 0 & 0 \\ 0 & q & 0 & \ddots & 0 & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & q & 0 & p \\ 0 & 0 & \cdots & 0 & 0 & 1 \end{pmatrix}.$$

It is the gambler's ruin: a gambler with $i-1$ units of capital betting one unit at a time until he is broke or has reached a capital of $N-1$ units.

Example: the politician's matrix. From island θ the politician moves east only if the coin proposes east (probability $\frac{1}{2}$) and the move is accepted; since $P(\theta+1) > P(\theta)$ it always is. He moves west only if the coin proposes west and the spinner accepts, which happens with probability $P(\theta-1)/P(\theta) = (\theta-1)/\theta$. So for an interior island

$$\begin{aligned} p_{\theta,\theta+1} &\overset{(1)}{=} \tfrac{1}{2}\min\!\left(1, \tfrac{\theta+1}{\theta}\right) = \tfrac{1}{2}, \\ p_{\theta,\theta-1} &\overset{(2)}{=} \tfrac{1}{2}\min\!\left(1, \tfrac{\theta-1}{\theta}\right) = \frac{\theta-1}{2\theta}, \\ p_{\theta\theta} &\overset{(3)}{=} 1 - \tfrac{1}{2} - \frac{\theta-1}{2\theta} = \frac{1}{2\theta}. \end{aligned} \tag{9.12}$$

(1) The probability of a move is (probability of proposing it) × (probability of accepting it); here the proposed island is larger, so acceptance is certain.
(2) The proposed island is smaller, so the spinner accepts with the population ratio.
(3) The rejected proposals keep him where he is, and the row must sum to one.

At the ends the formulas change only through the empty islands: island 7 never moves east, so $p_{77} = 1 - \frac{6}{14} = \frac{4}{7}$, and island 1 never moves west, so $p_{11} = 1 - \frac{1}{2} = \frac{1}{2}$. Written out,

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & \frac{1}{6} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & \frac{3}{8} & \frac{1}{8} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & \frac{2}{5} & \frac{1}{10} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & \frac{5}{12} & \frac{1}{12} & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 & \frac{3}{7} & \frac{4}{7} \end{pmatrix}. \tag{9.13}$$

This is the matrix Kruschke builds symbolically in his Eq. (7.3) [Kruschke §7.2.5]. The script builds it in four lines, which are the rule itself:

Listing 9.2: The politician's transition matrix, built from the proposal and the acceptance rule (`code/ch09/03_island_hopping.py`)

```
# --- transition matrix: propose a neighbour (prob 1/2 each), accept with min(1, ratio) ------
P = np.zeros((K, K))
for i in range(K):
    for j in (i - 1, i + 1):
        if 0 <= j < K:                          # islands outside 1..7 have population 0: never accepted
            P[i, j] = 0.5 * min(1.0, pop[j] / pop[i])
    P[i, i] = 1.0 - P[i].sum()                  # rejected proposals: stay put
```

Line 27 is "propose with probability $\frac{1}{2}$, accept with $\min(1, \text{ratio})$"; line 26 skips the empty islands, whose proposals are never accepted; line 28 puts every refused proposal on the diagonal.

His key equation, in our symbols. Kruschke's notation and ours name the same things differently. His island θ is our state i; his relative population $P(\theta)$ is our target π_i, up to the factor 28; his "probability of moving from θ to $\theta+1$" is our matrix entry $p_{\theta,\theta+1}$. His Eq. (7.2) compares the two directions across one bridge between neighbouring islands. Dividing the first line of (9.12) by the second line written for island $\theta+1$,

$$\frac{p_{\theta,\theta+1}}{p_{\theta+1,\theta}} = \frac{\frac{1}{2}}{\frac{1}{2}\,\theta/(\theta+1)} = \frac{P(\theta+1)}{P(\theta)} :$$

the ratio of the two hopping probabilities equals the ratio of the two populations, and the total 28 never appears. Why this ratio forces the long-run fractions to be $\theta/28$ is the subject of detailed balance (Section 9.4.2), where the same equation is rearranged into a statement about traffic across each bridge.

9.2.3 How a distribution moves: $p_{t+1} = p_t P$

We now answer the first question. Write the probabilities of being in each state on day t as a *row* vector $p_t = (p_t(1), \ldots, p_t(K))$, with $p_t(j) = \mathrm{P}(X_t = j)$. Today's distribution and the transition matrix fix tomorrow's: to be in j tomorrow the walker must be somewhere today and then jump to j, and the possible "somewheres" are disjoint, so

$$\begin{aligned} p_{t+1}(j) &\overset{(1)}{=} \sum_i \mathrm{P}(X_t = i,\ X_{t+1} = j) \overset{(2)}{=} \sum_i \mathrm{P}(X_t = i)\ \mathrm{P}(X_{t+1} = j \mid X_t = i) \\ &\overset{(3)}{=} \sum_i p_t(i)\, p_{ij}, \qquad \text{i.e.} \qquad p_{t+1} = p_t\, P. \end{aligned} \tag{9.14}$$

(1) Law of total probability over today's state i (Chapter 1).
(2) Product rule.
(3) Definitions of p_t and of p_{ij}; homogeneity lets us drop the t from p_{ij}. The sum over i of a row vector times the ith row of P is the vector–matrix product.

One step of the chain is one multiplication by P from the right. A stochastic matrix maps probability vectors to probability vectors: the entries of $p_t P$ are sums of non-negative terms, and they add up to $\sum_j \sum_i p_t(i) p_{ij} = \sum_i p_t(i) \sum_j p_{ij} = 1$.

Watch out: Rows or columns?

Probabilists write distributions as row vectors and multiply by P on the right, so the *rows* of P sum to one. Much of the physics literature (master equations, quantum channels) writes column vectors and uses the transpose. Before using a formula from elsewhere, check which index is "from" and which is "to". The stationary distribution is a *left* eigenvector of P in our convention.

Example: the politician's first two nights. On day 1 he is on island 4 for sure, $p_1 = (0,0,0,1,0,0,0)$. Multiplying a vector with a single 1 in position 4 by P picks out row 4 of (9.13):

$$p_2 = p_1 P = \left(0,\ 0,\ \tfrac{3}{8},\ \tfrac{1}{8},\ \tfrac{1}{2},\ 0,\ 0\right) = (0, 0, 0.375, 0.125, 0.5, 0, 0),$$

the three bars of Kruschke's panel $t = 2$ [Kruschke §7.2.3]. On day 3,

$$p_3(2) \overset{(1)}{=} \tfrac{3}{8}\cdot\tfrac{1}{3} = \tfrac{1}{8}, \qquad p_3(3) \overset{(2)}{=} \tfrac{3}{8}\cdot\tfrac{1}{6} + \tfrac{1}{8}\cdot\tfrac{3}{8} = \tfrac{7}{64}, \qquad p_3(4) \overset{(3)}{=} \tfrac{3}{8}\cdot\tfrac{1}{2} + \tfrac{1}{8}\cdot\tfrac{1}{8} + \tfrac{1}{2}\cdot\tfrac{2}{5} = \tfrac{129}{320},$$
$$p_3(5) \overset{(4)}{=} \tfrac{1}{8}\cdot\tfrac{1}{2} + \tfrac{1}{2}\cdot\tfrac{1}{10} = \tfrac{9}{80}, \qquad p_3(6) \overset{(5)}{=} \tfrac{1}{2}\cdot\tfrac{1}{2} = \tfrac{1}{4}, \qquad p_3(1) = p_3(7) = 0.$$

(1) Island 2 can be reached on day 3 only from island 3 (moving west, probability $p_{32} = \frac{1}{3}$).
(2) From island 3 by staying ($p_{33} = \frac{1}{6}$) or from island 4 by moving west ($p_{43} = \frac{3}{8}$).
(3) From 3 eastwards, from 4 by staying, from 5 westwards ($p_{54} = \frac{2}{5}$).
(4) From 4 eastwards and from 5 by staying ($p_{55} = \frac{1}{10}$).
(5) From 5 eastwards.

These are 0.125, 0.1094, 0.4031, 0.1125, 0.25, and they add up to one. After two nights he can be at most two islands from where he started, so islands 1 and 7 still have probability zero.

What p_t means, and how to simulate a chain. p_t is not what one politician does; one politician is on exactly one island each day. It is the histogram of where a large number of independent politicians would be on day t, all following the same rule from the same starting distribution [Wasserman p. 386]. Simulating one of them needs only the initial distribution p_0 and P: draw X_0 from p_0; if it came out i, draw X_1 from row i of P; if that came out j, draw X_2 from row j; and so on. In code, each draw from a row is one uniform number compared with the row's cumulative sums (the inverse-CDF method of Chapter 6); that is all `walk` and `step_many` in `lib_markov.py` do.

What we compute. Two things, for the same days Kruschke drew. The exact crowd histogram comes from repeated multiplication, $p_t = p_1 P^{t-1}$, with P the matrix (9.13) and p_1 the spike on island 4. The simulated one comes from running 20 000 independent politicians from island 4 and counting where they sleep on day t. The two are drawn together in Figure 9.5 for days 2, 3, 5, 13 and 99. The early distributions are lopsided, with a bulge over the starting island; by day 13 the shape is already close to a straight incline.

Side by side with Kruschke. Kruschke's Fig. 7.3 shows the same sequence of histograms as bars, read off his matrix [Kruschke §7.2.3]. At $t = 2$ there is something exact to compare; at later days the natural comparison is with the incline $\theta/28$ itself, measured by the total-variation distance $\frac{1}{2}\sum_\theta |p_t(\theta) - \theta/28|$ (the largest difference in probability that any set of islands can show, defined in Section 9.5).

quantity	Kruschke, Fig. 7.3	ours
$p_2(3)$, $p_2(4)$, $p_2(5)$	0.375, 0.125, 0.5	0.375, 0.125, 0.5
day 13, distance from $\theta/28$	close to the incline (by eye)	0.048
day 99, distance from $\theta/28$	indistinguishable (by eye)	1.5×10^{-6}

The day-2 values agree exactly, as they must: both are row 4 of the same matrix. What the calculation adds is the size of the remaining difference on later days, a number instead of a picture; how fast it shrinks is the subject of Section 9.5.

Many steps at once: the Chapman–Kolmogorov equations. Let $p_{ij}(n) = \mathrm{P}(X_{m+n} = j \mid X_m = i)$ be the probability of going from i to j in exactly n steps, and P_n the matrix of these **n-step**

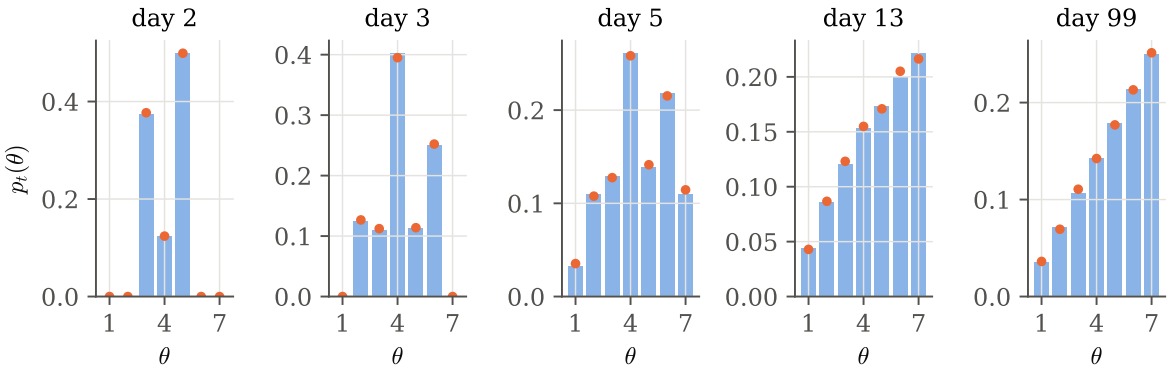


Figure 9.5: Where the politician is likely to be, starting from island 4. Bars are the exact $p_t = p_1 P^{t-1}$; dots are the fractions of 2×10^4 simulated politicians. The distribution spreads from its starting spike, overshoots, and settles into the incline $p(\theta) = \theta/28$, the same for every later day.

transition probabilities [Wasserman p. 385]. To go from i to j in $m + n$ steps the chain must be somewhere, say k, after the first m of them:

$$\begin{aligned} p_{ij}(m+n) = \mathrm{P}(X_{m+n} = j \mid X_0 = i) &\overset{(1)}{=} \sum_k \mathrm{P}(X_{m+n} = j,\ X_m = k \mid X_0 = i) \\ &\overset{(2)}{=} \sum_k \mathrm{P}(X_{m+n} = j \mid X_m = k,\ X_0 = i)\ \mathrm{P}(X_m = k \mid X_0 = i) \\ &\overset{(3)}{=} \sum_k \mathrm{P}(X_{m+n} = j \mid X_m = k)\ \mathrm{P}(X_m = k \mid X_0 = i) \overset{(4)}{=} \sum_k p_{ik}(m)\, p_{kj}(n). \end{aligned} \tag{9.15}$$

(1) Law of total probability over the intermediate state, inside the conditional probability given $X_0 = i$.
(2) Product rule in its conditional form, $\mathrm{P}(A, B \mid C) = \mathrm{P}(B \mid C)\,\mathrm{P}(A \mid B, C)$.
(3) The Markov property: once $X_m = k$ is known, X_0 is irrelevant for the future. (For $n > 1$ this uses (9.9) repeatedly, one step at a time.)
(4) Homogeneity: the probabilities depend only on the number of steps.

The last line is matrix multiplication, so $P_{m+n} = P_m P_n$. Since $P_1 = P$, we get $P_2 = P^2$, $P_3 = P^3$ and in general

$$P_n = P^n, \qquad p_n = p_0 P^n. \tag{9.16}$$

The second statement follows from the first by one more use of total probability, $p_n(j) = \sum_i \mathrm{P}(X_n = j \mid X_0 = i)\,\mathrm{P}(X_0 = i) = \sum_i p_0(i)\, p_{ij}(n)$, or simply by applying (9.14) n times. Every question about a homogeneous finite chain is therefore a question about the powers of one matrix, and the long-run behaviour is the behaviour of P^n as $n \to \infty$, which is a question about eigenvalues.

Example: the weather a week ahead. For Wasserman's weather (9.11), starting on a sunny day, $p_0 = (1, 0)$: $p_1 = (0.4, 0.6)$, $p_2 = (0.4 \cdot 0.4 + 0.6 \cdot 0.8,\ \ldots) = (0.64, 0.36)$, $p_3 = (0.544,\ \ldots)$, and $p_6(\text{sunny}) = 0.5732$. The forecast swings above and below a value near 0.571 with a shrinking amplitude. Section 9.5 shows that the value is $4/7$ and that the swing shrinks by the factor -0.4 each day.

9.2.4 The long run, seen in one trajectory

The second question is about a single politician over a long time, not about a crowd on one day. Figure 9.6a shows the first 500 days of one simulated trip on a logarithmic time axis, like Kruschke's Fig. 7.2: the politician wanders down to island 1 early on, then drifts east and spends most of his time among the larger islands. Counting his days over a trip of 100 000 days (Figure 9.6b), the fraction spent on each island differs from $\theta/28$ by at most 0.0025. The time he spends on an island tracks its population.

Two different facts. Figures 9.5 and 9.6 look alike but answer different questions. The first is a histogram over many politicians on one day t: it approaches $\theta/28$ as t grows, which is convergence of p_t (Section 9.5). The second is a histogram over the days of one politician: it approaches $\theta/28$ as the trip grows, which is the ergodic theorem (Section 9.6). Neither implies the other on sight: a crowd can settle while each member keeps a lopsided personal history, and one long history can be fair even when the crowd never settles (Section 9.4.1). Kruschke stops at showing that $\theta/28$ is *stable*, leaving the rest as "far beyond" his book [Kruschke §7.2.5]; both facts are proved in Sections 9.5 and 9.6.

The lesson. Reproducing the island figures turns a picture that "looks settled" into two separate limits, one over walkers and one over time, and only the second is what a single MCMC run relies on.

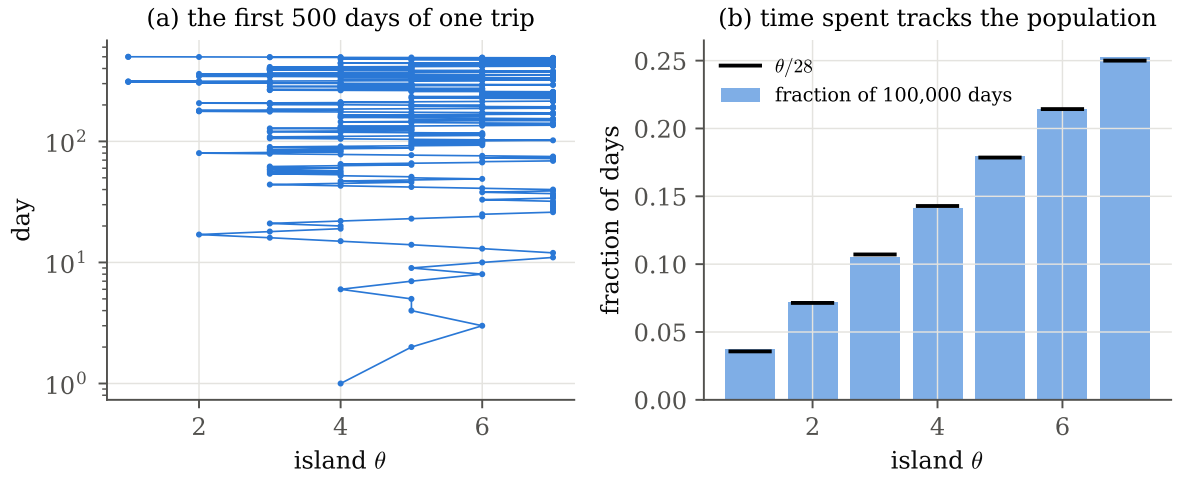


Figure 9.6: (a) One politician's first 500 days, with the day on a logarithmic axis so that the early wandering is visible. (b) The fraction of 10^5 days he spends on each island (bars) against the population shares $\theta/28$ (black ticks).

9.2.5 Reading the transition matrix off a history

So far P was given and we predicted the walk. The inverse question also arises: we observe one history $x_0, x_1, \ldots, x_n$ of a chain (a weather record, a sequence of detector states) and want to infer P. By (9.10) the probability of the observed history is a product of one-step factors, one per observed transition. Collecting equal factors, with n_{ij} the number of observed transitions from i to j,

$$\mathcal{L}(p_0, P) = p_0(x_0) \prod_{r=1}^{n} p_{x_{r-1}x_r} = p_0(x_0) \prod_{i,j} p_{ij}^{n_{ij}} \tag{9.17}$$

[Wasserman p. 394]. There is only one observation of the starting state, so p_0 cannot be learned; we maximise over P. Take the logarithm of (9.17),

$$\ln \mathcal{L} = \ln p_0(x_0) + \sum_i \Big(\sum_j n_{ij} \ln p_{ij} \Big).$$

The first term does not contain P at all, so it plays no part in the maximisation. The rest is a sum of one bracket per row i, and each bracket contains only the entries of its own row. The constraints are also one per row, $\sum_j p_{ij} = 1$. So changing row i changes only its own bracket, and each row can be maximised on its own, with its own Lagrange multiplier μ_i (the method of Chapter 3 for maximising under a constraint):

$$0 \overset{(1)}{=} \frac{\partial}{\partial p_{ij}} \Big[\sum_k n_{ik} \ln p_{ik} - \mu_i \Big(\sum_k p_{ik} - 1 \Big) \Big] = \frac{n_{ij}}{p_{ij}} - \mu_i \quad \Longrightarrow \quad p_{ij} = \frac{n_{ij}}{\mu_i},$$

$$1 \overset{(2)}{=} \sum_j p_{ij} = \frac{n_i}{\mu_i} \quad \Longrightarrow \quad \hat{p}_{ij} = \frac{n_{ij}}{n_i}, \qquad n_i = \sum_j n_{ij}. \tag{9.18}$$

(1) The part of $\ln \mathcal{L}$ that involves row i, plus the constraint for that row; set the derivative to zero. Only the $k = j$ terms depend on p_{ij}, and $\partial(n_{ij} \ln p_{ij})/\partial p_{ij} = n_{ij}/p_{ij}$, so the condition is $n_{ij}/p_{ij} = \mu_i$, that is $p_{ij} = n_{ij}/\mu_i$. It is a maximum: the second derivative of each term, $-n_{ij}/p_{ij}^2$, is negative, so the bracket is concave in the row and its only stationary point on the constraint is the top.

(2) Impose the constraint to find μ_i: it equals n_i, the number of departures from state i.

The estimate is the obvious one: of the n_i times we left state i, the fraction that went to j.

How uncertain is it? Fix a row i and look only at the n_i nights the chain spent in state i. By the Markov property each departure from i is a fresh draw from row i of P, whatever happened before, so the n_i destinations are independent draws from the same pmf $(p_{i1}, \ldots, p_{iK})$. Counting how many went to each j is then exactly the multinomial setting of Section 2.3, with n_i trials and cell probabilities p_{ij}:

$$\operatorname{Var}(n_{ij}) = n_i p_{ij}(1 - p_{ij}), \qquad \operatorname{Cov}(n_{ij}, n_{i\ell}) = -n_i p_{ij} p_{i\ell} \quad (j \neq \ell).$$

Dividing the counts by n_i gives the estimate $\hat{p}_{ij} = n_{ij}/n_i$, whose variance is $p_{ij}(1 - p_{ij})/n_i$, so its error bar is $\sqrt{\hat{p}_{ij}(1 - \hat{p}_{ij})/n_i}$. Different rows use different nights, so their errors are uncorrelated. For large n_i the multinomial counts are close to Normal (the central limit theorem applied to the sum of n_i indicator vectors), which gives $\sqrt{n_i}(\hat{p}_{ij} - p_{ij}) \rightsquigarrow \mathcal{N}$ with variance $p_{ij}(1 - p_{ij})$ [Wasserman p. 394, Thm 23.31].

One thing differs from ordinary multinomial counting: n_i is not fixed in advance but is itself random, since it depends on how often the chain happened to visit i. For an ergodic chain the fraction of time spent in i settles to π_i (Section 9.6), so n_i grows like $n\pi_i$ as the record lengthens, every row gets more and more departures, and the estimate is consistent with the error bar above shrinking as $1/\sqrt{n\pi_i}$.

Example: a thousand days of weather. The script `08_markov_mle.py` simulates 1000 days of the three-state weather chain used throughout Section 9.5 (sunny, cloudy, rainy; true rainy row $(0.6, 0, 0.4)$). The record visits sun, cloud and rain 518, 342 and 140 times, and gives for the rainy row $\hat{p}_{RS} = 0.664 \pm 0.04$ and $\hat{p}_{RR} = 0.336 \pm 0.04$, the error bars being $\sqrt{\hat{p}(1 - \hat{p})/n_R}$. Rain is the rarest state, so its row is the least well measured. Repeating the experiment on 2000 independent histories, the standardised error $(\hat{p}_{RS} - p_{RS})/\sqrt{p_{RS}(1 - p_{RS})/n_R}$ has mean 0.01 and standard deviation 1.013, and its histogram is a standard Normal (Figure 9.7): the multinomial error bar is right even though the visits to rain arrive in correlated clumps.

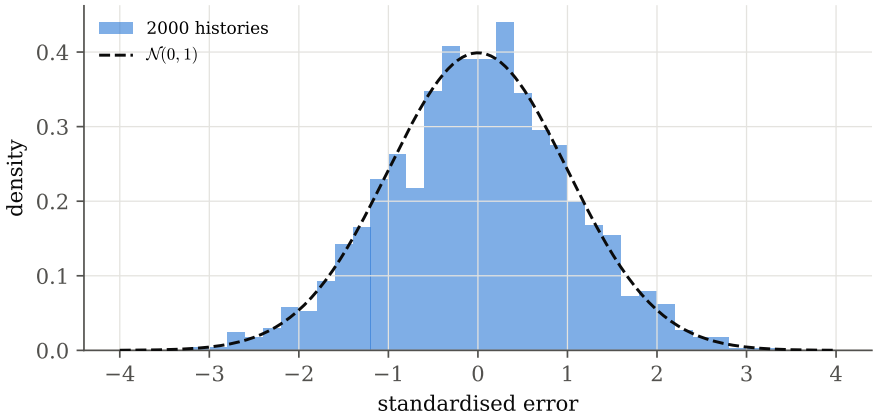


Figure 9.7: The error of the estimated rain-to-sun probability, in units of its multinomial standard error, over 2000 simulated thousand-day records. The histogram follows the standard Normal curve.

9.3 Who can reach whom: classifying the states

Not every Markov chain settles down. A fish that swims into a net never visits the rest of the ocean again; a rollercoaster car runs round its track forever and is at the top of the loop only at regular instants [Lambert §13.6.2]. Both are perfectly good Markov chains, but neither has a long-run histogram that is independent of where it started: the fish's histogram depends on when it was caught, and the car's distribution at time t keeps rotating instead of settling. Wasserman's Fig. 23.2 shows the same contrast for two simulated chains, one wandering off for good, one fluctuating about a fixed level [Wasserman p. 384]. Before proving that a chain converges we need words for the ways in which it can fail to: some states cannot be reached from others, some states are left and never revisited, and some are revisited only in a fixed rhythm.

The questions. For a given transition matrix P:

1. From state i, can the chain ever get to state j?
2. Having left i, is it certain to come back, and how long does that take on average?
3. Can it come back at any time, or only at multiples of some period?

9.3.1 Communication and irreducibility

State i **reaches** j (or j is **accessible** from i), written $i \to j$, if $p_{ij}(n) > 0$ for some $n \geq 0$; we count $n = 0$, so every state reaches itself. If $i \to j$ and $j \to i$ the two states **communicate**, $i \leftrightarrow j$ [Wasserman p. 387].

Communication is an equivalence relation: it is reflexive ($n = 0$), symmetric by definition, and transitive, because if $i \to j$ in m steps and $j \to k$ in n steps then

$$p_{ik}(m+n) \overset{(1)}{=} \sum_{\ell} p_{i\ell}(m)\, p_{\ell k}(n) \overset{(2)}{\geq} p_{ij}(m)\, p_{jk}(n) \overset{(3)}{>} 0.$$

(1) Chapman–Kolmogorov (9.15).
(2) All terms are non-negative; keep only the term $\ell = j$.

(3) Both factors are positive by assumption.

An equivalence relation cuts the state space into disjoint **communicating classes**, $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2 \cup \cdots$, with two states in the same class exactly when they communicate [Wasserman p. 387, Thm 23.12].

Working definition 9.3 (irreducible, closed, absorbing)

A chain is **irreducible** if all its states communicate [one single class, every state reachable from every other in some number of steps]. A set of states is **closed** if the chain can never leave it once inside [$p_{ij} = 0$ for every i in the set and j outside it]. A closed set consisting of one state is an **absorbing state** [$p_{ii} = 1$]. *Operational test:* draw the transition graph, an arrow $i \to j$ for every nonzero p_{ij}. The chain is irreducible if and only if from every node there is a path of arrows to every other node.

Example: four states, three classes. Wasserman's Ex. 23.13 has $\mathcal{X} = \{1, 2, 3, 4\}$ and

$$P = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} & 0 & 0 \\ \frac{2}{3} & \frac{1}{3} & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Reading the graph (Figure 9.8a): 1 and 2 send arrows only to each other, so $\{1, 2\}$ is a class, and it is closed. State 3 sends arrows everywhere, but nothing comes back to 3 from any other state, so $\{3\}$ is a class on its own, and not closed. State 4 has $p_{44} = 1$: $\{4\}$ is a closed class and 4 is absorbing. A walker that starts in 3 stays there for a geometric number of steps (mean $1/(1 - \frac{1}{4}) = \frac{4}{3}$) and then falls for good into $\{1, 2\}$ or into 4. The odds of the two exits follow from row 3. Each night in state 3 the walker leaves with probability $1 - \frac{1}{4} = \frac{3}{4}$; it goes to 1 or 2 with probability $\frac{1}{4} + \frac{1}{4} = \frac{1}{2}$ and to 4 with probability $\frac{1}{4}$. Given that it leaves, the chances are therefore $(\frac{1}{2})/(\frac{3}{4}) = \frac{2}{3}$ for $\{1, 2\}$ and $(\frac{1}{4})/(\frac{3}{4}) = \frac{1}{3}$ for 4, the same on every night, so these are also the chances of the eventual exit. Where it ends up depends on luck, so this chain has no unique long-run histogram.

The politician's chain (9.13) is irreducible: every island sends an arrow to each neighbour, so any island reaches any other by a walk along the chain. The absorbing random walk of Section 9.2.2 is not: its classes are $\{1\}$, $\{N\}$ (both absorbing) and the interior $\{2, \ldots, N-1\}$, which is not closed.

9.3.2 Coming back: recurrence and transience

Start the chain in state i. Will it ever come back? State i is **recurrent** (or persistent) if

$$\mathrm{P}(X_n = i \text{ for some } n \geq 1 \mid X_0 = i) = 1,$$

and **transient** otherwise [Wasserman p. 388]. The test is the expected number of visits.

Key idea: Recurrence ⇔ infinitely many expected returns

State i is recurrent if and only if $\sum_n p_{ii}(n) = \infty$, and transient if and only if $\sum_n p_{ii}(n) < \infty$ [Wasserman p. 388, Thm 23.15].

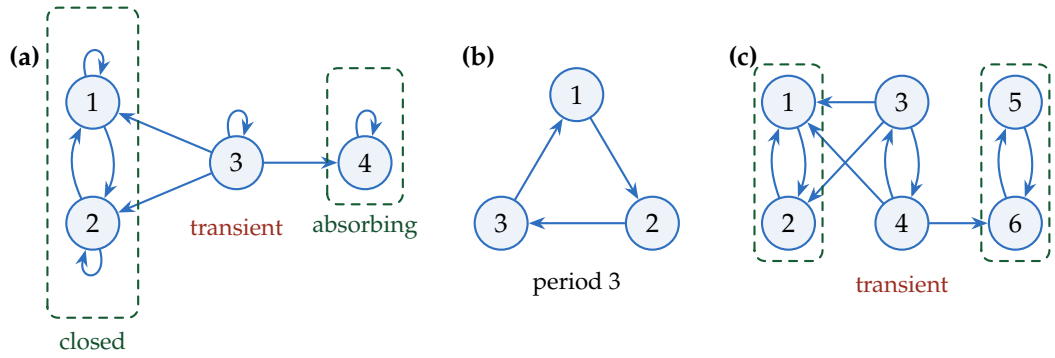


Figure 9.8: Transition graphs, an arrow for every nonzero transition probability. (a) Wasserman's Ex. 23.13: $\{1,2\}$ is a closed class, 4 is absorbing, and 3 leaks into both and is never re-entered. (b) A deterministic cycle: the chain returns to each state only every third step. (c) Ex. 23.28 (self-loops not drawn): two closed classes $\{1,2\}$ and $\{5,6\}$, and states 3 and 4 that eventually drain into one of them.

Why. Let $I_n = 1$ if $X_n = i$ and 0 otherwise, so that $Y = \sum_{n\geq 0} I_n$ counts the visits to i. Its mean, for a chain started at i, is

$$\mathbb{E}(Y \mid X_0 = i) \overset{(1)}{=} \sum_{n=0}^{\infty} \mathbb{E}(I_n \mid X_0 = i) \overset{(2)}{=} \sum_{n=0}^{\infty} \mathrm{P}(X_n = i \mid X_0 = i) \overset{(3)}{=} \sum_{n=0}^{\infty} p_{ii}(n). \tag{9.19}$$

(1) Linearity of expectation, which holds for infinite sums of non-negative terms.
(2) The mean of an indicator is the probability of its event.
(3) Definition of the n-step probability.

Now count the visits differently. Let a_i be the probability of ever returning. Each time the chain is at i, the Markov property makes the future a fresh copy of the chain started at i, so it returns again with probability a_i, independently of the past. If $a_i = 1$ it returns, returns again, and so on forever: $Y = \infty$ with probability one, and (9.19) diverges. If $a_i < 1$, each visit is the last with probability $1 - a_i$, so Y is geometric:

$$\mathrm{P}(Y = k \mid X_0 = i) = a_i^{k-1}(1 - a_i), \qquad \mathbb{E}(Y \mid X_0 = i) \overset{(1)}{=} \sum_{k\geq 1} k\, a_i^{k-1}(1 - a_i) \overset{(2)}{=} \frac{1 - a_i}{(1 - a_i)^2} = \frac{1}{1 - a_i} < \infty.$$

(1) Mean of the geometric distribution with "success" (no return) probability $1 - a_i$.
(2) $\sum_k k a^{k-1} = \frac{\mathrm{d}}{\mathrm{d}a} \sum_k a^k = 1/(1-a)^2$ for $|a| < 1$.

So a finite sum means transience and an infinite one means recurrence. For a finite chain there is a shortcut for the general facts [Wasserman p. 389, Thm 23.16]: recurrence and transience are properties of a whole class (if i is recurrent and $i \leftrightarrow j$, then j is recurrent, and likewise for transient); a finite chain has at least one recurrent state (it runs forever on finitely many states, so some state is visited infinitely often); and in a finite irreducible chain every state is recurrent. The state space always splits as $\mathcal{X} = \mathcal{X}_T \cup \mathcal{X}_1 \cup \mathcal{X}_2 \cup \cdots$, the transient states $\mathcal{X}_T$ plus closed irreducible classes of recurrent states (the **decomposition theorem**, [Wasserman p. 389, Thm 23.17]). In Figure 9.8a, $\mathcal{X}_T = \{3\}$ and the closed classes are $\{1,2\}$ and $\{4\}$.

Example: the random walk on the integers. A walker on $\{\ldots, -1, 0, 1, \ldots\}$ steps $+1$ with probability p and -1 with $q = 1 - p$ [Wasserman p. 389, Ex. 23.18]. All states communicate, so either all are recurrent or all are transient; test state 0. A return can happen only after an even number of steps, $2n$, with exactly n steps each way, in any order, so $p_{00}(2n) = \binom{2n}{n} p^n q^n$.

Stirling's formula $n! \simeq n^n e^{-n}\sqrt{2\pi n}$ gives the size of the binomial coefficient:

$$\binom{2n}{n} = \frac{(2n)!}{(n!)^2} \overset{(1)}{=} \frac{(2n)^{2n}e^{-2n}\sqrt{4\pi n}}{n^{2n}e^{-2n}\,2\pi n} \overset{(2)}{=} \frac{4^n}{\sqrt{\pi n}}, \qquad \text{so} \qquad p_{00}(2n) \simeq \frac{(4pq)^n}{\sqrt{\pi n}}. \tag{9.20}$$

(1) Stirling for $(2n)!$ in the numerator and for each $n!$ in the denominator.
(2) $(2n)^{2n}/n^{2n} = 4^n$ and $\sqrt{4\pi n}/(2\pi n) = 1/\sqrt{\pi n}$.

Since $pq \le \frac{1}{4}$ with equality only at $p = \frac{1}{2}$, two cases appear. For $p \neq \frac{1}{2}$, $4pq < 1$ and the terms decrease geometrically, the sum converges, and the walk is transient: it drifts away in the direction of the bias and returns only finitely often. For $p = \frac{1}{2}$, $p_{00}(2n) \simeq 1/\sqrt{\pi n}$, whose sum diverges like $\sqrt{n}$, and the walk is recurrent. Stirling is already good at $n = 10$, where exact over approximate is 0.9876 (Figure 9.9a).

The two cases can also be checked against the geometric law of the visits. The expected number of visits to 0, counting the start, is the full sum (9.19), and the binomial series $\sum_{n\ge0}\binom{2n}{n}x^n = (1-4x)^{-1/2}$ (valid for $|x| < \frac{1}{4}$) sums it in closed form with $x = pq$:

$$\sum_{n\ge0} p_{00}(2n) \overset{(1)}{=} (1-4pq)^{-1/2} \overset{(2)}{=} \big((p-q)^2\big)^{-1/2} \overset{(3)}{=} \frac{1}{|p-q|}.$$

(1) The binomial series with $x = pq$.
(2) $1 - 4pq = (p+q)^2 - 4pq = p^2 - 2pq + q^2 = (p-q)^2$, using $p + q = 1$.
(3) The positive square root of a square is the absolute value.

The same mean is $1/(1-a)$ by the geometric law of the visits found above, so $1/(1-a) = 1/|p-q|$, and the probability of ever returning is $a = 1 - |p-q|$. For $p = 0.6$: $x = 0.24$, $(1-0.96)^{-1/2} = 0.04^{-1/2} = 5$ visits on average including the start, so $\sum_{n\ge1} p_{00}(2n) = 4$ after removing the $n = 0$ term, and the probability of ever returning is $a = 1 - \frac{1}{5} = 0.8 = 1 - |0.6 - 0.4|$. Simulating 10 000 walkers for 10 000 steps (Figure 9.9b), the fraction that has come back at least once is 0.801 for $p = 0.6$, and 0.9916 for $p = \frac{1}{2}$, still creeping towards one.

How fast the symmetric walker comes home. For $p = \frac{1}{2}$ there is an exact statement: the probability of no return by step $2n$ equals the probability $p_{00}(2n)$ of being at the origin at step $2n$. A quick check at $n = 1$: the walker fails to return at step 2 with probability $\frac{1}{2}$, and $p_{00}(2) = \binom{2}{1}\frac{1}{4} = \frac{1}{2}$. The general proof uses the series above once more. Write $u_n = p_{00}(2n)$ and f_n for the probability that the *first* return happens at step $2n$. Being at 0 at step $2n \ge 2$ means a first return at some step $2k \le 2n$ followed, by the Markov property, by being back at 0 after $2n - 2k$ more steps, so $u_n = \sum_{k=1}^{n} f_k u_{n-k}$. Multiply by s^n and sum over n; with $U(s) = \sum_{n\ge0} u_n s^n$ and $F(s) = \sum_{n\ge1} f_n s^n$ this reads $U = 1 + FU$, and the binomial series with $x = s/4$ gives $U(s) = (1-s)^{-1/2}$. Then

$$\sum_{n\ge0} \mathrm{P}(T_{00} > 2n)\, s^n \overset{(1)}{=} \sum_{n\ge0}\Big(1 - \sum_{k=1}^{n} f_k\Big)s^n \overset{(2)}{=} \frac{1-F(s)}{1-s} \overset{(3)}{=} \frac{1}{(1-s)\,U(s)} \overset{(4)}{=} (1-s)^{-1/2} = U(s).$$

(1) No return by step $2n$ means the first return comes later than step $2n$.
(2) $\sum_n s^n = 1/(1-s)$, and $\sum_n s^n \sum_{k\le n} f_k = F(s)/(1-s)$ (each f_k appears in every term with $n \ge k$).
(3) $U = 1 + FU$ gives $1 - F = 1/U$.
(4) $U(s) = (1-s)^{-1/2}$.

Two power series that are equal have equal coefficients, so $\mathrm{P}(T_{00} > 2n) = p_{00}(2n) \simeq 1/\sqrt{\pi n}$ by (9.20), which predicts the returned fraction 0.992 seen in the simulation.

How long until it comes back? Recurrence says that the return happens; it does not say how soon. The **recurrence time** $T_{ii} = \min\{n > 0 : X_n = i\}$ (for $X_0 = i$) has distribution

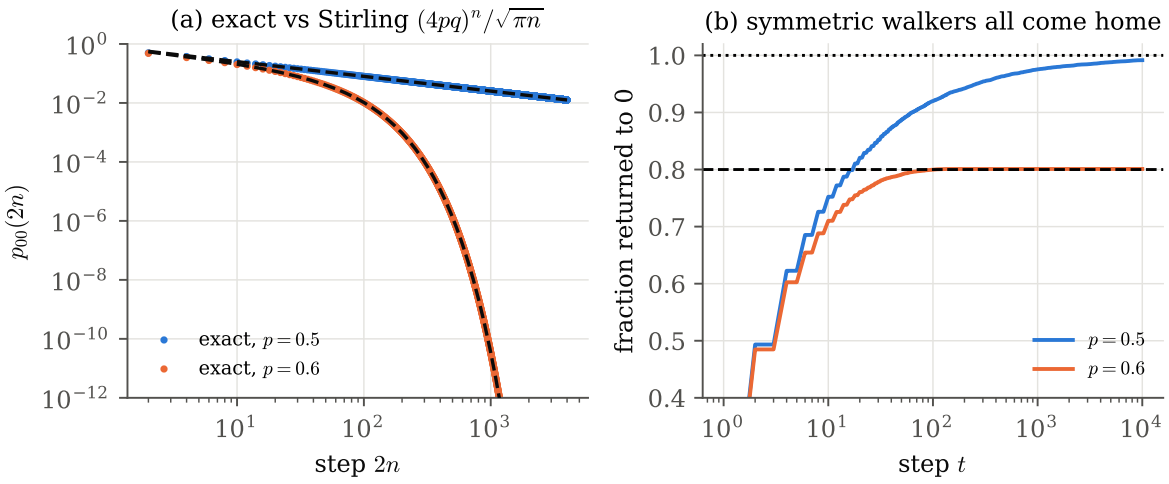


Figure 9.9: (a) The probability of being back at the origin after $2n$ steps: exact (dots) and Stirling's approximation (dashed). The symmetric walk decays only as $1/\sqrt{n}$; a bias of 0.6 makes it decay geometrically. (b) The fraction of 10^4 walkers that have returned to the origin at least once: all of them eventually for the symmetric walk, 80% for the biased one (dashed line).

$f_{ii}(n) = \mathrm{P}(X_1 \neq i, \ldots, X_{n-1} \neq i, X_n = i \mid X_0 = i)$ and the **mean recurrence time** is $m_i = \mathbb{E}(T_{ii}) = \sum_n n f_{ii}(n)$ [Wasserman p. 390]. A recurrent state with $m_i = \infty$ is **null recurrent**; otherwise it is **positive recurrent**.

A mean from the tail. For a waiting time T taking values $1, 2, 3, \ldots$ there is a convenient way to compute the mean. Write T as a count of the steps $k = 0, 1, \ldots, T-1$ before it, $T = \sum_{k\geq 0} I(T > k)$, with I the indicator (1 if true, 0 if not)„ and take expectations term by term (all terms are non-negative):

$$\mathbb{E}(T) = \sum_{k\geq 0} \mathrm{P}(T > k).$$

The mean is finite exactly when the tail probabilities $\mathrm{P}(T > k)$ have a finite sum.

The symmetric walk is null. The symmetric walk on the integers is the standard null-recurrent chain: every walker comes home, but the probability of not yet being home after $2n$ steps is $p_{00}(2n) \simeq 1/\sqrt{\pi n}$, as just shown. Each value $\mathrm{P}(T_{00} > 2n)$ appears twice in the tail sum (for $k = 2n$ and $k = 2n+1$, since returns happen only at even steps), so $\mathbb{E}(T_{00}) \approx \sum_n 2/\sqrt{\pi n}$, which diverges like $\sqrt{n}$. At a null-recurrent state the walker is certain to return, yet the chance of finding it at home at any given late time goes to zero, $p_{ii}(n) \to 0$, because it spends longer and longer excursions away [Wasserman p. 390, Lemma 23.19]; the symmetric walk shows it directly, with $p_{00}(2n) \simeq 1/\sqrt{\pi n}$. We will meet the null walker again as the free diffusion of Section 9.7.

A finite chain has no null states. Take a state i in a closed irreducible class of a finite chain. From every state j of the class some path leads to i, and we may take it of length at most M, the number of states (a shortest path never visits a state twice). Its probability is positive, and since there are only finitely many j, the smallest of these probabilities is some $\varepsilon > 0$. So wherever the walker stands, the chance of reaching i within the next M steps is at least ε, and by the Markov property this holds afresh for each block of M steps. Missing i in k blocks in a row therefore has probability at most $(1-\varepsilon)^k$:

$$\mathrm{P}(T_{ii} > kM) \leq (1-\varepsilon)^k, \qquad m_i = \sum_{n\geq 0} \mathrm{P}(T_{ii} > n) \leq M \sum_{k\geq 0} (1-\varepsilon)^k = \frac{M}{\varepsilon} < \infty,$$

using the tail formula and the fact that $\mathrm{P}(T_{ii} > n)$ can only decrease with n, so each block of M terms is at most M times its first term. A geometric tail leaves no room for the ever-

longer excursions of a null state: in a finite chain every recurrent state is positive recurrent [Wasserman p. 390, Lemma 23.20].

9.3.3 Rhythm: periodicity

The deterministic cycle $1 \to 2 \to 3 \to 1$ (Figure 9.8b),

$$P = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix},$$

started in state 1 is in state 1 at times $0, 3, 6, 9, \ldots$ and never in between [Wasserman p. 390]. The **period** of state i is $d(i) = \gcd\{n \geq 1 : p_{ii}(n) > 0\}$, the greatest common divisor of the possible return times; i is **periodic** if $d(i) > 1$ and **aperiodic** if $d(i) = 1$. Period is a class property: communicating states have the same period (Lemma 23.21). Two quick tests cover most chains met in practice. A single self-loop ($p_{ii} > 0$) makes the period 1, since then 1 is a possible return time. A walk that alternates between two sets, such as ± 1 steps on a ring of even length, has period 2: it returns only after an even number of steps. The politician's chain is aperiodic because every island has $p_{\theta\theta} = 1/(2\theta) > 0$; refused proposals act as self-loops. A "tilted" die that always moves to a neighbouring face round the ring of six faces [Lambert §12.10] has period 2, and we will see in Section 9.6 what that does and does not spoil.

Example: four lakes on a ring. Lambert's marooned fisherman camps each evening beside one of four lakes A, B, C, D joined in a ring [Lambert §13.3]. He wants to fish each lake in proportion to its fish stock, so he uses the politician's rule: flip a coin to propose the clockwise or the anticlockwise neighbour, go there if it holds more fish, and otherwise go with probability equal to the ratio of the stocks. If the four stocks were equal, every proposal would be accepted and he would alternate between $\{A, C\}$ and $\{B, D\}$: period 2. Take stocks in the ratio $1 : 2 : 4 : 3$ (our numbers). Then

$$P = \begin{pmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & 0 \\ 0 & \frac{1}{4} & \frac{3}{8} & \frac{3}{8} \\ \frac{1}{6} & 0 & \frac{1}{2} & \frac{1}{3} \end{pmatrix}$$

(rows and columns A, B, C, D; for example $p_{CD} = \frac{1}{2} \cdot \frac{3}{4}$ and $p_{CC} = 1 - \frac{1}{4} - \frac{3}{8}$). Lake A has no self-loop, since both neighbours are bigger and every proposal is accepted, but B, C and D have refused proposals, hence $p_{ii} > 0$ and period 1. Communicating states share their period, so the whole chain is aperiodic: unequal stocks break the rhythm that equal stocks would impose. The chain is also irreducible, so it is ergodic, and $\pi = (0.1, 0.2, 0.4, 0.3)$ balances every pair of neighbours (for instance $\pi_A p_{AB} = 0.05 = \pi_B p_{BA}$).

Working definition 9.4 (ergodic state and chain)

A state is **ergodic** if it is recurrent, non-null [finite mean return time] and aperiodic [return times have greatest common divisor 1]. A chain is ergodic if all its states are [Wasserman p. 390, Def. 23.22]. *Operational test for a finite chain:* it is ergodic if and only if it is irreducible and aperiodic, and then some power P^m has every entry strictly positive.

Why some power of P is strictly positive. The operational test in the box rests on a small fact about numbers. Fix a state i and let $R_i = \{n \geq 1 : p_{ii}(n) > 0\}$ be its possible return times.

Return times can be added. If m and n are in R_i, so is $m+n$: by Chapman–Kolmogorov (9.15), $p_{ii}(m+n) = \sum_k p_{ik}(m)p_{ki}(n) \geq p_{ii}(m)\,p_{ii}(n) > 0$, keeping only the term $k = i$ (go round one loop, then the other).

Coprime loops fill in every large number. Aperiodic means $\gcd(R_i) = 1$. Take the simplest case, return times 3 and 5. Adding them in all combinations gives $3a + 5b$ for whole numbers $a, b \geq 0$, and these include $8 = 3+5$, $9 = 3\cdot 3$, $10 = 5\cdot 2$, and then every larger number, because adding 3 to $8, 9, 10$ gives $11, 12, 13$, and so on. Only $1, 2, 4, 7$ are missed. The same holds for any set of positive integers with greatest common divisor 1 that is closed under addition (a classical fact of elementary number theory, of which the $3, 5$ case is the pattern): it contains every integer beyond some N_i. So the walker can return to i in *exactly* n steps for every $n \geq N_i$.

Then go anywhere. For any pair i, j, irreducibility gives a path from i to j of some length ℓ_{ij} with $p_{ij}(\ell_{ij}) > 0$. Waste $N_i + k$ steps looping at i, then take the path:

$$p_{ij}(N_i + k + \ell_{ij}) \geq p_{ii}(N_i + k)\,p_{ij}(\ell_{ij}) > 0 \qquad \text{for every } k \geq 0.$$

One power for all pairs. There are finitely many pairs, so take m larger than every $N_i + \ell_{ij}$. Then $p_{ij}(m) > 0$ for all i, j, which is $P^m > 0$ entrywise. Finally, a finite irreducible chain is automatically positive recurrent (Section 9.3.2, the geometric-tail argument), so irreducible and aperiodic is all the test needs.

Example: six states. Wasserman's Ex. 23.28 has

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{3}{4} & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{4} \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

[Wasserman p. 392]. Rows 1–2 and 5–6 send probability only within $\{1, 2\}$ and $\{5, 6\}$, so these are closed irreducible classes. States 3 and 4 communicate with each other, but the path $3 \to 4 \to 6$ leads into $\{5, 6\}$, from which there is no way back, so 3 and 4 are transient. Every state has a self-loop, so all are aperiodic: 1, 2, 5 and 6 are ergodic, 3 and 4 are transient (Figure 9.8c). The chain as a whole is not irreducible, and its long-run behaviour depends on which closed class swallows the walker.

Irreducible is not the same as ergodic. Some texts say "irreducible (or ergodic)" [Lambert §13.6.2], but these are different conditions. The deterministic 3-cycle is irreducible but periodic, so its distribution at time t never settles; the symmetric random walk on the integers is irreducible and recurrent but null, so it has no equilibrium at all. Irreducibility says the chain can get everywhere; ergodicity adds that it comes back often enough and without a fixed rhythm. The vocabulary of classification, collected:

property	meaning	example
transient	left for good with positive probability	state 3 of Figure 9.8a
null recurrent	returns surely, mean return time infinite	symmetric walk on $\mathbb{Z}$
positive recurrent	returns, finite mean return time	any state of a finite irreducible chain
periodic ($d > 1$)	returns only at multiples of d	deterministic 3-cycle
ergodic	positive recurrent and aperiodic	every island of the politician

9.4 Equilibrium: stationary distributions and detailed balance

Go back to the crowd of independent politicians of Section 9.2.3, all following the island-hopping rule. Suppose that tonight the crowd happens to be spread over the islands in proportion to their populations, a fraction $\theta/28$ on island θ. Tomorrow some of them move east, some west, some stay. Kruschke's "climactic implication" is that the crowd is still spread as $\theta/28$ [Kruschke §7.2.5]: individuals move, the histogram does not. This is what equilibrium means for a random walk, and it is the property that makes the target distribution something a chain can settle into.

The question. Is there a distribution of the walkers that one step of the chain leaves unchanged? If so, is there only one, and how do we find it without solving for it?

Working definition 9.5 (stationary distribution)

A probability vector $\pi = (\pi_1, \dots, \pi_K)$ [non-negative entries that sum to one] is a **stationary** (or invariant) distribution of the chain if

$$\pi = \pi P, \qquad \text{that is} \qquad \pi_j = \sum_i \pi_i \, p_{ij} \quad \text{for every } j. \tag{9.21}$$

It is a left eigenvector of P with eigenvalue 1 [Wasserman p. 390, Def. 23.23]. *Operational recipe:* solve the linear system $\pi(P - \mathbb{1}) = 0$ together with $\sum_i \pi_i = 1$, or compute the left eigenvector of P for eigenvalue 1 and normalise it (`stationary` in `lib_markov.py`).

If the chain starts in a stationary distribution it stays there forever. With $X_0 \sim \pi$, (9.16) gives $p_1 = \pi P = \pi$, $p_2 = \pi P^2 = (\pi P)P = \pi P = \pi$, and by induction $p_n = \pi P^n = \pi$ for all n [Wasserman p. 391]. Each individual walker keeps moving; only the distribution of the crowd is frozen.

A stationary vector always exists for a finite chain, because the eigenvalue 1 always occurs: the rows of P sum to one, which says $P\mathbf{1} = \mathbf{1}$ for the column vector $\mathbf{1}$ of ones, so 1 is an eigenvalue of P, and a matrix and its transpose have the same eigenvalues, so there is also a left eigenvector. That this left eigenvector can be chosen with non-negative entries, and that it is unique when the chain is irreducible, are the content of the Perron–Frobenius theorem; Section 9.5 proves both by a direct argument.

Balance of flows. Think of probability as a conserved fluid on the states, with $\pi_i p_{ij}$ the amount that flows from i to j in one step. Write the stationarity condition (9.21) for state j in two ways, separating in each the part that stays put at j:

$$\pi_j \overset{(1)}{=} \sum_i \pi_i \, p_{ij} = \pi_j \, p_{jj} + \sum_{i \neq j} \pi_i \, p_{ij},$$
$$\pi_j \overset{(2)}{=} \pi_j \sum_k p_{jk} = \pi_j \, p_{jj} + \sum_{k \neq j} \pi_j \, p_{jk}.$$

(1) Stationarity: what is at j tomorrow is what arrives from everywhere, including from j itself; take the $i = j$ term out of the sum.

(2) Row j of P sums to one, so multiplying π_j by that sum changes nothing; take the $k = j$ term out.

The left sides are equal, so the right sides are. Cancel the common term $\pi_j p_{jj}$, the probability that sits at j and stays:

$$\underbrace{\sum_{i \neq j} \pi_i \, p_{ij}}_{\text{flow into } j} \;=\; \underbrace{\sum_{k \neq j} \pi_j \, p_{jk}}_{\text{flow out of } j} \qquad \text{for every } j. \tag{9.22}$$

Stationarity is **global balance**: at every state the inflow equals the outflow, like Kirchhoff's current law at every node of a circuit. It does not say that the flow along each wire vanishes. That stronger statement is detailed balance, below.

Example: Wasserman's weather. For (9.11) the first component of $\pi = \pi P$ reads $\pi_S = 0.4\pi_S + 0.8\pi_C$, so $0.6\pi_S = 0.8\pi_C$, and with $\pi_S + \pi_C = 1$, $\pi = (\frac{4}{7}, \frac{3}{7})$. In the long run four days in seven are sunny. For a general two-state chain with $a = p_{12}$ and $b = p_{21}$ the same two lines give

$$\pi = \Big(\frac{b}{a+b}, \; \frac{a}{a+b} \Big), \tag{9.23}$$

the flow balance $\pi_1 a = \pi_2 b$ in disguise: the state that is harder to leave gets more weight.

Example: a three-state weather chain. Our running example of a chain has states sunny (S), cloudy (C) and rainy (R) and

$$P = \begin{pmatrix} 0.6 & 0.2 & 0.2 \\ 0.3 & 0.7 & 0 \\ 0.6 & 0 & 0.4 \end{pmatrix}. \tag{9.24}$$

Cloud tends to persist; rain clears straight to sun, and there is no direct cloud–rain transition in either direction. The three equations of $\pi = \pi P$ are

$$\pi_S = 0.6\pi_S + 0.3\pi_C + 0.6\pi_R, \qquad \pi_C = 0.2\pi_S + 0.7\pi_C, \qquad \pi_R = 0.2\pi_S + 0.4\pi_R.$$

The second gives $\pi_C = \frac{2}{3}\pi_S$, the third $\pi_R = \frac{1}{3}\pi_S$ (the first is then automatic, since the three equations add up to an identity), and normalisation gives

$$\pi = \big(\tfrac{1}{2}, \; \tfrac{1}{3}, \; \tfrac{1}{6} \big). \tag{9.25}$$

Example: the Hardy–Weinberg law. A gene comes in two types A and a, and each person carries two copies, so people are of genotype AA, Aa or aa [Wasserman p. 392, Ex. 23.29]. Each parent passes on one of its two copies at random, and mates are chosen at random.

The gene pool. Let a population have genotype fractions r (AA), s (Aa) and $1 - r - s$ (aa). Pool all the genes: an AA person contributes two A's, an Aa person one A out of two, so the fraction of A genes in the pool is $g = r + s/2$, and the fraction of a genes is $h = 1 - g$. A child gets one gene from each parent, each an independent draw from the pool, so it is AA with probability g^2, Aa with $2gh$ (an A from either parent) and aa with h^2. Its own genes then carry the fraction $g^2 + \frac{1}{2}(2gh) = g(g + h) = g$ of A: the gene frequency does not change, and from the second generation on the genotype fractions are $(g^2, 2gh, h^2)$, the Hardy–Weinberg law.

One family line. Now follow a single person and their descendants, one child per generation, each child's other parent drawn at random from the population. The genotype of the nth descendant depends only on the genotype of the $(n-1)$th (and on the fixed pool), so it is a Markov chain on $\{AA, Aa, aa\}$. Its rows come from asking what each parent passes on.

- An AA parent always passes A; the mate passes A with probability g. The child is AA with probability g and Aa with h.

- An Aa parent passes A or a with probability $\frac{1}{2}$ each. The child is AA if both pass A, probability $\frac{1}{2}g$; aa if both pass a, probability $\frac{1}{2}h$; and Aa otherwise, probability $\frac{1}{2}h+\frac{1}{2}g=\frac{1}{2}$ (using $g+h=1$).
- An aa parent always passes a. The child is Aa with probability g and aa with h.

So

$$P=\begin{pmatrix} g & h & 0 \\ \frac{g}{2} & \frac{1}{2} & \frac{h}{2} \\ 0 & g & h \end{pmatrix}.$$

Hardy–Weinberg is stationary. Take $\pi=(g^2,2gh,h^2)$ and check $\pi P=\pi$ component by component, using $g+h=1$,

$$\begin{aligned}(\pi P)_{AA} &= g^2\cdot g+2gh\cdot\tfrac{g}{2}=g^2(g+h)=g^2,\\ (\pi P)_{Aa} &= g^2\cdot h+2gh\cdot\tfrac{1}{2}+h^2\cdot g=gh(g+h)+gh=2gh,\\ (\pi P)_{aa} &= 2gh\cdot\tfrac{h}{2}+h^2\cdot h=h^2(g+h)=h^2.\end{aligned}$$

From the crowd to one line. The chain is irreducible when $0<g<1$ (every genotype can reach every other within two generations) and aperiodic (the diagonal entries are positive), so by the ergodic theorem (Section 9.6) the fraction of its generations that one family line spends as AA tends to $\pi_{AA}=g^2$: the same as the fraction of AA people in any one generation.

9.4.1 Stationary does not mean convergent

A stationary distribution is an equilibrium. Whether the chain ever reaches it is a different question [Wasserman p. 392, Ex. 23.27]. The deterministic 3-cycle of Section 9.3.3 has the uniform distribution $\pi=(\frac{1}{3},\frac{1}{3},\frac{1}{3})$ as a stationary distribution (πP just permutes the three equal entries). Started in π it stays in π. Started in state 1 it is in state 1, then 2, then 3, then 1 again: p_t is a spike that rotates forever, and its distance from π never shrinks (Figure 9.11, blue). What does converge is the *time average* of p_t over a period, which is uniform; we will see in Section 9.6 that time averages need less than distributions do.

The chain has a **limiting distribution** π if the rows of P^n all converge to the same vector π,

$$\lim_{n\to\infty} p_{ij}(n)=\pi_j \quad \text{for every } i, \qquad \text{i.e.} \qquad P^n\to\begin{pmatrix}\pi\\ \pi\\ \vdots\\ \pi\end{pmatrix} \tag{9.26}$$

[Wasserman p. 391, Def. 23.24]. Then $p_n=p_0P^n\to\sum_i p_0(i)\pi=\pi$ for every initial distribution: the chain forgets where it started. The 3-cycle has a stationary distribution but no limiting one.

9.4.2 Detailed balance

Kruschke's argument for why the island rule works starts not from $\pi=\pi P$ but from pairs of neighbouring islands [Kruschke §7.2.5, Eq. (7.2)]. The ratio of the probability of hopping east from θ to that of hopping back west from $\theta+1$ is, by (9.12),

$$\frac{p_{\theta,\theta+1}}{p_{\theta+1,\theta}}=\frac{\frac{1}{2}}{\frac{1}{2}\,\theta/(\theta+1)}=\frac{\theta+1}{\theta}=\frac{P(\theta+1)}{P(\theta)}.$$

Rearranged, $P(\theta)\,p_{\theta,\theta+1}=P(\theta+1)\,p_{\theta+1,\theta}$: with the crowd spread in proportion to the populations, as many politicians cross from θ to $\theta+1$ each night as cross back. Every bridge carries equal traffic both ways.

Working definition 9.6 (detailed balance)

A distribution π satisfies **detailed balance** with respect to P if

$$\pi_i \, p_{ij} = \pi_j \, p_{ji} \qquad \text{for every pair } i, j \tag{9.27}$$

[the probability flow from i to j equals the flow from j to i, for each pair separately] [Wasserman p. 391, (23.15)]. A chain with such a π is called **reversible**. *Operational test:* for each pair with $p_{ij} > 0$, check that $p_{ji} > 0$ and that $p_{ij}/p_{ji} = \pi_j/\pi_i$. Only ratios of π enter.

Detailed balance is stronger than stationarity, and it implies it. Sum (9.27) over i:

$$(\pi P)_j = \sum_i \pi_i \, p_{ij} \overset{(1)}{=} \sum_i \pi_j \, p_{ji} \overset{(2)}{=} \pi_j \sum_i p_{ji} \overset{(3)}{=} \pi_j . \tag{9.28}$$

(1) Detailed balance, term by term.
(2) π_j does not depend on the summation index.
(3) Row j of P sums to one.

That is the whole proof [Wasserman p. 391, Thm 23.26]. Compare it with Kruschke's verification of stationarity for the islands, which needs the four cases of whether each neighbour is larger or smaller [Kruschke §7.2.5, Eq. (7.4)]: detailed balance does all four at once.

Key idea: Detailed balance ⇒ stationary

If $\pi_i p_{ij} = \pi_j p_{ji}$ for all pairs, then $\pi P = \pi$. To build a chain with a prescribed stationary distribution it is enough to make every pair of moves balance, and that needs only the ratios π_j/π_i, never the normalisation of π.

Example: the politician's equilibrium. For the islands, (9.27) holds with $\pi_\theta \propto \theta$ for each pair of neighbours (just shown), and trivially for non-neighbours, where both sides are zero. So $\pi_\theta = \theta/28$ is stationary. The politician never needed the total population 28: his rule uses only the ratio of two neighbouring populations, and the ratio is all detailed balance asks for. That is the whole secret of the Metropolis algorithm, and why it can sample a posterior whose evidence Z is unknown [36, 170].

Example: the three-state weather chain is reversible. For (9.24) with $\pi = (\frac{1}{2}, \frac{1}{3}, \frac{1}{6})$: $\pi_S p_{SC} = \frac{1}{2} \cdot 0.2 = 0.1 = \frac{1}{3} \cdot 0.3 = \pi_C p_{CS}$; $\pi_S p_{SR} = \frac{1}{2} \cdot 0.2 = 0.1 = \frac{1}{6} \cdot 0.6 = \pi_R p_{RS}$; and $\pi_C p_{CR} = 0 = \pi_R p_{RC}$. Each of the three "bridges" carries 0.1 of the probability each way per day (Figure 9.10a).

The physical picture: equilibrium without currents. In statistical mechanics, detailed balance is the statement that in thermal equilibrium every elementary process is balanced by its reverse: as many molecules react $A \to B$ per second as react $B \to A$, as many photons are absorbed in a transition as are emitted in the reverse one. This principle of microscopic reversibility is what made Einstein's A and B coefficients consistent with the Planck spectrum. There are no net probability currents anywhere. A stationary state can exist with currents, however, as a steady current flows round a circuit driven by a battery: global balance (9.22) holds at every node, but the flow on each wire does not vanish. Such states are *non-equilibrium steady states*.

Example: a biased cycle has a stationary distribution but no detailed balance. On a ring of three states let the walker step clockwise with probability 0.6, anticlockwise with 0.2, and stay

with 0.2:

$$P = \begin{pmatrix} 0.2 & 0.6 & 0.2 \\ 0.2 & 0.2 & 0.6 \\ 0.6 & 0.2 & 0.2 \end{pmatrix}.$$

Its columns sum to one as well as its rows (a **doubly stochastic** matrix), so the uniform $\pi = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ is stationary: $(\pi P)_j = \frac{1}{3}\sum_i p_{ij} = \frac{1}{3}$. But $\pi_1 p_{12} = 0.2 \neq \pi_2 p_{21} = 0.0\overline{6}$. Each edge carries a net clockwise current $J = \frac{1}{3}(0.6 - 0.2) = 0.1333$ per step, the same on all three edges, so the inflow and outflow balance at every node while probability circulates for ever (Figure 9.10b). A reversible chain could not do this. Going round any closed loop of states, detailed balance forces the product of the transition probabilities one way to equal the product the other way, because the π's cancel ($p_{12}p_{23}p_{31} = \frac{\pi_2}{\pi_1}p_{21} \cdot \frac{\pi_3}{\pi_2}p_{32} \cdot \frac{\pi_1}{\pi_3}p_{13}$), and here $0.6^3 \neq 0.2^3$. (The converse also holds, and this loop test is Kolmogorov's criterion for reversibility.)

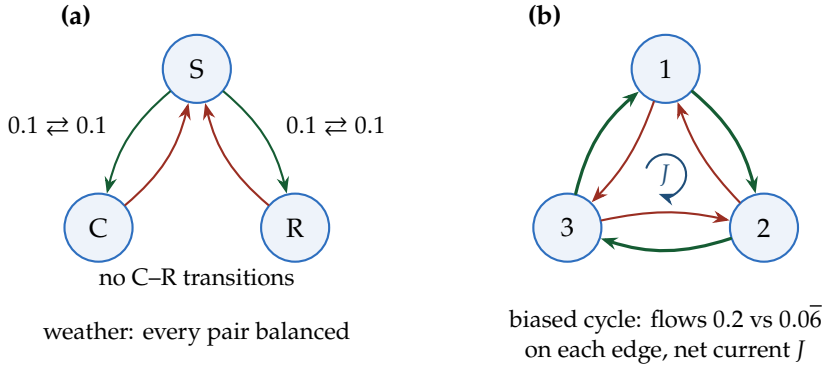


Figure 9.10: Probability flows $\pi_i p_{ij}$ in two stationary chains. (a) The three-state weather chain: each pair of states exchanges 0.1 of the probability per day in each direction, so no net current flows anywhere (equilibrium). (b) The biased three-state cycle: the clockwise flow (thick green) exceeds the anticlockwise one (red) on every edge, and a net current circulates round the ring, although inflow equals outflow at each state (a steady state that is not an equilibrium).

The biased cycle nevertheless converges to its uniform π (Figure 9.11, orange): it is irreducible and aperiodic (it has self-loops), and that is all convergence needs. Detailed balance is sufficient for stationarity, not necessary. Its signature in the dynamics is visible in the figure: the approach to π oscillates.

The eigenvalues of the biased cycle. P is **circulant**: each row is the row above shifted one place to the right, because the ring looks the same from every state. For such a matrix the eigenvectors can be written down at once. Let $w = e^{2\pi i/3}$, a cube root of one, and try $v_k = (1, w^k, w^{2k})$ for $k = 0, 1, 2$. Row 1 of Pv_k is $0.2 + 0.6w^k + 0.2w^{2k}$; row 2 is $0.2 + 0.2w^k + 0.6w^{2k} = w^k(0.2w^{-k} + 0.2 + 0.6w^k)$, which is w^k times the same number because $w^{-k} = w^{2k}$, and row 3 works the same way. So v_k is a right eigenvector with eigenvalue

$$\lambda_k = 0.2 + 0.6\,w^k + 0.2\,w^{-k}$$

(the same step as for the ring in (9.56)). For $k = 0$, $\lambda_0 = 1$, the uniform mode. For $k = 1$, with $w = -\frac{1}{2} + i\frac{\sqrt{3}}{2}$ and $w^{-1} = -\frac{1}{2} - i\frac{\sqrt{3}}{2}$,

$$\lambda_1 \overset{(1)}{=} 0.2 + 0.6\left(-\tfrac{1}{2} + i\tfrac{\sqrt{3}}{2}\right) + 0.2\left(-\tfrac{1}{2} - i\tfrac{\sqrt{3}}{2}\right) \overset{(2)}{=} -0.2 + 0.4 \cdot \tfrac{\sqrt{3}}{2}\,i = -0.2 + 0.346\,i,$$
$$|\lambda_1| \overset{(3)}{=} \sqrt{0.2^2 + 0.346^2} = \sqrt{0.04 + 0.12} = 0.4.$$

(1) Substitute w and w^{-1}.

(2) Real part $0.2 - 0.3 - 0.1 = -0.2$; imaginary part $(0.6 - 0.2)\frac{\sqrt{3}}{2}$.

(3) Modulus of a complex number; $0.346^2 = 0.4^2 \cdot \frac{3}{4} = 0.12$.

λ_2 is the complex conjugate. Clockwise and anticlockwise steps enter with different weights, 0.6 and 0.2, so the imaginary parts do not cancel; with equal weights λ_1 would be real.

A complex eigenvalue is a rotation. Write $\lambda_1 = 0.4\, e^{i\alpha}$ with $\alpha = 120°$. Then $\lambda_1^t = 0.4^t e^{i\alpha t}$: each step multiplies the size by 0.4 and advances the phase by 120°, one third of a turn, as the walker drifts one state clockwise. The deviation $p_t - \pi$ is a combination of this mode and its conjugate, so each of its components is 0.4^t times a cosine of αt plus a phase that differs from state to state: the excess probability is carried round the ring while it shrinks, a spiral into π, with distance decaying as 0.4^t (Figure 9.11b). A reversible chain cannot rotate: with $D = \mathrm{diag}(\sqrt{\pi_i})$, detailed balance says that DPD^{-1} is a symmetric matrix ($\sqrt{\pi_i}p_{ij}/\sqrt{\pi_j} = \sqrt{\pi_j}p_{ji}/\sqrt{\pi_i}$), so all eigenvalues of P are real.

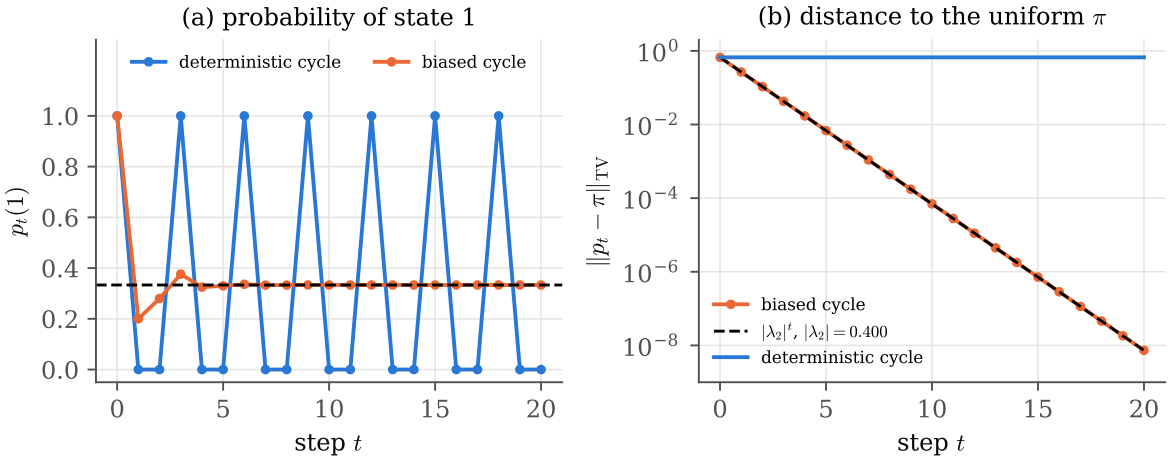


Figure 9.11: Two chains on a ring of three states, both with the uniform distribution as equilibrium, started in state 1. (a) The deterministic cycle (blue) carries its spike round forever; the biased cycle (orange) spirals in towards $\frac{1}{3}$. (b) The distance to the uniform distribution: constant for the deterministic cycle, and decaying as 0.4^t for the biased one.

Reversibility as a movie run backwards. The name "reversible" has a literal meaning. In equilibrium, $X_t \sim \pi$, the probability of seeing the step $i \to j$ is $\mathrm{P}(X_t = i, X_{t+1} = j) = \pi_i p_{ij}$, and of seeing $j \to i$ it is $\pi_j p_{ji}$. Detailed balance says these are equal, so a film of the stationary chain has the same statistics played forwards or backwards [Lambert §13.6.3]. For the biased cycle the film run backwards shows a walker that prefers anticlockwise steps; you can tell the direction of time.

9.4.3 Continuous states

Nothing changes for a continuous state space except that sums become integrals. The transition matrix becomes a **transition kernel** $p(x, y)$, the density of the next state y given the current state x, with $\int p(x, y)\,\mathrm{d}y = 1$ [Wasserman p. 413]. A density f_t of walkers moves to

$$f_{t+1}(y) = \int f_t(x)\, p(x, y)\,\mathrm{d}x \tag{9.29}$$

(the transition operator applied to f_t, [Lambert §13.6.1]), a density f is stationary if $f(y) = \int f(x)p(x, y)\,\mathrm{d}x$, and detailed balance reads $f(x)\,p(x, y) = f(y)\,p(y, x)$. The proof that it implies stationarity is (9.28) with integrals [Wasserman p. 414, (24.7)]: $\int f(x)p(x, y)\,\mathrm{d}x = \int f(y)p(y, x)\,\mathrm{d}x = f(y)\int p(y, x)\,\mathrm{d}x = f(y)$.

Example: a Gaussian walker with a pull. Let the next state be $Y = \rho x + \sqrt{1-\rho^2}\,Z$ with $Z \sim \mathcal{N}(0,1)$ and $|\rho| < 1$, so that $p(x,y)$ is the $\mathcal{N}(\rho x, 1-\rho^2)$ density in y: a fraction ρ of the current position is kept and fresh noise is added. This is the **AR(1)** chain (autoregressive of order 1), the discrete-time version of a Brownian particle in a harmonic well (the Ornstein–Uhlenbeck process). Is $f = \mathcal{N}(0,1)$ stationary? Check detailed balance: the exponent of $f(x)p(x,y)$ is

$$-\frac{x^2}{2} - \frac{(y-\rho x)^2}{2(1-\rho^2)} \overset{(1)}{=} -\frac{x^2(1-\rho^2) + y^2 - 2\rho xy + \rho^2 x^2}{2(1-\rho^2)} \overset{(2)}{=} -\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)},$$

(1) Common denominator, and $(y-\rho x)^2 = y^2 - 2\rho xy + \rho^2 x^2$.
(2) The $\rho^2 x^2$ terms cancel.

which is symmetric under $x \leftrightarrow y$, and the prefactor $1/(2\pi\sqrt{1-\rho^2})$ is symmetric too. So $f(x)p(x,y) = f(y)p(y,x)$: the AR(1) chain is reversible with stationary law $\mathcal{N}(0,1)$. The joint law of two consecutive states in equilibrium is the bivariate Gaussian with correlation ρ, the same whichever way round you read it. We will watch this chain forget its starting point in Section 9.7.

9.5 Forgetting where you started

Five weather forecasters use the same three-state weather chain (9.24) (sunny, cloudy, rainy, with $p_{SS} = 0.6$, $p_{CC} = 0.7$, $p_{RR} = 0.4$, rain clearing straight to sun and no direct cloud–rain changes), but they disagree violently about today. One is certain it is sunny, one certain it is cloudy, one certain it is raining; one has no idea and says $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$; one is 90% sure of rain. Each forecasts day t by $p_t = p_0 P^t$. Figure 9.12a shows their forecasts of rain: wildly different on day 0, nearly indistinguishable by day 5, all on the line $\pi_R = \frac{1}{6}$ of the stationary distribution (9.25) soon after. The chain has forgotten where it started.

This is the property an MCMC run relies on. We cannot start a sampler from the target distribution (if we could draw from it we would not need the sampler), so we start it anywhere, and we need a guarantee that the start is forgotten, together with a rate.

The questions.

1. Does $p_t = p_0 P^t$ converge to the same π for every p_0?
2. How fast does the memory of p_0 fade?
3. What property of P guarantees it?

To answer "how fast" we need a distance between distributions.

Working definition 9.7 (total-variation distance)

For two distributions p, q on the same states,

$$\|p - q\|_{\mathrm{TV}} = \tfrac{1}{2}\sum_i |p_i - q_i| = \max_A |p(A) - q(A)|\,, \tag{9.30}$$

where A runs over all events [all subsets of states]. It is the largest disagreement between p and q about the probability of any event, so it lies between 0 (identical) and 1 (no overlap at all). *Operational recipe:* half the sum of absolute differences (`tv` in `lib_markov.py`).

The two expressions in (9.30) agree. Let $A^* = \{i : p_i > q_i\}$. For any event A, $p(A) - q(A) = \sum_{i\in A}(p_i - q_i)$ is largest when A contains exactly the states where the difference is positive, so the

maximum is $p(A^*) - q(A^*)$ (and the most negative value is $-(p(A^*) - q(A^*))$, at the complement). Because $\sum_i (p_i - q_i) = 1 - 1 = 0$, the positive and the negative differences have equal total size, and each is half of $\sum_i |p_i - q_i|$.

9.5.1 Two states: the whole story in one line

Take a two-state chain with $a = p_{12}$, $b = p_{21}$ and stationary distribution $\pi = (\frac{b}{a+b}, \frac{a}{a+b})$ (9.23). Let $d_t = p_t(1) - \pi_1$ be the forecaster's excess belief in state 1. Then

$$d_{t+1} = p_{t+1}(1) - \pi_1 \overset{(1)}{=} \big[p_t(1)(1-a) + (1-p_t(1))\,b\big] - \big[\pi_1(1-a) + (1-\pi_1)\,b\big]$$
$$\overset{(2)}{=} (1-a-b)\big(p_t(1) - \pi_1\big) = \lambda\, d_t, \qquad \lambda \equiv 1 - a - b. \tag{9.31}$$

(1) One step of (9.14) for the first component, for p_t and for the stationary π (for which one step changes nothing).
(2) Collect the terms in $p_t(1)$ and π_1: the constant b cancels.

So $d_t = \lambda^t d_0$, and since $p_t(2) - \pi_2 = -d_t$,

$$p_t = \pi + \lambda^t\,(p_0 - \pi), \qquad \|p_t - \pi\|_{\mathrm{TV}} = |\lambda|^t\ \|p_0 - \pi\|_{\mathrm{TV}}\,. \tag{9.32}$$

The memory of the start decays geometrically with ratio $|\lambda| = |1 - a - b|$, and λ is the second eigenvalue of P (the trace of P is $2 - a - b = 1 + \lambda$). The two ways for $|\lambda|$ to equal 1 are exactly the failures of Section 9.3: $a = b = 0$, two absorbing states and a reducible chain, which remembers its start forever; and $a = b = 1$, a deterministic flip with period 2, for which $\lambda = -1$ and the excess d_t changes sign every day without shrinking. Every other two-state chain forgets.

Example: Wasserman's weather. For (9.11), $a = 0.6$, $b = 0.8$, so $\lambda = -0.4$ and $\pi = (\frac{4}{7}, \frac{3}{7})$. Starting on a sunny day, $d_0 = 1 - \frac{4}{7} = \frac{3}{7}$, and $p_t(\text{sunny}) = \frac{4}{7} + \frac{3}{7}(-0.4)^t$: 1, 0.4, 0.64, 0.544, ..., the sequence found by brute force in Section 9.2.3. The negative λ explains the overshoot: a sunny day makes cloud likely tomorrow, which makes sun likely the day after, and the alternation damps by a factor 0.4 a day.

9.5.2 Many states: the eigenvalues of P

For a K-state chain the role of λ is played by the eigenvalues of P. Let us assume that P has K independent eigenvectors (this is always true for a reversible chain, whose eigenvalues were shown to be real in Section 9.4.2; for the rest see the end of this subsection). There are right eigenvectors, $Pv_k = \lambda_k v_k$ (columns), and left eigenvectors, $\ell_k P = \lambda_k \ell_k$ (rows), with the same eigenvalues. A left and a right eigenvector with different eigenvalues are orthogonal: $\ell_k P v_m$ equals both $\lambda_k \ell_k v_m$ and $\lambda_m \ell_k v_m$, so $\ell_k v_m = 0$ when $\lambda_k \neq \lambda_m$. Label $\lambda_1 = 1$, with right eigenvector $v_1 = \mathbf{1}$ (rows sum to one) and left eigenvector $\ell_1 = \pi$, and order the rest by modulus, $|\lambda_2| \geq |\lambda_3| \geq \cdots$. Expand the starting distribution in left eigenvectors, $p_0 = \sum_k c_k \ell_k$; multiplying by v_k and using the orthogonality picks out one coefficient, $c_k = p_0 v_k / (\ell_k v_k)$. Then

$$p_t = p_0 P^t \overset{(1)}{=} \sum_k c_k\, \ell_k P^t \overset{(2)}{=} \sum_k c_k\, \lambda_k^t\, \ell_k \overset{(3)}{=} \pi + \sum_{k \geq 2} c_k\, \lambda_k^t\, \ell_k. \tag{9.33}$$

(1) Expand p_0; the product is linear.
(2) Each ℓ_k is a left eigenvector, so $\ell_k P^t = \lambda_k^t \ell_k$.
(3) $\lambda_1 = 1$, $\ell_1 = \pi$ and $c_1 = p_0 \mathbf{1} / (\pi \mathbf{1}) = 1/1 = 1$: every probability vector contains exactly one unit of the stationary distribution.

Every other component is multiplied by λ_k^t each step. If $|\lambda_k| < 1$ for all $k \geq 2$, all of them die and $p_t \to \pi$ from any start. At late times the slowest one dominates,

$$\|p_t - \pi\|_{\mathrm{TV}} \simeq C\ |\lambda_2|^t\,, \qquad \ln \|p_t - \pi\|_{\mathrm{TV}} \simeq \ln C + t \ln |\lambda_2|\,, \tag{9.34}$$

a straight line of slope $\ln|\lambda_2|$ on a log plot. The number of steps for the memory to fall by a factor e is the **relaxation time** $t_{\mathrm{rel}} = -1/\ln|\lambda_2|$, and $1 - |\lambda_2|$ is called the **spectral gap**. This is exactly the physics of normal modes in a damped system: the initial condition is a superposition of modes, each decays at its own rate, and after a while only the slowest one is visible.

Why $|\lambda| \leq 1$, and why only one eigenvalue equals 1. This is the Perron–Frobenius theorem for stochastic matrices, and for our purposes it has a two-line proof. Take a right eigenvector, $Pv = \lambda v$, and look at its largest entry in modulus, at state i^*. Row i^* of Pv is an *average* of the entries of v with weights $p_{i^*j} \geq 0$ that sum to one, and an average cannot exceed the largest thing being averaged:

$$|\lambda|\ |v_{i^*}| = |(Pv)_{i^*}| \overset{(1)}{=} \Big|\sum_j p_{i^*j} v_j\Big| \overset{(2)}{\leq} \sum_j p_{i^*j}\,|v_j| \overset{(3)}{\leq} \sum_j p_{i^*j}\,|v_{i^*}| \overset{(4)}{=} |v_{i^*}|\,. \tag{9.35}$$

(1) Definition of the matrix–vector product.
(2) Triangle inequality.
(3) $|v_j| \leq |v_{i^*}|$ by the choice of i^*.
(4) Row i^* sums to one.

So $|\lambda| \leq 1$. Now suppose the chain is ergodic, so that some power P^m has all entries positive (Section 9.3.3), and apply the same argument to P^m, whose eigenvalue for v is λ^m. If $|\lambda| = 1$, both inequalities in (9.35) must be equalities. With every weight $p_{i^*j}(m)$ positive, equality in (3) forces $|v_j| = |v_{i^*}|$ for *all* j, and equality in the triangle inequality (2) forces all the v_j to have the same phase. So v is constant, $v \propto \mathbf{1}$, and then $Pv = v$ says $\lambda = 1$. Hence for an ergodic chain the eigenvalue 1 belongs only to the constant vector and every other eigenvalue has $|\lambda_k| < 1$. The failures are again the ones of Section 9.3: the deterministic 3-cycle has the eigenvalues $1, e^{\pm 2\pi i/3}$, all of modulus 1, which rotate the spike forever; a chain with two closed classes has the eigenvalue 1 twice, one stationary distribution living on each class.

When there are not enough eigenvectors. The expansion (9.33) assumed K independent eigenvectors. A matrix can fall short of that, and the simplest example shows what changes. Take the 2×2 block

$$B = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}, \qquad B^t = \begin{pmatrix} \lambda^t & t\lambda^{t-1} \\ 0 & \lambda^t \end{pmatrix},$$

which has the single eigenvector $(1,0)^{\mathsf{T}}$. The formula for B^t holds at $t = 1$, and multiplying it by B once more gives the off-diagonal entry $\lambda^t + t\lambda^{t-1}\lambda = (t+1)\lambda^t$, so it holds for all t. The power still decays like $|\lambda|^t$, only with an extra factor t, which a geometric decay overwhelms. Every square matrix can be brought to blocks of this kind (its Jordan form), and a block of size $r+1$ gives factors up to t^r; the decay is again geometric with ratio $|\lambda_2|$, so the conclusion is unchanged. The general block is not worked out here, and nothing below depends on it: the Doeblin argument of Section 9.5.3 proves geometric forgetting without any eigenvectors.

Example: the three-state weather chain, exactly. Take the three-state weather chain (9.24) (sunny, cloudy, rainy), with $\pi = (\frac{1}{2}, \frac{1}{3}, \frac{1}{6})$ from (9.25), and find all of its eigenvalues and eigenvectors by hand.

The eigenvalues. The sum of the eigenvalues is the trace, $0.6+0.7+0.4 = 1.7$, and their product is the determinant. Expanding along the first row, each entry times the 2×2 determinant left when its row and column are struck out,

$$\det P = 0.6\begin{vmatrix}0.7 & 0\\ 0 & 0.4\end{vmatrix} - 0.2\begin{vmatrix}0.3 & 0\\ 0.6 & 0.4\end{vmatrix} + 0.2\begin{vmatrix}0.3 & 0.7\\ 0.6 & 0\end{vmatrix} = 0.6(0.28) - 0.2(0.12) + 0.2(-0.42) = 0.06.$$

With $\lambda_1 = 1$ the other two satisfy $\lambda_2 + \lambda_3 = 0.7$ and $\lambda_2\lambda_3 = 0.06$, the roots of $\lambda^2 - 0.7\lambda + 0.06 = 0$: $\lambda_2 = 0.6$ and $\lambda_3 = 0.1$.

The right eigenvectors. For $\lambda = 0.6$ the equation $(P - 0.6\,\mathbb{1})v = 0$ is three row equations,

$$0.2v_2 + 0.2v_3 = 0, \qquad 0.3v_1 + 0.1v_2 = 0, \qquad 0.6v_1 - 0.2v_3 = 0.$$

The third gives $v_3 = 3v_1$, the second $v_2 = -3v_1$, and the first is then satisfied ($-0.6v_1 + 0.6v_1 = 0$), as it must be for an eigenvalue. Choosing $v_1 = 1$ gives $v_2 = (1, -3, 3)^{\mathsf{T}}$. For $\lambda = 0.1$ the rows read $0.5v_1 + 0.2v_2 + 0.2v_3 = 0$, $0.3v_1 + 0.6v_2 = 0$ and $0.6v_1 + 0.3v_3 = 0$, so $v_2 = -\tfrac{1}{2}v_1$ and $v_3 = -2v_1$; with $v_1 = 2$, $v_3 = (2, -1, -4)^{\mathsf{T}}$ (the first row checks: $1 - 0.2 - 0.8 = 0$).

The left eigenvectors from reversibility. The chain satisfies detailed balance (Section 9.4.2), and that turns each right eigenvector into a left one. Define $\ell(i) = \pi_i v(i)$, each entry of v weighted by the stationary probability. Then

$$(\ell P)_j \overset{(1)}{=} \sum_i \pi_i v_i\, p_{ij} \overset{(2)}{=} \sum_i \pi_j p_{ji}\, v_i \overset{(3)}{=} \pi_j\,(Pv)_j \overset{(4)}{=} \lambda\,\pi_j v_j = \lambda\,\ell_j.$$

(1) Row vector times matrix, with $\ell_i = \pi_i v_i$.
(2) Detailed balance, $\pi_i p_{ij} = \pi_j p_{ji}$.
(3) π_j does not depend on i; what is left is row j of Pv.
(4) v is a right eigenvector.

So

$$v_2 = (1,-3,3)^{\mathsf{T}}, \qquad \ell_2 = \left(\tfrac{1}{2}, -1, \tfrac{1}{2}\right), \qquad \ell_2 v_2 = 5,$$
$$v_3 = (2,-1,-4)^{\mathsf{T}}, \qquad \ell_3 = \left(1, -\tfrac{1}{3}, -\tfrac{2}{3}\right), \qquad \ell_3 v_3 = 5.$$

(Check one: the first entry of $\ell_2 P$ is $\frac{1}{2}(0.6) - 1(0.3) + \frac{1}{2}(0.6) = 0.3 = 0.6\cdot\frac{1}{2}$.) For the forecaster certain of rain, $p_0 = (0,0,1)$, so $c_2 = p_0 v_2/5 = \frac{3}{5}$ and $c_3 = p_0 v_3/5 = -\frac{4}{5}$, and (9.33) becomes

$$p_t = \left(\tfrac{1}{2}, \tfrac{1}{3}, \tfrac{1}{6}\right) + \tfrac{3}{5}\,(0.6)^t\left(\tfrac{1}{2}, -1, \tfrac{1}{2}\right) - \tfrac{4}{5}\,(0.1)^t\left(1, -\tfrac{1}{3}, -\tfrac{2}{3}\right). \tag{9.36}$$

At $t = 0$ this is $(0,0,1)$, as it must be. The 0.1^t mode is gone within two or three days; after that $\|p_t - \pi\|_{\mathrm{TV}} \simeq \frac{3}{5}(0.6)^t \cdot \frac{1}{2}(\frac{1}{2} + 1 + \frac{1}{2}) = \frac{3}{5}(0.6)^t$. On day 10 this predicts 0.0036, and the exact distance is 0.0036; by day 20 it is 2.19×10^{-5}. Fitting a straight line to $\ln\|p_t - \pi\|$ over days 10–30 gives slope -0.5108, against $\ln 0.6 = -0.5108$ (Figure 9.12b). All five forecasters have the same slope; they differ only in the constant C, which is the size of their initial error along ℓ_2. The uniform forecaster starts closest and stays closest. Simulating 200 000 independent weather histories from a rainy day reproduces the exact curve until it hits the Monte Carlo floor $\approx 1.2\times 10^{-3}$, the typical distance between π and a histogram of that many draws from π.

Key idea: A finite ergodic chain forgets its start

If a finite chain is irreducible and aperiodic, it has exactly one stationary distribution π, and for every initial distribution p_0, $\left\|p_0 P^t - \pi\right\|_{\mathrm{TV}} \to 0$ geometrically. The rate is set by the second-largest eigenvalue modulus: $\|p_t - \pi\|_{\mathrm{TV}} \simeq C\,|\lambda_2|^t$.

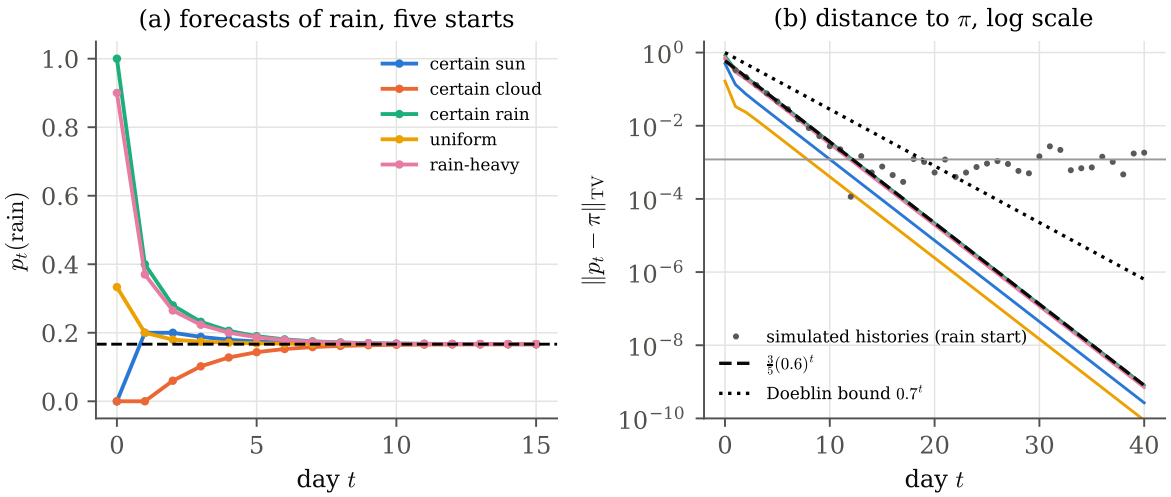


Figure 9.12: Five forecasters with very different beliefs about day 0, one weather chain. (a) Their forecasts of rain converge to $\pi_R = \frac{1}{6}$ (dashed) within a few days. (b) The total-variation distance to π on a log scale: after a short transient from the fast 0.1^t mode, every curve is a straight line of slope $\ln 0.6$, parallel to $\frac{3}{5}(0.6)^t$ (dashed). Grey dots are 2×10^5 simulated histories started on a rainy day; they follow the exact curve down to the sampling-noise floor (horizontal line). The dotted line is the Doeblin bound 0.7^t of Section 9.5.3, which lies above every curve.

9.5.3 The easy-to-see proof: a contraction

The eigenvalue argument gives the exact rate, but it needs the eigenvectors. There is a cruder argument that needs nothing, works directly with probability, and proves convergence, existence and uniqueness of π at once. It goes back to Doeblin.

The idea. Look at the rows of the weather matrix (9.24) as three forecasts for tomorrow, one for each kind of today (Figure 9.13). Whatever today is, tomorrow is sunny with probability at least 0.3 (the smallest entry of the first column). That 0.3 is common to all three forecasts: it does not depend on today at all. So we may think of tomorrow's weather as drawn in two stages. With probability $\delta = 0.3$, ignore today completely and draw tomorrow from a fixed distribution ν (here, sun for certain). With probability $1 - \delta = 0.7$, draw it from what is left of today's row. Whenever the first branch is taken, the memory of today is wiped out. After t days the memory survives only if the second branch was taken every time, which has probability 0.7^t.

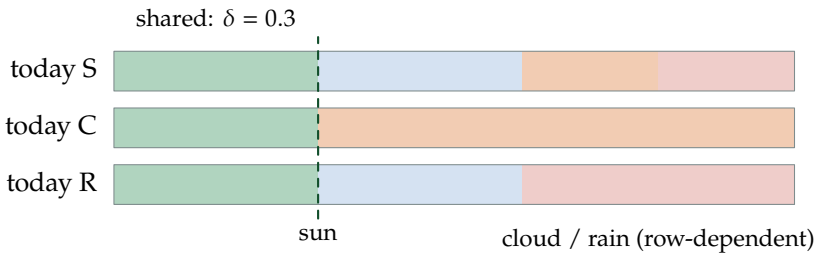


Figure 9.13: The rows of the weather matrix as stacked bars: tomorrow's sun (green and blue), cloud (orange) and rain (red), for each kind of today. The green part, a probability 0.3 of sun, is the same in every row; the rest depends on today. Each day, the shared part erases the dependence on the past with probability $\delta = 0.3$.

The calculation. Define the **Doeblin coefficient** $\delta = \sum_j \min_i p_{ij}$, the total mass that all rows have in common, and suppose $\delta > 0$. Then $v_j = \min_i p_{ij}/\delta$ is a probability vector, and $R = (P - \delta\,\mathbf{1}v)/(1-\delta)$ is a stochastic matrix (its entries are non-negative because $p_{ij} \geq \min_k p_{kj}$, and its rows sum to $(1-\delta)/(1-\delta) = 1$). So

$$P = \delta\,\mathbf{1}v + (1-\delta)\,R, \tag{9.37}$$

where $\mathbf{1}v$ is the matrix with every row equal to v. Take any two distributions p and q:

$$\begin{aligned} \|pP - qP\|_{\mathrm{TV}} &\overset{(1)}{=} \tfrac{1}{2}\big\|\delta\,(p-q)\mathbf{1}\,v + (1-\delta)(p-q)R\big\|_1 \overset{(2)}{=} (1-\delta)\,\tfrac{1}{2}\,\|(p-q)R\|_1 \\ &\overset{(3)}{\leq} (1-\delta)\,\tfrac{1}{2}\,\|p-q\|_1 = (1-\delta)\;\|p-q\|_{\mathrm{TV}}\,. \end{aligned} \tag{9.38}$$

(1) Insert (9.37); $\|x\|_1 = \sum_j |x_j|$.
(2) $(p-q)\mathbf{1} = \sum_i p_i - \sum_i q_i = 0$: the shared part sends both distributions to the same place, so it cancels in the difference.
(3) A stochastic matrix cannot increase the ℓ_1 norm: $\sum_j \big|\sum_i x_i R_{ij}\big| \leq \sum_j \sum_i |x_i|\,R_{ij} = \sum_i |x_i|$, using the triangle inequality and then the row sums of R.

One step of the chain shrinks the distance between any two distributions by at least the factor $1-\delta$. Three consequences follow at once.

- *Uniqueness.* If π and π' were both stationary, $\|\pi - \pi'\| = \|\pi P - \pi' P\| \leq (1-\delta)\,\|\pi - \pi'\|$, which forces $\|\pi - \pi'\| = 0$.
- *Existence.* The sequence $p_0, p_0P, p_0P^2, \ldots$ has steps that shrink geometrically, so it is a Cauchy sequence and converges; its limit is unchanged by one more step, so it is stationary. (This is the contraction-mapping theorem.)
- *Forgetting, with a guaranteed rate.* Taking $q = \pi$ and iterating,

$$\big\|p_0P^t - \pi\big\|_{\mathrm{TV}} \leq (1-\delta)^t\;\|p_0 - \pi\|_{\mathrm{TV}} \leq (1-\delta)^t. \tag{9.39}$$

Example: two states, where the bound is exact. For the two-state chain, $\delta = \min(1-a, b) + \min(a, 1-b)$. If $a+b \leq 1$ then $1-a \geq b$ and $1-b \geq a$, so $\delta = a+b$ and $1-\delta = 1-a-b$. If $a+b > 1$ the minima are $1-a$ and $1-b$, so $1-\delta = a+b-1$. Either way $1-\delta = |1-a-b| = |\lambda|$: for two states the contraction factor is exactly the eigenvalue modulus, and (9.39) is the exact solution (9.32). For Wasserman's weather, $\delta = \min(0.4, 0.8) + \min(0.6, 0.2) = 0.6$ and $1-\delta = 0.4 = |-0.4|$.

Example: three states, where it is only a bound. For the three-state weather chain the column minima are 0.3, 0 and 0, so $\delta = 0.3$ and the guaranteed factor is 0.7 per day, against the true 0.6 (Figure 9.12b, dotted line). The bound is honest but pessimistic: it only uses the part of the rows that is literally identical, while the eigenvalue sees all the averaging.

When no column is shared. For the politician, no island can be reached from every island in one night, so every column of (9.13) has a zero and $\delta = 0$. The argument then applies to P^m instead of P: for an ergodic chain some power P^m is strictly positive (Section 9.3.3), so its Doeblin coefficient δ_m is positive, and the distance shrinks by $1-\delta_m$ every m steps.

The seven islands. After three nights, can every island have sent the politician to island 4? From island 1 the only way is east, east, east, with probability $\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2} = \frac{1}{8}$; from island 7 it is west, west, west, with probability $\frac{3}{7}\cdot\frac{5}{12}\cdot\frac{2}{5} = \frac{1}{14}$. The islands nearer to 4 get there in fewer moves and must spend the remaining nights somewhere, which they can because every island has a

refused proposal, $p_{\theta\theta} = 1/(2\theta) > 0$, by (9.12). Adding the probabilities of all such three-night paths gives column 4 of P^3, from islands $1, \dots, 7$:

$$\big(0.125,\ 0.135,\ 0.309,\ 0.150,\ 0.254,\ 0.051,\ 0.071\big).$$

Every entry is positive, and the smallest, 0.051 from island 6 (two moves plus one wasted night on an island that rarely refuses), is the mass that all rows of P^3 share in column 4. No other column of P^3 is positive in every row (island 1 cannot reach island 5 or beyond in three nights, island 7 cannot reach 3 or below), so $\delta_3 \approx 0.05$. Longer blocks share more:

block length m	1	2	3	4	5	6
δ_m	0	0	0.051	0.152	0.229	0.316

At $m = 6$ every column has a positive minimum, so P^6 is strictly positive. Geometric forgetting holds; only the guaranteed rate is per block of m steps, for example a factor $1 - 0.316$ every six days.

The actual rate. It is again $|\lambda_2|$, which is 0.887 for the politician. The memory falls as $|\lambda_2|^t$, so it halves after the number of days t with $|\lambda_2|^t = \frac{1}{2}$, that is $t = \ln 2/(-\ln|\lambda_2|) = \ln 2/(-\ln 0.887) \approx 6$ days, slower than the weather, because he can move only one island at a time.

9.5.4 Coupling: two walkers who meet

The two-stage picture of the Doeblin step has a vivid probabilistic version, which turns the inequality into a story about two walkers. Walker A starts from our arbitrary p_0. Walker B starts from the stationary π, so B's distribution is π on every day. Let them move together in a correlated way, each one still obeying the chain P on its own, and arrange that once they stand on the same state they move together forever. Let T be the day they first meet. After T the two walkers are identical, so for any event A,

$$\big|\mathrm{P}(X_t^A \in A) - \mathrm{P}(X_t^B \in A)\big| \overset{(1)}{=} \big|\mathrm{P}(X_t^A \in A,\ T > t) - \mathrm{P}(X_t^B \in A,\ T > t)\big| \overset{(2)}{\le} \mathrm{P}(T > t), \tag{9.40}$$

(1) Split each probability according to whether $T \le t$ or $T > t$. On $T \le t$ the two walkers are at the same place, so those parts are equal and cancel.

(2) Both remaining terms lie between 0 and $\mathrm{P}(T > t)$, so their difference does too.

and, since $X_t^B \sim \pi$, maximising over A gives the **coupling inequality** $\|p_t - \pi\|_{\mathrm{TV}} \le \mathrm{P}(T > t)$. Forgetting the start is the same as meeting a walker who never had one to forget.

The Doeblin split (9.37) tells us how to make them meet.

The construction. Each day, flip one coin, shared by both walkers, that comes up heads with probability δ. On heads, draw one state from ν and move *both* walkers there. On tails, each walker draws its next state from its own row of R, independently of the other. Once they stand on the same state, let them use the same random numbers from then on, so they never separate.

Each walker on its own is still the chain P. Look at walker A alone, standing at i. It goes to j either by heads and then ν choosing j, or by tails and then row i of R choosing j:

$$\mathrm{P}(\text{A moves } i \to j) = \delta\,\nu_j + (1 - \delta)\,R_{ij} = p_{ij},$$

which is exactly row i of (9.37). The same holds for B, and also after the meeting, since then both follow one walker's moves. So the coupling changes how the two walkers are correlated but not how either moves, and $X_t^B \sim \pi$ on every day, as the coupling inequality needs.

The meeting time. Every heads day ends with the walkers together, so they can still be apart on day t only if all t coins came up tails: $\mathrm{P}(T > t) \le (1-\delta)^t$, which is (9.39) again, now as a statement about meeting times. The bound can be sharpened a little at the start. A and B may already stand on the same state on day 0; with their starting draws arranged to agree as often as possible, this happens with probability $\sum_i \min(p_0(i), \pi_i)$. For A starting on a rainy day this is π_R, so the two are apart on day 0 with probability $1-\pi_R$, and $\mathrm{P}(T > t) \le (1-\delta)^t(1-\pi_R)$.

The script `05_coupling.py` runs 200 000 pairs from a rainy day (A) and from π (B), once with this Doeblin coupling and once with the naive coupling in which the two walkers move independently until they happen to meet (Figure 9.14). On day 10 the exact distance is 0.0036; the probability of not having met is 0.0046 for the Doeblin coupling and 0.018 for independent moves; and the guaranteed bound $(0.7)^{10}(1-\pi_R)$ is 0.0235. Each coupling gives a valid upper bound on the distance; a cleverer coupling gives a tighter one.

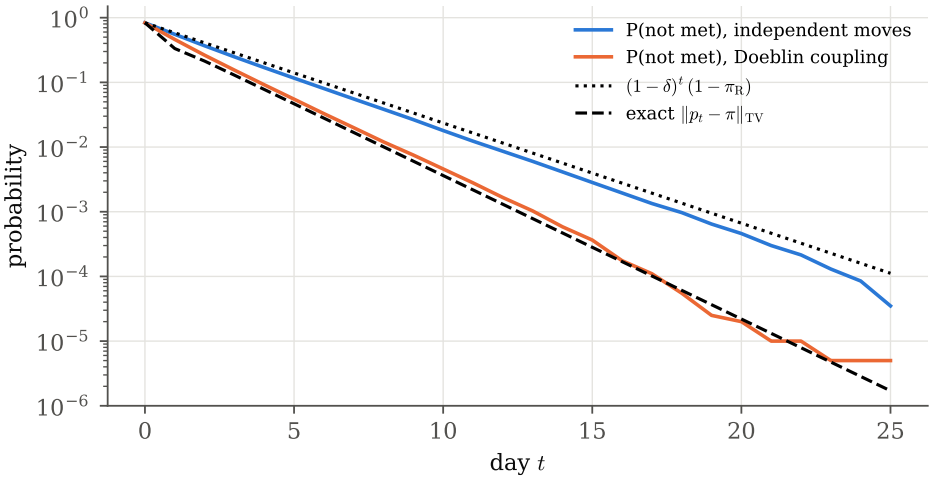


Figure 9.14: Coupling a walker started on a rainy day with one started from the stationary distribution. The probability that they have not yet met, for walkers moving independently (blue) and for the Doeblin coupling that sometimes forces them onto the same state (orange), always lies above the true distance $\|p_t-\pi\|_{\mathrm{TV}}$ (dashed), and the Doeblin coupling lies below its guaranteed bound $(1-\delta)^t(1-\pi_R)$ (dotted).

Watch out: A bound is not a rate

$(1-\delta)^t$ and $\mathrm{P}(T > t)$ guarantee forgetting; they do not measure it. The actual decay of $\|p_t-\pi\|$ is $|\lambda_2|^t$, which can be much faster. In the other direction, a chain can look converged long before it is: a walker trapped in one of two well-separated regions has $|\lambda_2|$ just below 1, a tiny Doeblin coefficient, and a histogram that looks perfectly stable for a long time while it misses half of the target.

9.6 One long walk is a sample: the ergodic theorem

Section 9.5 was about a crowd: the histogram of many independent walkers on day t approaches π. An MCMC run is not a crowd. It is one walker (or a handful), followed for a long time, and what we use is its diary: the fraction of steps it spent in each region, or the average of some quantity g over all the steps. One politician, followed for 100 000 days, spent his days in proportions within 0.0025 of $\theta/28$ (Figure 9.6b). Why should a single walker's diary reproduce the crowd's equilibrium? The successive entries in the diary are not independent, so the ordinary law of large numbers of Chapter 3 does not apply.

The questions.

1. Does the fraction of time one walker spends in state j converge to π_j, and the average of $g(X_t)$ along the walk to $\mathbb{E}_\pi\, g = \sum_j g(j)\pi_j$?
2. How accurate is the average after N steps, given that the steps are correlated?

The first question is the oscillator's question of Section 9.1, time average versus ensemble average, now for a random dynamics.

Key idea: Ergodic theorem for Markov chains

For an irreducible, ergodic Markov chain with stationary distribution π, started anywhere, and for any bounded function g of the state,

$$\frac{1}{N}\sum_{n=1}^{N} g(X_n) \xrightarrow[N\to\infty]{} \mathbb{E}_\pi(g) = \sum_j g(j)\,\pi_j \qquad \text{with probability one} \tag{9.41}$$

[Wasserman p. 391, Thm 23.25]. The time average along one walk equals the average over the stationary ensemble.

It is a law of large numbers for dependent variables, and the iid law is the special case of a chain that ignores its present state [CasellaBerger §5.8.5]. Markov introduced his chains for this very purpose: the law of large numbers does not need independence, only forgetting (in 1913 he marked the two-hundredth anniversary of the law with a celebration of his own [Lambert §12.9]). It is the reason MCMC works at all: estimate any posterior expectation by averaging along the chain [Wasserman p. 411]. Two arguments make it believable at our level: one counts returns, the other computes the variance.

9.6.1 Why: counting returns

Take g to be the indicator of one state j, so that the time average is the fraction of time spent in j. Every time the walker arrives at j, the Markov property says the future is a fresh copy of the chain started at j, independent of how it got there. So the waiting times between successive visits, $T^{(1)}, T^{(2)}, \dots$, are independent copies of the return time T_{jj}, with mean m_j (Section 9.3.2). After the kth visit the elapsed time is $T^{(1)} + \cdots + T^{(k)}$ (plus the time to the first visit, which is finite and does not grow with k), so

$$\frac{\text{visits to } j \text{ in } n \text{ steps}}{n} \overset{(1)}{=} \frac{k}{T^{(1)}+\cdots+T^{(k)}} \overset{(2)}{=} \left(\frac{T^{(1)}+\cdots+T^{(k)}}{k}\right)^{-1} \xrightarrow[k\to\infty]{} \frac{1}{m_j}. \tag{9.42}$$

(1) Choose n at the kth visit; between visits the count does not change, and this changes the fraction by at most $1/k$.

(2) The ordinary law of large numbers for the iid return times: their average tends to m_j.

The long-run fraction of time in j exists and equals $1/m_j$, whatever the start, since an irreducible finite chain reaches j from anywhere.

Which number is $1/m_j$*?* Now run the same walker from a different start: draw X_0 from a stationary distribution π (one exists for every finite chain, Section 9.4). Then $p_t = \pi$ on every day, and the *expected* fraction of time in j over n steps is exactly

$$\mathbb{E}\left(\frac{\text{visits to } j \text{ in } n \text{ steps}}{n}\right) = \frac{1}{n}\sum_{t=1}^{n} \mathrm{P}(X_t = j) = \frac{1}{n}\sum_{t=1}^{n} \pi_j = \pi_j$$

for every n (the expectation of a count of visits is the sum of the probabilities of a visit, as in (9.19)). The fraction itself tends to $1/m_j$ with probability one, and it always lies between 0 and 1. A sequence of random variables that stays between fixed bounds and converges with probability one also has converging expectations (the bounded convergence theorem: no rare, huge values can hold the mean away from the limit). So the constant expectation π_j equals the limit $1/m_j$, which gives **Kac's formula**

$$\pi_j = \frac{1}{m_j} : \tag{9.43}$$

the equilibrium probability of a state is one over its mean recurrence time. A rare state is one that is revisited only after long excursions. Two by-products come free. The argument never asked whether p_t converges from other starts, so it holds for periodic chains too. And since m_j is fixed by the chain, any stationary distribution must equal $(1/m_1, \dots, 1/m_K)$: an irreducible finite chain has only one.

From one state to any function. On a finite state space any function g is a sum of indicators, $g(x) = \sum_j g(j)\, I(x = j)$, so its time average is a weighted sum of time fractions,

$$\frac{1}{N}\sum_{n=1}^{N} g(X_n) = \sum_j g(j) \times \Big(\text{fraction of the } N \text{ steps spent in } j\Big) \xrightarrow[N\to\infty]{} \sum_j g(j)\,\pi_j = \mathbb{E}_\pi(g),$$

each fraction tending to $\pi_j = 1/m_j$ and the finite sum following term by term. That is (9.41).

Example: how long until it rains again? For the three-state weather chain $\pi_R = \frac{1}{6}$, so the mean time from one rainy day to the next is 6 days. In a simulated history of 10^6 days the average gap between rainy days is 6.007. For the politician, $\pi_1 = \frac{1}{28}$: having left island 1, he is back there on average 28 nights later.

Example: the cycle. The deterministic 3-cycle of Section 9.4.1 returns to each state after exactly 3 steps, so $m_j = 3$ and the time fractions are exactly $\frac{1}{3} = \pi_j$. Yet p_t never converges. The return argument used irreducibility and positive recurrence, but not aperiodicity. That is a general fact: periodicity spoils the convergence of the distribution at a fixed time, not the convergence of time averages. The tilted die that always moves to a neighbouring face (Section 9.3.3) has period 2, and its average face value still tends to 3.5.

9.6.2 How accurate: the variance of a time average

Suppose the chain has already reached equilibrium, $X_1 \sim \pi$, so that every X_t has the same distribution and the covariance $C_k = \mathrm{Cov}\big(g(X_t), g(X_{t+k})\big)$ depends only on the lag k. Write $\bar g_N = \frac{1}{N}\sum_{t=1}^{N} g(X_t)$, $\sigma^2 = C_0$, and $\rho_k = C_k/C_0$ for the autocorrelation at lag k. Then

$$\begin{aligned}\mathrm{Var}(\bar g_N) &\overset{(1)}{=} \frac{1}{N^2}\sum_{s=1}^{N}\sum_{t=1}^{N} C_{|s-t|} \overset{(2)}{=} \frac{1}{N^2}\Big[N C_0 + 2\sum_{k=1}^{N-1}(N-k)\,C_k\Big] \\ &\overset{(3)}{=} \frac{\sigma^2}{N}\Big[1 + 2\sum_{k=1}^{N-1}\Big(1-\frac{k}{N}\Big)\rho_k\Big] \;\simeq\; \frac{\sigma^2}{N}\,\tau, \qquad \tau \equiv 1 + 2\sum_{k=1}^{\infty}\rho_k. \end{aligned} \tag{9.44}$$

(1) Variance of a sum is the sum of all covariances (Chapter 1), with stationarity.
(2) There are N pairs with $s = t$ and $2(N-k)$ ordered pairs with $|s-t| = k$.

(3) Factor out N; for N much larger than the correlation length, $k/N \to 0$ in the terms that matter, and the sum extends to infinity.

Is τ finite? The covariances can be computed from P. Centre g on its equilibrium mean, $g_c = g - \mathbb{E}_\pi\, g$. Given $X_t = i$, the expected value of g_c after k more steps is $\sum_j p_{ij}(k) g_c(j) = (P^k g_c)(i)$, so averaging over $X_t \sim \pi$,

$$C_k = \mathbb{E}\big[g_c(X_t)\, g_c(X_{t+k})\big] = \sum_i \pi_i\, g_c(i)\, (P^k g_c)(i).$$

Expand g_c in the right eigenvectors of P, $g_c = \sum_m b_m v_m$, the same expansion as (9.33) with columns in place of rows (assuming, as there, a full set of eigenvectors). There is no $m = 1$ term, since $v_1 = \mathbf{1}$ and the coefficient along it is $\pi g_c = 0$. Each P^k multiplies v_m by λ_m^k, so

$$C_k = \sum_{m \geq 2} a_m\, \lambda_m^k, \qquad a_m = b_m \sum_i \pi_i\, g_c(i)\, v_m(i),$$

a sum of geometric sequences, one per mode. For a finite ergodic chain every $|\lambda_m| < 1$ ($m \geq 2$), so each sums, and

$$\tau = 1 + 2 \sum_{k \geq 1} \frac{C_k}{C_0} = 1 + 2 \sum_{m \geq 2} \frac{a_m}{C_0}\, \frac{\lambda_m}{1 - \lambda_m},$$

using $\sum_{k \geq 1} \lambda^k = \lambda/(1-\lambda)$. Slow modes ($\lambda_m$ near 1) inflate τ in proportion to how much of g lies along them; negative λ_m reduce it. With τ finite, the variance of the time average falls as $1/N$, and $\bar{g}_N \to \mathbb{E}_\pi\, g$ in mean square. That is a second, quantitative route to (9.41). The price of dependence is the factor τ, the **integrated autocorrelation time**: the chain average is as accurate as an average of N/τ independent draws, which is why N/τ is called the effective number of samples. Estimating τ from the chain itself, when no formula is available, is one of the MCMC diagnostics of this chapter; here we compute it exactly where we can.

Example: two states, exactly. For the two-state chain with $a = p_{12}$, $b = p_{21}$ and $\lambda = 1 - a - b$, let g be the indicator of state 1. From (9.32) with $p_0 = (1, 0)$, $d_0 = 1 - \pi_1 = \pi_2$, so $\mathrm{P}(X_k = 1 \mid X_0 = 1) = \pi_1 + \pi_2 \lambda^k$, and

$$\begin{aligned} C_k &\overset{(1)}{=} \mathrm{P}(X_0 = 1, X_k = 1) - \pi_1^2 \overset{(2)}{=} \pi_1(\pi_1 + \pi_2 \lambda^k) - \pi_1^2 = \pi_1 \pi_2\, \lambda^k, \\ \tau &\overset{(3)}{=} 1 + 2 \sum_{k \geq 1} \lambda^k = 1 + \frac{2\lambda}{1 - \lambda} = \frac{1 + \lambda}{1 - \lambda}. \end{aligned} \tag{9.45}$$

(1) $\mathrm{Cov}(I, I') = \mathbb{E}(II') - \mathbb{E}\, I\ \mathbb{E}\, I'$ for indicators, in equilibrium.
(2) $\mathrm{P}(X_0 = 1) = \pi_1$ times the k-step return probability just found.
(3) $\rho_k = \lambda^k$ since $C_0 = \pi_1 \pi_2$; geometric series, $|\lambda| < 1$.

So the fraction of time in state 1 over N steps has variance $\pi_1 \pi_2 \tau / N$. Two cases make the point (Figure 9.15a). Wasserman's weather has $\lambda = -0.4$ and $\tau = 0.429$: consecutive days are *anti*correlated, sun tends to follow cloud, and the chain average is *more* accurate than an average of independent days. A sticky chain with $a = 0.1$, $b = 0.2$ has $\lambda = 0.7$ and $\tau = 5.667$: it lingers, and needs almost six times as many steps for the same accuracy. Over 4000 independent chains of 1000 steps, the measured N Var of the time fraction is 0.1044 against the predicted 0.105 for the weather, and 1.282 against 1.259 for the sticky chain.

Example: the tilted die. Lambert's tilted die makes the dependence tangible [Lambert §12.10]. A fair six-sided die is "thrown" by tilting it: with a probability ε, the *tilt probability*, it rolls

to one of the two neighbouring faces round the ring $1, 2, \dots, 6, 1$ (each with chance $\frac{1}{2}$), and otherwise it is thrown afresh. Call W the matrix of the roll to a neighbour, with $W_{j,j\pm1} = \frac{1}{2}$ (faces counted round the ring) and zeros elsewhere; a fresh throw is the matrix $\frac{1}{6}\mathbf{1}\mathbf{1}^{\mathsf{T}}$ with every entry $\frac{1}{6}$. The transition matrix is the mixture

$$P = (1-\varepsilon)\,\tfrac{1}{6}\mathbf{1}\mathbf{1}^{\mathsf{T}} + \varepsilon\, W.$$

The uniform distribution is stationary. Each column of $\frac{1}{6}\mathbf{1}\mathbf{1}^{\mathsf{T}}$ holds six entries $\frac{1}{6}$ and sums to 1; each column of W holds two entries $\frac{1}{2}$ and also sums to 1. So every column of P sums to $(1-\varepsilon)+\varepsilon = 1$, and then $(\pi P)_j = \frac{1}{6}\sum_i p_{ij} = \frac{1}{6}$ for π uniform. The time average of the face value therefore converges to 3.5 for every ε.

The slowest mode alternates. Try the vector $v_j = (-1)^j$, $+1$ on even faces and -1 on odd ones. Both neighbours of face j have the opposite sign, so $(Wv)_j = \frac{1}{2}(-v_j) + \frac{1}{2}(-v_j) = -v_j$; and v sums to zero, so the fresh-throw part gives 0. Hence $Pv = (1-\varepsilon)\cdot 0 + \varepsilon(-v) = -\varepsilon v$: an eigenvalue $-\varepsilon$. The other non-trivial modes are the smooth waves round the ring of (9.56), with eigenvalues $\varepsilon\cos(2\pi k/6) = \pm\varepsilon/2$. With $\varepsilon = 0.9$ the largest non-trivial modulus is therefore 0.9, from the alternating mode.

Why τ is small anyway. The face value $g(j) = j$ rises steadily round most of the ring, so it lies mostly along the smoothest waves: about 69% of its variance is in the two modes with $\lambda = +0.45$, 23% in those with $\lambda = -0.45$, and only 9% in the alternating mode with $\lambda = -0.9$. Inserted in the formula for τ above, the positive modes add about 1.1, the negative ones subtract about 0.2, and $\tau = 1.9$, as computed by the function below. A negative eigenvalue anticorrelates successive throws, which helps rather than hurts. After 1000 throws the rms error of the average is 0.0744, against 0.055 for a fair die: a ratio close to $\sqrt{\tau}$ (Figure 9.15b). The lesson is that τ depends on the function being averaged, through how it projects on the slow and fast modes, and not on $|\lambda_2|$ alone.

Listing 9.3: The integrated autocorrelation time of $g(X)$ for a stationary finite chain (`code/ch09/06_ergodic_average.py`)

```
56 def tau_of(P, g, kmax=4000):
57     """Integrated autocorrelation time 1 + 2 sum_k rho_k of g(X) for the stationary
       chain."""
58     pi = stationary(P)
59     gc = g - pi @ g
60     var = pi @ gc**2
61     v, s = gc.copy(), 0.0
62     for _ in range(kmax):
63         v = P @ v                                   # v = E[g(X_k) - mean | X_0]
64         s += (pi * gc) @ v
65     return 1 + 2 * s / var
```

Line 59 centres g on its equilibrium mean. Line 63 is the whole trick: $v = P^k g_c$ is the conditional expectation $\mathbb{E}\big(g_c(X_k) \mid X_0 = i\big)$ for every starting state i, so averaging it against $\pi_i g_c(i)$ in line 64 gives the lag-k covariance C_k. The loop sums C_k over lags, and line 65 assembles τ from (9.44).

One long history. Figure 9.16 follows the fraction of rainy days along a single weather history of 10^6 days that starts on a rainy day. The early fraction is dominated by the start and by chance; after a few hundred days it lives inside the band $\pi_R \pm 2\sqrt{\pi_R(1-\pi_R)\tau/N}$, with $\tau = 2.222$ for the rain indicator, and at the end it is 0.1665, against $\frac{1}{6} = 0.1667$, with a predicted standard error of 5.6×10^{-4}. The starting point has been forgotten twice over: its influence on the distribution decays as 0.6^t, and its weight in the running average is diluted as $1/N$.

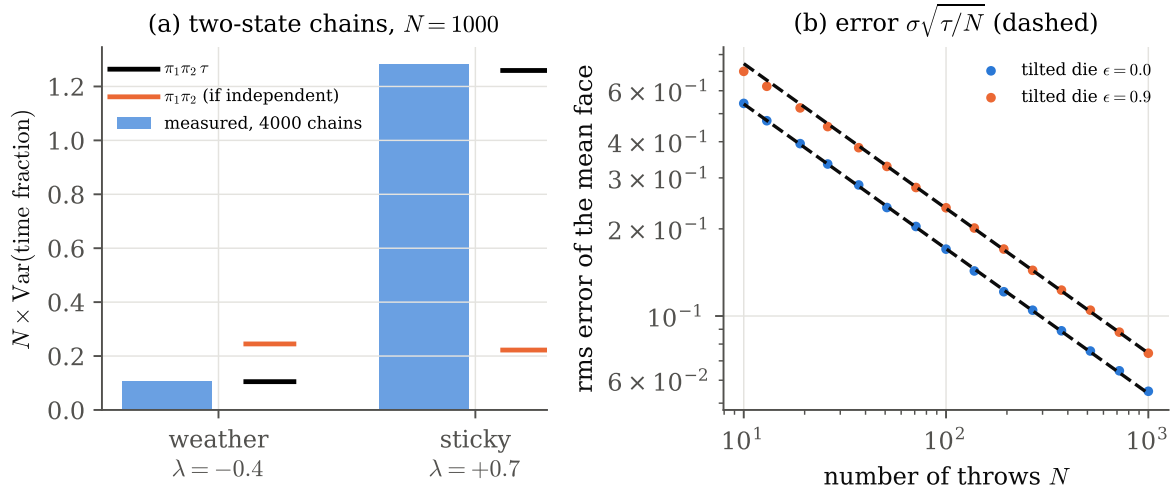


Figure 9.15: The cost (or benefit) of correlated steps. (a) N times the variance of the fraction of time spent in state 1, measured over 4000 chains of 1000 steps (bars), against $\pi_1\pi_2\tau$ (black ticks) and the value for independent draws (orange ticks). The anticorrelated weather chain beats independent sampling; the sticky chain is almost six times worse. (b) The rms error of the average face of the tilted die after N throws, for a fair die ($\varepsilon = 0$) and a strongly tilted one ($\varepsilon = 0.9$); both follow $\sigma\sqrt{\tau/N}$ (dashed).

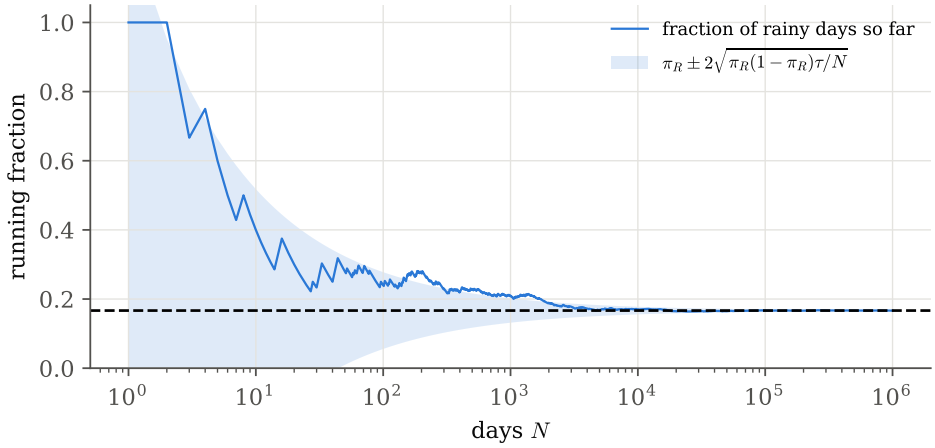


Figure 9.16: The fraction of rainy days so far along one simulated weather history, started on a rainy day, on a logarithmic time axis. It settles on $\pi_R = \frac{1}{6}$ (dashed) and stays within the two-standard-error band predicted by (9.44).

How much the start biases a finite run. Started out of equilibrium, from a distribution p_0, the chain's expected value of g at step t is $p_t g = \sum_j p_t(j) g(j)$. By the spectral expansion (9.33), $p_t = \pi + \sum_{k\geq 2} c_k \lambda_k^t \ell_k$, with c_k the weights of the starting distribution along the left eigenvectors ℓ_k, so

$$\mathbb{E}\, g(X_t) - \mathbb{E}_\pi\, g = \sum_{k\geq 2} c_k\, \lambda_k^t\, (\ell_k g) \simeq c\, \lambda_2^t, \qquad c \equiv c_2\, (\ell_2 g),$$

where at late times the slowest mode dominates. The amplitude c is the product of how far the start sits along the slowest mode (c_2) and how strongly g responds to that mode ($\ell_2 g$). (For a

reversible chain λ_2 is real, Section 9.4.2.) Averaging over the N steps of the run,

$$\mathbb{E}(\bar{g}_N) - \mathbb{E}_\pi\, g \overset{(1)}{=} \frac{1}{N}\sum_{t=1}^{N}\big(\mathbb{E}\, g(X_t) - \mathbb{E}_\pi\, g\big) \overset{(2)}{=} \frac{c}{N}\sum_{t=1}^{N}\lambda_2^t \overset{(3)}{=} \frac{c}{N}\sum_{t=1}^{\infty}\lambda_2^t \overset{(4)}{=} \frac{c\,\lambda_2}{N(1-\lambda_2)}. \tag{9.46}$$

(1) The mean of an average is the average of the means.
(2) The transient of each step, keeping the slowest mode.
(3) The terms beyond $t = N$ add $\lambda_2^{N+1}/(1-\lambda_2)$, negligible once N is many relaxation times.
(4) Geometric series, $\sum_{t\geq 1}\lambda^t = \lambda/(1-\lambda)$ for $|\lambda| < 1$.

The start causes a bias of order $1/N$, while the statistical error $\sigma\sqrt{\tau/N}$ of (9.44) is of order $1/\sqrt{N}$. For a long chain the start is harmless: the bias shrinks faster than the noise. For a short chain started far from the bulk of π, where c is large and λ_2 close to 1, it is not, and practitioners discard the first part of the chain, the **burn-in**.

The conventions of a CMB analysis. Dunkley et al. asked a practical question [171]: how long must a chain for cosmological parameters be, and how much of its start must be thrown away, before the table of means and errors it produces stops changing when the chain is run longer? Their answers are two rules of thumb, and both follow from the formulas above.

Stop when the chain mean is good to a tenth of an error bar. They ask for the variance of the chain mean to be below one percent of the posterior variance. Their convergence ratio r, the variance of the chain mean divided by the variance of the parameter, is in our symbols $\mathrm{Var}(\bar{g}_N)/\sigma^2 = \tau/N$ by (9.44), with g the parameter. So

$$\frac{\sigma^2\tau}{N} \leq 0.01\,\sigma^2 \quad\Longleftrightarrow\quad \frac{N}{\tau} \geq 100:$$

at least about 100 effective samples per parameter, and then the Monte Carlo error of a posterior mean is a tenth of its posterior error bar. Their chain length is our N, and their correlation length plays the role of our τ.

Discard until the density is within a factor 10 of its peak. Their burn-in rule removes the start of the chain until the target density has risen to a tenth of its maximum. For a Gaussian posterior in one parameter the density relative to the peak is $e^{-z^2/2}$ at z standard deviations, so

$$e^{-z^2/2} = 0.1 \quad\Longleftrightarrow\quad z = \sqrt{2\ln 10} \approx 2.15.$$

A walker that has reached that level is within about two standard deviations of the peak, in the bulk of the posterior, where c in (9.46) is no longer large. In d parameters the same rule means $\Delta\chi^2 = 2\ln 10 \approx 4.6$ from the best fit, and since typical posterior points sit at $\Delta\chi^2 \approx d$, for a handful of parameters this is already among typical points.

The toy chains of this section check the ingredients of both rules, τ and the λ_2^t transient, against exact values; a CMB chain has to estimate them from its own output.

How this is used in real research: Chain averages as posterior summaries

Every number in a cosmological parameter table produced by MCMC (CosmoMC, Cobaya, MontePython, `emcee`) is a time average (9.41) along a few long chains: a posterior mean is the chain mean of θ, a credible interval is a pair of chain quantiles, a marginal is the histogram of one column. Its Monte Carlo error is $\sigma\sqrt{\tau/N}$, and the stopping rule $N/\tau \gtrsim 100$ effective samples per parameter derived above keeps it at a tenth of the posterior error. In the toy chains above, the Monte Carlo *verifies* formulas we derived (τ, Kac, $|\lambda_2|^t$); in a

real analysis τ and λ_2 are unknown and must be estimated from the chain itself. Scale is the other difference. Our chains have 10^3–10^6 steps of a model that costs nanoseconds, and run in seconds on a laptop; a Planck analysis runs 4–8 chains of 10^5–10^6 steps with a Boltzmann code costing about a second per step, which takes days on a cluster. A laptop-scale CMB analysis keeps the run short by replacing the Boltzmann code with an emulated or Taylor-expanded spectrum, at the price of a restricted parameter range and some loss of accuracy.

9.7 The walker that never settles: free diffusion

Take Kruschke's politician of Section 9.2 and remove two things: the ends of the archipelago and his habit of comparing populations. He now lives on an endless chain of islands, flips a coin every morning, and moves to the left or right neighbour as the coin says, whatever the islands look like. Make the islands into a continuous coastline and the coin into a Gaussian, and this is the simplest continuous Markov chain there is: the **unbiased random walk**

$$X_{t+1} = X_t + s\,Z_{t+1}, \qquad Z_1, Z_2, \ldots \text{ iid } \mathcal{N}(0,1), \tag{9.47}$$

where s is the step size and each Z_t is a fresh standard Gaussian draw. Its transition kernel (Section 9.4.3, the density of the next position y given the current position x) is $p(x,y) = \varphi_s(y-x)$, with φ_s the $\mathcal{N}(0, s^2)$ density. It is what a Markov chain sampler does if we keep its proposals and throw away every rule that refers to the target. It is also a Brownian particle observed once per time step.

Every result of the previous sections needed some pull towards a target distribution π: the politician's acceptance rule, the weather's fixed transition probabilities on a finite set of states. Here there is none, and the walker is free to wander over the whole real line.

The questions.

1. Where is the walker after t steps, and does the histogram of many walkers settle down?
2. Is there a stationary density, one that the walk would leave unchanged?
3. Does the walk at least forget its starting point?

9.7.1 Where the walker is after t steps

Unrolling (9.47) from a fixed start $X_0 = x_0$ gives the position as the start plus a sum of independent kicks:

$$\begin{aligned} X_t &\overset{(1)}{=} x_0 + s\,(Z_1 + Z_2 + \cdots + Z_t), \\ \mathbb{E}\,X_t &\overset{(2)}{=} x_0, \qquad \operatorname{Var} X_t \overset{(3)}{=} s^2\big(\operatorname{Var} Z_1 + \cdots + \operatorname{Var} Z_t\big) = s^2 t, \\ X_t &\sim \mathcal{N}\big(x_0,\, s^2 t\big). \end{aligned} \tag{9.48}$$

(1) Apply (9.47) t times: $X_t = X_{t-1} + sZ_t = X_{t-2} + sZ_{t-1} + sZ_t = \ldots$
(2) Each Z_k has mean zero.
(3) The kicks are independent, so their variances add (Chapter 1); each Z_k has variance 1.

The last line holds because a sum of independent Gaussians is Gaussian (Chapter 2). The centre never moves, because the walk is unbiased; the width grows as $s\sqrt{t}$, without limit. This is the

random-walk law of Brownian motion, $\langle (X_t - x_0)^2 \rangle \propto t$, which Einstein used to connect the jiggling of pollen grains to the size of molecules.

The same result can be read as an equation for the density of walkers. One step convolves the density with the kernel, $f_{t+1}(y) = \int f_t(x)\, \varphi_s(y-x)\, \mathrm{d}x$ (Eq. (9.29)). Substitute $x = y - u$ and expand f_t around y for small steps:

$$\begin{aligned} f_{t+1}(y) &\overset{(1)}{=} \int f_t(y-u)\, \varphi_s(u)\, \mathrm{d}u \\ &\overset{(2)}{=} \int \left[f_t(y) - u\, f_t'(y) + \tfrac{1}{2} u^2 f_t''(y) - \dots \right] \varphi_s(u)\, \mathrm{d}u \\ &\overset{(3)}{=} f_t(y) + \tfrac{1}{2} s^2 f_t''(y) + O\left(s^4 f_t''''\right). \end{aligned} \tag{9.49}$$

(1) Change of variable $u = y - x$; φ_s is even, so $\varphi_s(y-x) = \varphi_s(u)$.
(2) Taylor expansion of f_t about y, valid when f_t varies little over one step s.
(3) The kernel integrates to 1, has mean $\int u\, \varphi_s\, \mathrm{d}u = 0$ and variance $\int u^2 \varphi_s\, \mathrm{d}u = s^2$; the next surviving term is the fourth moment, $3s^4$.

From steps to a clock. Suppose the walker takes one step every Δt seconds. Call the step number n (it was t in (9.49)), so that after n steps the physical time is $t = n\,\Delta t$, and write $f(x,t)$ for the density of walkers at that time. For small steps taken often, the density changes little from one step to the next, and

$$\Delta t\, \frac{\partial f}{\partial t} \overset{(1)}{=} f(x, t+\Delta t) - f(x,t) \overset{(2)}{=} f_{n+1}(x) - f_n(x) \overset{(3)}{=} \tfrac{1}{2} s^2\, \frac{\partial^2 f}{\partial x^2}.$$

(1) The first-order Taylor expansion in time, valid when f changes little during one step.
(2) One step of the clock is one step of the walk.
(3) (9.49), dropping the higher-order term.

Dividing by Δt gives the **diffusion equation**

$$\frac{\partial f}{\partial t} = D\, \frac{\partial^2 f}{\partial x^2}, \qquad D = \frac{s^2}{2\,\Delta t}, \tag{9.50}$$

the heat equation, with the **diffusion coefficient** D collecting the step size and the step rate. Its solution from a point is the spreading Gaussian of (9.48), and the two descriptions agree on the width: after $n = t/\Delta t$ steps the variance is $s^2 n = s^2 t/\Delta t = 2Dt$. A density of walkers behaves exactly like a drop of ink in still water.

Example: twenty thousand walkers. Figure 9.17 follows 20 000 independent walkers with $s = 1$, all started at the origin. The histograms at $t = 10, 100, 1000$ lie on the Gaussians $\mathcal{N}(0, t)$, and the sample variance is 100.7 at $t = 100$ and 1014 at $t = 1000$, against $s^2 t = 100$ and 1000. The core of the simulation is eight lines:

Listing 9.4: The unbiased random walk for many walkers at once (`code/ch09/07_random_walk_diffusion.py`)

```
x = np.zeros(W)
var_t, keep = [], {}
for t in range(1, T + 1):
    x += s * rng.standard_normal(W)
    var_t.append(x.var())
    if t in snaps:
        keep[t] = x.copy()
var_t = np.array(var_t)
```

Line 20 puts every walker at the origin. Line 23 is (9.47) applied to all of them in one vectorised step: each walker gets its own fresh Gaussian kick. Line 24 records the spread of the crowd, and lines 25–26 keep three snapshots for the histograms. Nothing in the loop refers to any target; the walkers obey the kernel and nothing else.

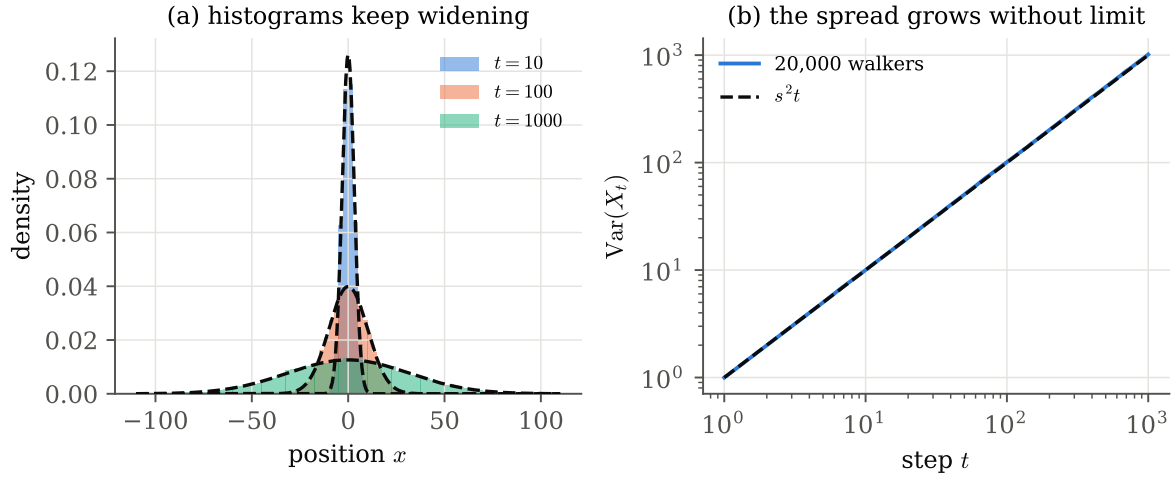


Figure 9.17: The unbiased walker never settles. (a) Histograms of 20 000 walkers started at the origin, after 10, 100 and 1000 steps of size $s = 1$, with the Gaussians $\mathcal{N}(0, s^2 t)$ (dashed): the crowd keeps flattening. (b) The variance of the crowd grows in proportion to the number of steps, on a log–log scale, with no sign of levelling off.

9.7.2 No stationary density

The histograms flatten for ever, which suggests that there is nothing for them to converge to. Let us prove it. A stationary density f would satisfy $f = f * \varphi_s$, a convolution. The convolution theorem turns this into a product for the characteristic functions $\chi_f(k) = \mathbb{E}\, e^{ikX} = \int f(x) e^{ikx}\, \mathrm{d}x$, the Fourier transforms of the densities, and the characteristic function of the Gaussian kick is $e^{-s^2k^2/2}$ (Chapter 2). So

$$\chi_f(k) \overset{(1)}{=} \chi_f(k)\, e^{-s^2k^2/2} \quad \Longrightarrow \quad \chi_f(k)\big(1 - e^{-s^2k^2/2}\big) \overset{(2)}{=} 0 \quad \Longrightarrow \quad \chi_f(k) \overset{(3)}{=} 0 \ \text{ for every } k \neq 0. \tag{9.51}$$

(1) $X_{t+1} = X_t + sZ$ with $X_t \sim f$ independent of Z, so $\mathbb{E}\, e^{ikX_{t+1}} = \mathbb{E}\, e^{ikX_t}\ \mathbb{E}\, e^{iksZ}$, and stationarity demands $X_{t+1} \sim f$.
(2) Move everything to one side.
(3) The bracket is nonzero for $k \neq 0$.

But every probability density has $\chi_f(0) = \int f = 1$ and a characteristic function that is continuous in k. One that is zero for every $k \neq 0$ cannot jump to 1 at $k = 0$. So no probability density is stationary for the free walk.

Detailed balance points to the same answer from another side. The kernel is symmetric, $p(x, y) = \varphi_s(y - x) = p(y, x)$, so the detailed-balance condition $f(x)p(x, y) = f(y)p(y, x)$ of Section 9.4.3 reduces to $f(x) = f(y)$: the only balanced "distribution" is a constant, the flat measure over the whole line. It cannot be normalised. The unbiased walker is, in this precise sense, trying to sample a uniform distribution on $\mathbb{R}$, and there is no such distribution. In the language of Section 9.3.2, the one-dimensional walk is null recurrent: it returns to every neighbourhood, but only after excursions so long that the mean return time m is infinite, and Kac's formula (9.43), $\pi = 1/m$, assigns every region probability zero.

9.7.3 Forgetting without settling

The third question is subtler. Start one crowd at $x_0 = 0$ and another at $x_0 = a$. After t steps they are $\mathcal{N}(0, \sigma_t^2)$ and $\mathcal{N}(a, \sigma_t^2)$ with $\sigma_t = s\sqrt{t}$. How different are they? For two Gaussians of equal width the densities cross at the midpoint $a/2$, so the event on which the first exceeds the second is $\{x < a/2\}$, and the total-variation distance (9.30) (written for densities, $\frac{1}{2}\int |p - q| \, \mathrm{d}x$) is

$$\begin{aligned} \big\| \mathcal{N}(0, \sigma_t^2) - \mathcal{N}(a, \sigma_t^2) \big\|_{\mathrm{TV}} &\overset{(1)}{=} \mathrm{P}_0\Big(X < \frac{a}{2}\Big) - \mathrm{P}_a\Big(X < \frac{a}{2}\Big) \overset{(2)}{=} \Phi\Big(\frac{a}{2\sigma_t}\Big) - \Phi\Big(-\frac{a}{2\sigma_t}\Big) \\ &\overset{(3)}{=} 2\,\Phi\Big(\frac{a}{2\sigma_t}\Big) - 1 \;\simeq\; \frac{a}{\sigma_t\sqrt{2\pi}} = \frac{a}{s\sqrt{2\pi t}}. \end{aligned} \tag{9.52}$$

(1) The maximum of $|p(A) - q(A)|$ is reached on the set where $p > q$ (Section 9.5), here $\{x < a/2\}$.
(2) Standardise: under $\mathcal{N}(\mu, \sigma^2)$, $\mathrm{P}(X < c) = \Phi\big((c - \mu)/\sigma\big)$, with Φ the standard normal CDF.
(3) $\Phi(-z) = 1 - \Phi(z)$; then for small z, $\Phi(z) \simeq \frac{1}{2} + z/\sqrt{2\pi}$, since the standard normal density at 0 is $1/\sqrt{2\pi}$.

The distance goes to zero: the walk does forget where it started. Two crowds launched from different islands become indistinguishable, because both spread so widely that the initial offset is a negligible fraction of their width. But the forgetting is slow, like $1/\sqrt{t}$ rather than geometrically, and, more importantly, the two crowds do not converge to anything. They just become equally flat.

Watch out: Forgetting the start is not converging

A chain can lose all memory of its initial distribution and still have no equilibrium to offer. The free walk does exactly that. For a sampler both properties are needed: a normalisable stationary distribution π (for a countable chain, positive recurrence) and convergence to it. Time averages along a free walk do not estimate anything: the fraction of time spent in any bounded interval goes to zero.

9.7.4 The walker with a pull

Add a restoring pull to the free walk and everything comes back. The AR(1) chain of Section 9.4.3 keeps only a fraction ρ of the current position before the kick,

$$X_{t+1} = \rho\, X_t + \sqrt{1 - \rho^2}\; Z_{t+1}, \qquad |\rho| < 1, \tag{9.53}$$

which is a Brownian particle in a harmonic well, observed at discrete times; we showed there that it satisfies detailed balance with respect to $\mathcal{N}(0, 1)$. Unrolling it from $X_0 = x_0$, the same way as (9.48),

$$\begin{aligned} X_t &\overset{(1)}{=} \rho^t x_0 + \sqrt{1 - \rho^2} \sum_{k=0}^{t-1} \rho^k Z_{t-k}, \\ \mathbb{E}\, X_t &\overset{(2)}{=} \rho^t x_0, \qquad \operatorname{Var} X_t \overset{(3)}{=} (1 - \rho^2) \sum_{k=0}^{t-1} \rho^{2k} \overset{(4)}{=} 1 - \rho^{2t}. \end{aligned} \tag{9.54}$$

(1) Substitute (9.53) into itself t times: the start is multiplied by ρ at every step, and the kick received k steps ago by ρ^k.
(2) The kicks have mean zero.

(3) Independent kicks, each of variance 1, multiplied by $\sqrt{1-\rho^2}\,\rho^k$.
(4) Geometric sum, $\sum_{k=0}^{t-1}\rho^{2k} = (1-\rho^{2t})/(1-\rho^2)$.

So $X_t \sim \mathcal{N}(\rho^t x_0,\, 1-\rho^{2t}) \to \mathcal{N}(0,1)$ for every start. Compare with the free walk: the width now saturates at 1, and the offset of the start shrinks as ρ^t. Two crowds started at x_0 and x_0' have the same width $\sqrt{1-\rho^{2t}}$ and centres $\rho^t x_0$ and $\rho^t x_0'$, so (9.52) with offset $a = \rho^t\,|x_0 - x_0'|$ gives a distance $\simeq \rho^t\,|x_0 - x_0'|\,/\sqrt{2\pi}$ once $\rho^{2t} \ll 1$: geometric forgetting at rate ρ.

The rate ρ plays the part of λ_2 of Section 9.5.2. For a finite chain λ_2 belonged to a right eigenvector v, with $(Pv)_i = \mathbb{E}\big(v(X_1)\,|\,X_0 = i\big) = \lambda_2 v_i$. The continuous analogue is a function g whose conditional expectation one step later is a multiple of itself:

$$\begin{aligned}
\mathbb{E}\big(X_1 \,|\, X_0 = x\big) &\overset{(1)}{=} \rho\, x,\\
\mathbb{E}\big(X_1^2 - 1 \,|\, X_0 = x\big) &\overset{(2)}{=} \rho^2 x^2 + (1-\rho^2) - 1 = \rho^2\,(x^2-1),\\
\mathbb{E}\big(X_1^3 - 3X_1 \,|\, X_0 = x\big) &\overset{(3)}{=} \rho^3 x^3 + 3\rho x(1-\rho^2) - 3\rho x = \rho^3\,(x^3 - 3x).
\end{aligned} \tag{9.55}$$

(1) The kick has mean zero.
(2) $\mathbb{E}\big((\rho x + \sqrt{1-\rho^2}Z)^2\big) = \rho^2x^2 + 2\rho x\sqrt{1-\rho^2}\,\mathbb{E}\,Z + (1-\rho^2)\,\mathbb{E}\,Z^2$, with $\mathbb{E}\,Z = 0$ and $\mathbb{E}\,Z^2 = 1$.
(3) Write $\sigma = \sqrt{1-\rho^2}$ and expand the cube: $(\rho x + \sigma Z)^3 = \rho^3x^3 + 3\rho^2x^2\sigma Z + 3\rho x\sigma^2Z^2 + \sigma^3Z^3$, with $\mathbb{E}\,Z = \mathbb{E}\,Z^3 = 0$ (the Gaussian is symmetric) and $\mathbb{E}\,Z^2 = 1$; the $-3X_1$ term has expectation $-3\rho x$. The terms in x then combine to $3\rho x(1-\rho^2-1) = -3\rho^3 x$.

So $g_1(x) = x$ is an eigenfunction with eigenvalue ρ, $g_2(x) = x^2 - 1$ one with eigenvalue ρ^2, and $g_3(x) = x^3 - 3x$ one with eigenvalue ρ^3. Each is the power x^n minus lower powers chosen to cancel the extra terms the kick produces. These are the Hermite polynomials, and the pattern continues: the nth Hermite polynomial has eigenvalue ρ^n (Mehler's formula, not proved here). The slowest non-trivial mode decays as ρ^t, exactly the rate found for the mean in (9.54). The variance lives in the $x^2 - 1$ mode: by (9.54) its distance from the target value 1 is $1 - \operatorname{Var} X_t = \rho^{2t} = (\rho^2)^t$, the eigenvalue ρ^2 of that mode raised to the power t, so it relaxes twice as fast in the exponent.

Example: three starts, one law. With $\rho = 0.9$, Figure 9.18 starts three crowds of 20 000 walkers at $x_0 = -20$, 0 and $+20$, twenty standard deviations of the target apart. The means follow $\rho^t x_0$: from $x_0 = 20$ the crowd's mean after 40 steps is 0.29, against $20 \times 0.9^{40} = 0.296$. After 80 steps the three histograms lie on top of $\mathcal{N}(0,1)$; the crowd started at $+20$ has mean -0.005 and variance 1.016, against 0 and $1 - 0.9^{160} \simeq 1$. Everything a sampler needs is present: a normalisable stationary law, and geometric forgetting of the start.

Example: the free walk on a ring. Is it the randomness or the endless line that stops the free walker from settling? Wrap the line into a ring of K islands, and let the walker stay put with probability $\frac{1}{2}$ and step to each neighbour with probability $\frac{1}{4}$ (staying put removes the period 2 of Section 9.3.3). There is still no bias anywhere. The rows of the transition matrix sum to one and so do its columns (each island receives $\frac{1}{2}$ from itself and $\frac{1}{4}$ from each neighbour), so the uniform distribution $\pi_j = 1/K$ is stationary: $\sum_i \pi_i p_{ij} = \frac{1}{K}\sum_i p_{ij} = \frac{1}{K}\cdot 1$. The eigenvectors are the Fourier modes $v_j = e^{2\pi i\, jk/K}$, $k = 0, \dots, K-1$, the normal modes of a ring of coupled oscillators:

$$\begin{aligned}
(Pv)_j &\overset{(1)}{=} \tfrac{1}{2}v_j + \tfrac{1}{4}v_{j+1} + \tfrac{1}{4}v_{j-1} \overset{(2)}{=} v_j\Big[\tfrac{1}{2} + \tfrac{1}{4}\big(e^{2\pi ik/K} + e^{-2\pi ik/K}\big)\Big]\\
&\overset{(3)}{=} v_j\,\frac{1+\cos(2\pi k/K)}{2} \overset{(4)}{=} v_j\cos^2\frac{\pi k}{K}.
\end{aligned} \tag{9.56}$$

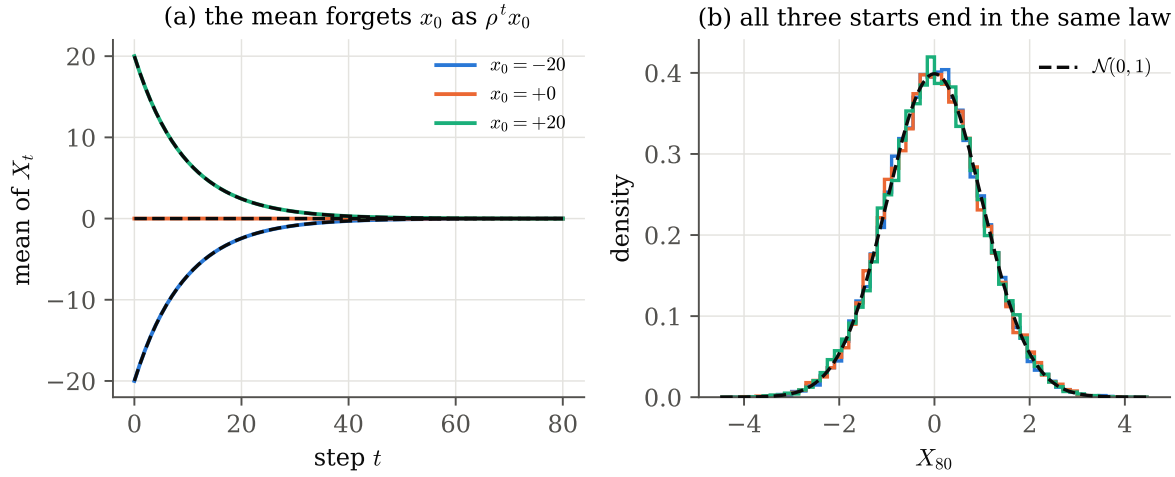


Figure 9.18: The walker with a pull, $X_{t+1} = 0.9X_t + \sqrt{1-0.81}\,Z$, forgets where it started. (a) The mean position of three crowds started at −20, 0 and +20 (solid) follows $\rho^t x_0$ (dashed) back to zero. (b) After 80 steps the three crowds have the same histogram, the standard Gaussian (dashed).

(1) One step: stay with probability $\frac{1}{2}$, move to either neighbour with $\frac{1}{4}$ (indices modulo K, which is the ring).
(2) $v_{j\pm1} = v_j\, e^{\pm 2\pi i k/K}$.
(3) $e^{i\phi} + e^{-i\phi} = 2\cos\phi$.
(4) $1 + \cos 2\alpha = 2\cos^2\alpha$.

Which mode is slowest? The eigenvalue $\cos^2(\pi k/K)$ lies between 0 and 1, and it is largest when the angle $\pi k/K$ is closest to 0 or to π. The mode $k = 0$ gives exactly 1: it is the constant vector, the uniform distribution itself. Among the others, the angle is closest to 0 or π for $k = 1$ and $k = K - 1$, the longest waves round the ring, so $\lambda_2 = \cos^2(\pi/K)$ (shared by the two). The distance from uniform decays as $\cos^{2t}(\pi/K)$, and the relaxation time is

$$\frac{1}{1-\lambda_2} = \frac{1}{\sin^2(\pi/K)} \simeq \frac{K^2}{\pi^2} \text{ steps},$$

using $1 - \cos^2 = \sin^2$ and, for large K, $\sin x \simeq x$ for the small angle $x = \pi/K$. For $K = 10$ islands, $\sin 18^\circ = 0.309$, $\sin^2 18^\circ = 0.0955$, and $1/0.0955 \simeq 10.5$ steps; for $K = 100$, $1/\sin^2 1.8^\circ \simeq 1013$, against $K^2/\pi^2 = 1013$. On a finite ring the free walker does settle, and its target is the flat distribution it was "trying" to sample all along. But the time it takes grows as the *square* of the size of the space, the signature of diffusion: to cross K sites with steps of size one takes about K^2 steps.

What diffusion costs a sampler. The ring makes a point that every random-walk sampler pays for. A chain that explores by unbiased local steps of size s needs about $(L/s)^2$ steps to cross a region of size L. Large steps would cross faster, but on a peaked target most of them land where the probability is tiny. The politician's acceptance rule of Section 9.2 and the Metropolis–Hastings rule built from detailed balance keep the unbiased proposal of (9.47) and add exactly the bias that is missing here, turning the unbiased walker into the biased one of Section 9.1: a step towards higher probability is always taken, a step away from it only sometimes, so that the walker lingers where the target is large [Kruschke §7.2.1], [Wasserman p. 411]. The diffusive cost then sets the length of every MCMC run. In the laptop examples of this chapter it means 10^4–10^5 steps of a fast model; a full cosmological analysis, with a dozen

correlated parameters and an expensive likelihood, needs the same order of steps per chain but at a second per step, which is why it runs on a cluster.

Recap

- A sample is the distribution for every practical purpose. An oscillator of fixed energy samples its own orbit when looked at at random instants, because its motion covers the orbit at the right rate; time average equals ensemble average. Many coupled degrees of freedom, a canonical ensemble or a posterior offer no such motion, so we build one.
- A Markov chain remembers one step: the next state depends only on the current one, through a transition matrix P (rows are pmfs) or kernel $p(x, y)$. A distribution of walkers evolves as $p_{t+1} = p_t P$, and P^n gives the n-step probabilities (Chapman–Kolmogorov). The MLE of p_{ij} from one history is n_{ij}/n_i.
- States split into communicating classes. A good sampler is irreducible (every state reachable from every other), positive recurrent (finite mean return time) and aperiodic. The symmetric walk on the integers is recurrent but null; biased, it is transient.
- π is stationary if $\pi P = \pi$: probability flowing in balances probability flowing out. Detailed balance, $\pi_i p_{ij} = \pi_j p_{ji}$, is a sufficient condition with no net current between any pair, the analogue of thermodynamic equilibrium; a circulating current can be stationary without it.
- A finite irreducible aperiodic chain forgets its starting distribution geometrically: $\|p_t - \pi\|_{\mathrm{TV}} \simeq C\,|\lambda_2|^t$, and the Doeblin contraction bounds it by $(1-\delta)^t$; coupling gives the same result by letting two walkers meet.
- The ergodic theorem: time averages along one walk converge to averages over π, also for periodic chains. Their variance is $\sigma^2\tau/N$ with $\tau = 1 + 2\sum_k \rho_k$, so N steps are worth N/τ independent draws; and $\pi_j = 1/m_j$ (Kac).
- The unbiased random walk on the line spreads as $s\sqrt{t}$, obeys the diffusion equation, and has no normalisable stationary density: it forgets its start, like $1/\sqrt{t}$, but has nothing to converge to. A restoring pull (AR(1)) gives a stationary $\mathcal{N}(0,1)$ and forgetting at rate ρ^t. The missing bias is what Metropolis–Hastings supplies.

9.8 From the drunkard to the biased walker

The free walk of Section 9.7 has a walker on the real line who, at every tick of a clock, takes a step $x_{t+1} = x_t + s\,z_t$ with z_t a fresh standard Gaussian number and s a fixed step length. That walker never settles. His spread grows like $s\sqrt{t}$, his distribution flattens out forever, and there is no normalisable density for him to converge to. He forgets where he started, but he has nothing to remember instead. We will call him the **drunkard**, or the unbiased random walker: every direction is as good as every other, because nothing in his rule knows about any particular distribution.

Now give the walker a job. Suppose the data of an experiment have left us with a posterior density for a parameter x, and that this posterior has two bumps:

$$f(x) = 0.3\,\mathcal{N}\big(x; -2, 0.6^2\big) + 0.7\,\mathcal{N}\big(x; 2, 0.8^2\big), \tag{9.57}$$

where $\mathcal{N}(x; \mu, \sigma^2)$ is the Gaussian density with mean μ and variance σ^2. A sign ambiguity does

this in practice: the data fix $|x|$ well and the sign of x only weakly. In a real problem we would not know that f is a sum of two Gaussians. All we would have is a routine that returns $f(x)$, or rather $\ln f(x)$ up to an additive constant, for any x we ask about. We cannot draw from f directly, but we would like a long list of numbers whose histogram is f: then every average, interval and marginal we want is a count over the list, as in Chapter 6.

The question. Keep the drunkard's way of suggesting a move: from the present position x, propose $y = x + s\,z$. Add one decision, "go to y or stay at x". Which decision rule makes the walker spend, in the long run, a fraction of his time near each x that is proportional to $f(x)$?

Before trying candidates, it helps to list what we are able to do at each step [Kruschke §7.2.4]: we can generate the proposal y; we can evaluate f at x and at y, so in particular the ratio $f(y)/f(x)$; and we can draw a uniform random number U in $[0, 1]$. Any rule must be built from these three things.

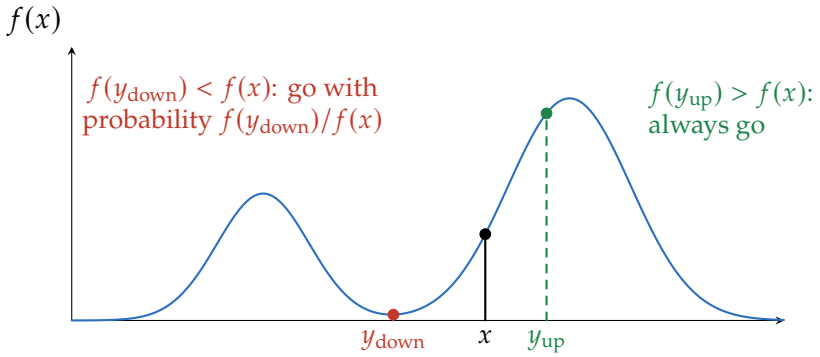


Figure 9.19: The Metropolis decision on the two-bump density. From the present point x the walker proposes a symmetric random step. A proposal to a higher density (green) is always taken. A proposal to a lower density (red) is taken with probability equal to the ratio of the two heights, so that small drops are usually taken and deep drops almost never.

Three candidate rules. We take the natural candidates one at a time and ask, for each, what its walker would do. Figure 9.20 shows the three walkers started at $x_0 = 0$ with step $s = 1.5$, for 50 000 steps.

Always go. This rule ignores f completely, so it cannot know where f is large. It is the drunkard again. Its walker drifts off and never comes back: after 50 000 steps it is at -38.75, a typical distance for a walk whose spread is $s\sqrt{t} \simeq 335.4$ (Figure 9.20a). Its histogram has nothing to do with f and keeps changing as the walk goes on.

Go only uphill. This rule uses f as much as possible: a proposal is taken only if $f(y) \geq f(x)$. Its walker climbs the nearest hill and stays at the top, because once it is at a maximum every proposal is downhill. Of all proposals only a fraction 2.200×10^{-4} were taken, and the walker ended at $x = 2.000$, the top of the right bump (Figure 9.20c). This is an optimiser, not a sampler: it finds the most probable value (here only the local one) and reports nothing about the spread around it.

Go uphill always, downhill sometimes. Between ignoring f and obeying it slavishly there is a middle way. Take every uphill proposal, and take a downhill one with a probability that shrinks the further down it goes: probability $f(y)/f(x)$, the ratio of the two heights (Figure 9.19). A step to a point half as probable is taken half the time; a step to a point a thousand times less probable is almost never taken. The walker now wanders, but it lingers where f is high. After discarding the first 1000 steps, its histogram lies on f (Figure 9.20b): it spent a fraction 0.7083 of its time at

$x > 0$, where (9.57) puts 0.7, its mean is 0.8518 against the exact $0.3 \times (-2) + 0.7 \times 2 = 0.8$, and its variance is 3.84 against 3.916. It accepted a fraction 0.5426 of its proposals.

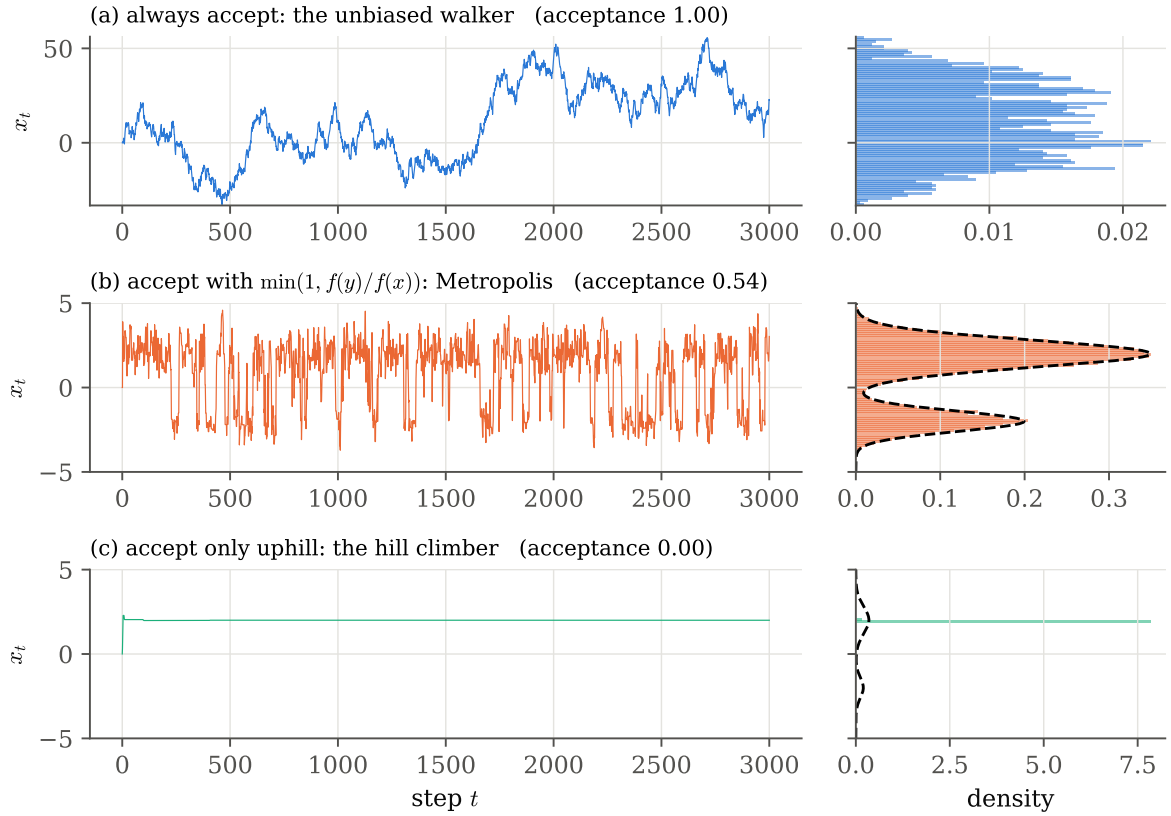


Figure 9.20: Three walkers with the same Gaussian proposals and three decision rules, on the two-bump density (9.57). Left: the first 3000 steps. Right: histograms, with the target dashed. (a) The drunkard wanders off; its histogram is not a density of anything stable. (b) The Metropolis walker hops between the bumps and its long-run histogram is the target. (c) The hill climber reaches the top of the nearest bump and freezes there.

Why the ratio $f(y)/f(x)$, and not, say, its square? Here is the plausibility argument; the proof comes with detailed balance (Section 9.9). Suppose many walkers are already spread out as f, the way the crowd of politicians was spread over the islands in proportion to the populations (Section 9.4). Take two points x and y with $f(y) < f(x)$. The number of walkers who move from x to y in one step is (number at x) $\times$ (chance of proposing y) $\times$ (chance of accepting). The number moving back is (number at y) $\times$ (chance of proposing x) $\times$ (chance of accepting). The proposal is symmetric, so the two proposal chances are equal. The walkers at x outnumber those at y by $f(x)/f(y)$. For the two flows to cancel, which is detailed balance, the acceptance from x to y must be smaller than the acceptance from y to x by exactly the factor $f(y)/f(x)$. Accepting every uphill move and downhill moves with probability $f(y)/f(x)$ does precisely that, and nothing else in the rule refers to f.

Key idea: The biased random walker

A random walk alone samples nothing: it diffuses forever. What turns it into a sampler is a bias, and the accept–reject step is that bias. Uphill proposals are always taken, a downhill proposal is taken with probability equal to the ratio of the target densities, and a refused proposal leaves the walker where he is for one more tick. The walker then spends more time where the probability is higher, in exact proportion to it. Without the accept–reject

step we would simply be simulating the random walk [Wasserman p. 415].

Written as a recipe, this is the algorithm that Metropolis et al. [36] published for the statistical mechanics of interacting particles, and that Hastings [170] generalised to asymmetric proposals (Section 9.9.1).

Working definition 9.8 (random-walk Metropolis)

Given a function returning $\ln f(x)$ up to a constant, a starting point x_0 with $f(x_0) > 0$, and a step length s:

1. *Propose* $y = x_t + s\,z$ with $z \sim \mathcal{N}(0,1)$ [the drunkard's step, symmetric in x and y].
2. *Compare*: compute $R = f(y)/f(x_t)$ [the only place the target enters].
3. *Decide*: draw $U \sim \text{Uniform}(0,1)$. If $U < R$ set $x_{t+1} = y$ [this happens with probability $\min(1, R)$]. Otherwise set $x_{t+1} = x_t$ [the old point is recorded again, it is not skipped].

The list $x_0, x_1, x_2, \ldots$, after the first stretch is discarded, is a correlated sample from f [Wasserman pp. 411–412], [Kruschke §7.3], [Padilla eq. 56]. *Operational test:* the fraction of accepted proposals is between about 0.2 and 0.5, and long runs from different starts give the same histograms (Sections 9.11 and 9.13).

9.8.1 Only ratios enter: the denominator of Bayes cancels

The rule needs $f(y)/f(x)$ and nothing else. To see why that is such good news, run once more through the reasoning of Chapter 5 that produces a posterior, in its usual order. *What did we observe?* The whole data vector $\boldsymbol{d}$, all of it at once, is the one observed event. *What could have produced it?* Each value of the parameter vector $\boldsymbol{\theta}$ is a candidate explanation, one hypothesis about how nature was set up. *How readily would each have produced it?* For every candidate the likelihood $p(\boldsymbol{d} \mid \boldsymbol{\theta})$ answers: had nature been set up this way, how plausible would it have been to get exactly this $\boldsymbol{d}$? *How do we update?* Bayes' theorem multiplies that likelihood by the prior plausibility of the candidate and divides by the evidence $Z = p(\boldsymbol{d})$, the sum of likelihood times prior over every candidate. That denominator is the one quantity that needs every hypothesis at once.

The walker never asks for a posterior on its own. At each tick it holds two candidate explanations side by side, the present point and the proposal, and asks only how much more weight one has than the other. A ratio of two weights that share the same denominator does not contain the denominator. Write the target as the posterior of a parameter vector $\boldsymbol{\theta}$ given data $\boldsymbol{d}$, with likelihood $\mathcal{L}(\boldsymbol{\theta}) = p(\boldsymbol{d} \mid \boldsymbol{\theta})$ and prior $p(\boldsymbol{\theta})$. For a current point $\boldsymbol{\theta}$ and a proposal $\boldsymbol{\theta}'$,

$$R \overset{(1)}{=} \frac{p(\boldsymbol{\theta}' \mid \boldsymbol{d})}{p(\boldsymbol{\theta} \mid \boldsymbol{d})} \overset{(2)}{=} \frac{\mathcal{L}(\boldsymbol{\theta}')\,p(\boldsymbol{\theta}')/Z}{\mathcal{L}(\boldsymbol{\theta})\,p(\boldsymbol{\theta})/Z} \overset{(3)}{=} \frac{\mathcal{L}(\boldsymbol{\theta}')\,p(\boldsymbol{\theta}')}{\mathcal{L}(\boldsymbol{\theta})\,p(\boldsymbol{\theta})},$$

$$\ln R \overset{(4)}{=} \big[\ln \mathcal{L}(\boldsymbol{\theta}') + \ln p(\boldsymbol{\theta}')\big] - \big[\ln \mathcal{L}(\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta})\big]. \tag{9.58}$$

(1) The target density f is the posterior.
(2) Bayes' theorem (Chapter 5) at each of the two points, with the same $Z = \int \mathcal{L}(\boldsymbol{\theta})\,p(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$ in both.
(3) Z does not depend on the point at which the posterior is evaluated, so it cancels.
(4) Logarithm of a ratio is a difference of logarithms; the decision $U < R$ becomes $\ln U < \ln R$.

The walker needs "likelihood times prior" at two points, both of which we can always compute, and never the integral over all points, which we usually cannot: the evidence is the same number for both candidates, so it drops out. The same cancellation will appear in statistical

mechanics, where the target is the Boltzmann weight $e^{-E/kT}/Z$ and the normalisation Z is the partition function (Section 9.15).

Example: the coin. Kruschke's coin [Kruschke §7.3.1] landed heads $z = 14$ times in $N = 20$ flips. With a flat prior on the bias $\theta \in (0, 1)$ the posterior is $p(\theta \mid \boldsymbol{d}) = \theta^z(1-\theta)^{N-z}/B(z+1, N-z+1)$, where B is the beta function. From $\theta = 0.6$ a proposal $\theta' = 0.75$ has

$$\ln R \overset{(1)}{=} 14 \ln \frac{0.75}{0.6} + 6 \ln \frac{0.25}{0.4} \overset{(2)}{=} 14 \times 0.2231 + 6 \times (-0.4700) = 3.124 - 2.820 = 0.304,$$

(1) The beta function, the evidence of this problem, cancels; the flat prior is 1 at both points.
(2) $\ln 1.25 = 0.2231$ and $\ln 0.625 = -0.4700$.

so $R = e^{0.304} = 1.36 > 1$ and the move is taken. The reverse proposal, from 0.75 to 0.6, has $R = 1/1.36 = 0.74$ and is taken 74 times in a hundred. The walker therefore sits near 0.75 a little more often than near 0.6, by the factor 1.36, which is the ratio of the posterior densities at the two points.

Example: why we never compute Z instead. The evidence is a D-dimensional integral. A grid with n points per axis costs n^D evaluations, 10^{12} for $n = 100$ and $D = 6$ (Chapter 5). Plain acceptance–rejection sampling (Chapter 6) does no better in many dimensions. To sample uniformly inside the unit ball in D dimensions by drawing points in the enclosing cube of side 2 and keeping those inside, the acceptance is the volume ratio $V_D/2^D = (\sqrt{\pi}/2)^D/\Gamma(1 + D/2)$: $\pi/4 \simeq 0.79$ for $D = 2$, about 0.0025 for $D = 10$, and about 10^{-70} for $D = 100$ [HobsonBMC §3.2.2, p. 61]. Most of a high-dimensional box is corners. A walker that moves locally and only compares neighbouring points never pays this price, because it never looks at the empty corners at all.[1]

Example: why logarithms. A likelihood built from many data points is a product of many factors and underflows. For a Gaussian likelihood of 10^4 data points with $\chi^2 \simeq 10^4$, $\mathcal{L} \propto e^{-\chi^2/2} = e^{-5000}$, far below the smallest positive double-precision number, $e^{-745} \simeq 5 \times 10^{-324}$. Every point would evaluate to zero and every ratio to $0/0$. The difference $\ln \mathcal{L}(\boldsymbol{\theta}') - \ln \mathcal{L}(\boldsymbol{\theta}) = -(\chi'^2 - \chi^2)/2$ is an ordinary number of order one. Every sampler below therefore works with $\ln f$ and accepts when $\ln U < \ln f(y) - \ln f(x)$, which is the advice of Chapter 5 to sum log-likelihoods and never multiply raw ones.

9.9 Why the rule is $\min(1, R)$: detailed balance

The walker of Section 9.8 took every uphill proposal and a downhill one with probability $f(y)/f(x)$, and on the two-bump density (9.57) its histogram came out right. We justified the rule with a crowd of walkers whose flows between two points cancel. The cancellation relied on a symmetric proposal. A boundary, such as the requirement that a rate be positive, can make an asymmetric proposal more convenient. We therefore need to see how the acceptance rule changes when the forward and reverse proposals have different probabilities, and how efficiently the resulting walker uses its proposals.

The question. Fix a way of proposing moves: from x, a candidate y is drawn from a density $q(y \mid x)$. With what probability $a(x, y)$ should we accept it so that the target f is the stationary distribution of the resulting chain?

[1] The source quotes ~ 0.025 for $n = 10$; the formula it gives evaluates to $(\sqrt{\pi}/2)^{10}/\Gamma(6) = 0.0025$.

The one-step law of the chain. Two things can happen in one step from x. Either a candidate y is proposed *and* accepted, which has density $q(y \mid x)\, a(x, y)$ in y; or the candidate is refused and the chain stays at x. The second possibility has total probability

$$r(x) = 1 - \int q(y \mid x)\, a(x, y)\, \mathrm{d}y, \tag{9.59}$$

because the proposal certainly produces some y and the step either accepts or refuses it. The transition kernel of Section 9.4.3 therefore has a smooth part and a spike at the starting point:

Working definition 9.9 (Metropolis–Hastings kernel)

The probability that one step from x lands in a small set $\mathrm{d}y$ is

$$P(x, \mathrm{d}y) = q(y \mid x)\, a(x, y)\, \mathrm{d}y \;+\; r(x)\, \delta_x(\mathrm{d}y), \tag{9.60}$$

where δ_x [the point mass at x, probability one on the point itself] carries the refusals, and the acceptance probability is

$$a(x, y) = \min\bigl(1, R_H(x, y)\bigr), \qquad R_H(x, y) = \frac{f(y)\, q(x \mid y)}{f(x)\, q(y \mid x)}. \tag{9.61}$$

R_H is the **Hastings ratio** [target ratio times the ratio of the reverse to the forward proposal density]; for a symmetric proposal, $q(x \mid y) = q(y \mid x)$, it reduces to $f(y)/f(x)$ [Wasserman p. 412], [HobsonBMC §3.2.4, p. 63], [Padilla eqs. 55–56].

Deriving the acceptance from detailed balance. Recall from Section 9.4.2 that a chain is in *detailed balance* with f when the probability flow from x to y equals the flow from y to x, for every pair. For $y \neq x$ only the smooth part of (9.60) moves probability, so we ask

$$\begin{aligned} & f(x)\, q(y \mid x)\, a(x, y) \overset{(1)}{=} f(y)\, q(x \mid y)\, a(y, x) \\ \Longleftrightarrow \quad & \frac{a(x, y)}{a(y, x)} \overset{(2)}{=} \frac{f(y)\, q(x \mid y)}{f(x)\, q(y \mid x)} \overset{(3)}{=} R_H(x, y). \end{aligned} \tag{9.62}$$

(1) Flow from x to y: density of being at x, times density of proposing y, times the chance of accepting; and the same for the reverse flow.
(2) Divide both sides by $f(x)\, q(y \mid x)\, a(y, x)$, assuming these are nonzero.
(3) Definition (9.61). Note that $R_H(y, x) = 1/R_H(x, y)$.

Detailed balance fixes only the *ratio* of the two acceptances, not each one. So let the acceptance be some function of the Hastings ratio, $a(x, y) = g\bigl(R_H(x, y)\bigr)$, with values in $[0, 1]$. Because $R_H(y, x) = 1/R_H(x, y)$, condition (9.62) becomes a condition on g alone:

$$g(R) = R\, g(1/R) \qquad \text{for every } R > 0. \tag{9.63}$$

Many functions satisfy it. Two classic ones are the Metropolis choice and Barker's choice:

$$\begin{aligned} g_{\mathrm{M}}(R) = \min(1, R): \quad & \text{for } R \geq 1, \;\; R\, g_{\mathrm{M}}(1/R) = R \cdot \tfrac{1}{R} = 1 = g_{\mathrm{M}}(R), \\ & \text{for } R < 1, \;\; R\, g_{\mathrm{M}}(1/R) = R \cdot 1 = R = g_{\mathrm{M}}(R); \\ g_{\mathrm{B}}(R) = \frac{R}{1 + R}: \quad & R\, g_{\mathrm{B}}(1/R) = R\, \frac{1/R}{1 + 1/R} = \frac{R}{R + 1} = g_{\mathrm{B}}(R). \end{aligned}$$

The Metropolis choice is the largest. Any admissible g satisfies two bounds: $g(R) \le 1$, because it is a probability, and $g(R) = R\,g(1/R) \le R \cdot 1 = R$, by (9.63) and the first bound applied to $g(1/R)$. Hence

$$g(R) \le \min(1, R) = g_{\mathrm{M}}(R) \qquad \text{for every admissible } g. \tag{9.64}$$

Metropolis accepts every proposal with the largest probability that detailed balance allows (Figure 9.21). Fewer refusals mean fewer wasted steps and less stickiness, and Peskun [172] showed that, for a given proposal, this choice also gives the smallest asymptotic variance of chain averages among such rules.

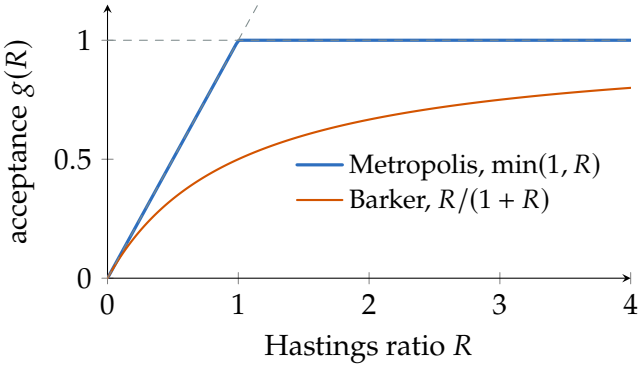


Figure 9.21: Two acceptance rules that both satisfy detailed balance. Every admissible rule lies below the line $g = R$ and below $g = 1$ (both dashed); the Metropolis rule runs along these two bounds, so it refuses as few proposals as possible.

From detailed balance to stationarity, with the refusals included. Detailed balance (9.62) concerns $y \neq x$. The refusals in (9.60) keep probability at x, and they must be counted when we check that f is stationary, that is, that a crowd of walkers spread as f is still spread as f after one step. The density of walkers at y after the step is the flow into y from everywhere else plus the walkers already at y who stayed:

$$\begin{aligned}\int f(x)\,q(y\,|\,x)\,a(x,y)\,\mathrm{d}x + f(y)\,r(y) &\overset{(1)}{=} \int f(y)\,q(x\,|\,y)\,a(y,x)\,\mathrm{d}x + f(y)\,r(y) \\ &\overset{(2)}{=} f(y)\Big[\int q(x\,|\,y)\,a(y,x)\,\mathrm{d}x + r(y)\Big] \overset{(3)}{=} f(y).\end{aligned} \tag{9.65}$$

(1) Detailed balance (9.62) for each pair (x, y) inside the integral.
(2) $f(y)$ does not depend on the integration variable x.
(3) By (9.59) at the point y, the probability of leaving y plus the probability of staying is one.

Wasserman checks (9.62) case by case [Wasserman pp. 413–414, (24.7)–(24.9)]: if $f(x)q(y\,|\,x) > f(y)q(x\,|\,y)$, then $a(x,y) = R_H < 1$ and $a(y,x) = 1$, so $f(x)\,q(y\,|\,x)\,a(x,y) = f(y)\,q(x\,|\,y) = f(y)\,q(x\,|\,y)\,a(y,x)$; the other case is the same with x and y exchanged.

Stationarity is half of what a sampler needs; the other half is that the chain actually gets there from any start. The results of Sections 9.5 and 9.6 supply it once the chain is irreducible and aperiodic. A Gaussian proposal has $q(y\,|\,x) > 0$ for every y, so any region where $f > 0$ can be reached in a single step: the chain is irreducible. And whenever some proposals are refused, the chain can stay put, which removes any cycle: it is aperiodic. Then the distribution of X_t converges to f from any starting point, and chain averages converge to averages over f (the ergodic theorem, Section 9.6).[2]

[2]For a continuous state space the full statement needs a little more measure theory (Harris recurrence); for the targets of this book, with f continuous and positive on a connected region and Gaussian proposals, it holds.

Example: two states, Metropolis against Barker. Take two states with target $f = (\frac{1}{3}, \frac{2}{3})$ and a proposal that always suggests the other state, so q is symmetric and the Hastings ratio is just the target ratio. A walker in state 1 is offered state 2, with $R = f_2/f_1 = 2$; a walker in state 2 is offered state 1, with $R = f_1/f_2 = \frac{1}{2}$. Each rule turns these two ratios into acceptance probabilities:

$$\begin{aligned} &\text{from state 1 } (R=2): \quad a_{\mathrm{M}} \overset{(1)}{=} \min(1,2) = 1, && a_{\mathrm{B}} \overset{(2)}{=} \frac{2}{1+2} = \frac{2}{3}, \\ &\text{from state 2 } (R=\tfrac{1}{2}): \quad a_{\mathrm{M}} \overset{(1)}{=} \min\bigl(1,\tfrac{1}{2}\bigr) = \tfrac{1}{2}, && a_{\mathrm{B}} \overset{(2)}{=} \frac{1/2}{1+1/2} = \frac{1}{3}. \end{aligned}$$

(1) The Metropolis rule $g_{\mathrm{M}}(R) = \min(1, R)$.
(2) Barker's rule $g_{\mathrm{B}}(R) = R/(1+R)$.

An accepted proposal switches state; a refused one keeps it, so each row of the transition matrix is (refuse, accept) or (accept, refuse). With rows labelled by where the walker is and columns by where it goes,

$$P_{\mathrm{M}} = \begin{pmatrix} 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}, \qquad P_{\mathrm{B}} = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix}.$$

Both leave f unchanged: $(\frac{1}{3}, \frac{2}{3})P_{\mathrm{M}} = (\frac{1}{3}, \frac{2}{3})$, and every row of P_{B} is already f.

Which is the better sampler? Both are correct, so the question is how many steps each needs for a given precision. The two-state chain with $p_{12} = a$ and $p_{21} = b$ has autocorrelation λ^k with $\lambda = 1 - a - b$ and integrated autocorrelation time $\tau = (1+\lambda)/(1-\lambda)$, from (9.45). Barker has $a = \frac{2}{3}$, $b = \frac{1}{3}$, $\lambda = 0$ and $\tau = 1$: its draws are independent. Metropolis has $a = 1$, $b = \frac{1}{2}$, $\lambda = -\frac{1}{2}$ and

$$\tau_{\mathrm{M}} = \frac{1 - \frac{1}{2}}{1 + \frac{1}{2}} = \frac{1}{3}.$$

What τ buys. The variance of a chain average is $\sigma^2\tau/N$ by (9.44), where σ^2 is the variance of one value. Independent draws have $\tau = 1$, so N_{ind} of them give $\sigma^2/N_{\mathrm{ind}}$. The two variances are equal when $N_{\mathrm{ind}} = N/\tau$, which for $\tau_{\mathrm{M}} = \frac{1}{3}$ is $3N$: N Metropolis steps estimate the fraction of time in state 1 as well as $3N$ independent draws would.

Why the chain beats independence. From state 1 the move to state 2 is uphill, and Metropolis never refuses an uphill move, so the walker leaves state 1 at every visit. The chain switches state as often as detailed balance allows, and a value in state 1 is followed by a value in state 2 more often than chance would have it. Successive values are anticorrelated ($\lambda < 0$), and their errors in the running average partly cancel instead of adding up. Barker refuses the uphill move one time in three and loses this. The general statement, proved by Peskun [172] for any target and any fixed proposal, is that a rule which refuses less never gives a larger asymptotic variance; the two-state matrices above are that theorem in its smallest case.

Example: the tempered family. What if the walker accepts with probability $\min\bigl(1, R^\beta\bigr)$ for some exponent $\beta \geq 0$? Since

$$R^\beta = \left(\frac{f(y)}{f(x)}\right)^\beta = \frac{f(y)^\beta}{f(x)^\beta},$$

this is exactly the Metropolis rule for the target f^β, so by (9.65) the walker samples $f^\beta / \int f^\beta$. The family interpolates between the two failures of Section 9.8: $\beta = 0$ is the drunkard ($f^0 = 1$ is flat and not normalisable on the line) and $\beta \to \infty$ is the hill climber (f^β piles up at the highest point). Figure 9.22 checks the claim on the two-bump density with 2×10^5 steps: the fraction of time at $x > 0$ is 0.5929 for $\beta = 0.3$ (exact, from $f^{0.3}$ normalised on a grid: 0.5915), 0.6937 for

$\beta = 1$ (0.6957) and 0.8742 for $\beta = 3$ (0.8772). If $f = e^{-E}$ then $f^\beta = e^{-\beta E}$, a Boltzmann weight at temperature $1/\beta$: lowering β heats the target and flattens its barriers. Simulated annealing (raising β slowly to find a global maximum) and parallel tempering (running several β at once and swapping) are built on this observation; the temperature returns as a physical temperature in Section 9.15.

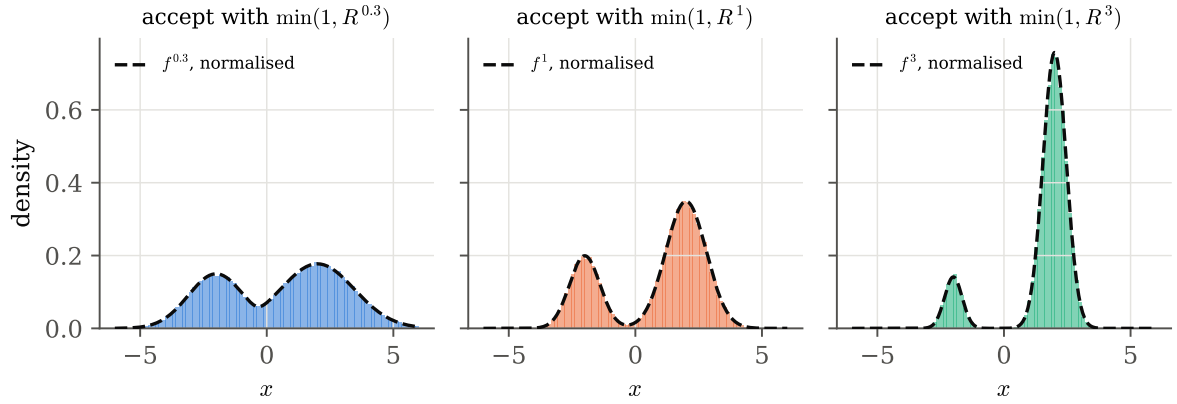


Figure 9.22: A walker that accepts with $\min(1, R^\beta)$ samples f^β, not f. Histograms of three such walkers on the two-bump density, with f^β normalised on a grid (dashed). Small β flattens the target towards the drunkard's flat line; large β sharpens it towards the hill climber's spike.

9.9.1 Asymmetric proposals: the Hastings factor

A Poisson process gave $n = 2$ events in unit exposure. With a flat prior on the rate $\lambda > 0$, the posterior is $f(\lambda) \propto \lambda^2 e^{-\lambda}$, a Gamma density with shape 3, mean 3, and probability 0.01439 below $\lambda = 0.5$. Rates, variances, masses and amplitudes all live on a half-line, and near the wall at $\lambda = 0$ a Gaussian proposal of width $s = 2$ often suggests a negative rate.

The question. What should the walker do with a proposal on the wrong side of the wall, and what changes if we choose proposals that never cross it?

We try five samplers, each run for 200 000 steps (Figure 9.23), and ask of each whether it satisfies (9.62).

A: Gaussian step, stay put if $y < 0$. The target is $f(y) = 0$ there, so $R = 0$ and the proposal is refused like any other bad proposal. Nothing about the rule has changed, so it is exact: mean 3.008, and a fraction 0.01464 of the time below 0.5.

B: Gaussian step, redraw until $y > 0$, no correction. This feels harmless but changes the proposal. *The redrawn proposal density.* An untruncated step $y = x + sZ$ has density $\phi\big((y - x)/s\big)/s$, where ϕ and Φ are the standard Gaussian density and cumulative distribution. The step lands on the allowed side with probability $\mathrm{P}(x + sZ > 0) = \mathrm{P}(Z > -x/s) = \Phi(x/s)$, the last step because Z is symmetric about zero. Redrawing until $y > 0$ keeps only those draws, so it renormalises the Gaussian over $y > 0$ by dividing by that probability:

$$q(y \mid x) = \frac{\phi\big((y - x)/s\big)}{s\,\Phi(x/s)}, \qquad y > 0. \tag{9.66}$$

Close to the wall, $\Phi(x/s)$ is near $\frac{1}{2}$ (half of all steps are thrown back); far from it, $\Phi(x/s)$ is near 1. So the denominator differs between the two directions of a move.

Which flow is unbalanced? Take a point x near the wall and a point y farther out. Sampler B accepts with $\min\big(1, f(y)/f(x)\big)$, as if the proposal were symmetric. The two flows of (9.62) are then

$$\frac{\text{flow } x \to y}{\text{flow } y \to x} \overset{(1)}{=} \frac{f(x)\,q(y\,|\,x)\min\big(1, f(y)/f(x)\big)}{f(y)\,q(x\,|\,y)\min\big(1, f(x)/f(y)\big)} \overset{(2)}{=} \frac{q(y\,|\,x)\ \min\big(f(x), f(y)\big)}{q(x\,|\,y)\ \min\big(f(y), f(x)\big)} \overset{(3)}{=} \frac{\Phi(y/s)}{\Phi(x/s)}.$$

(1) Density at the start, times the proposal density, times the acceptance, for each direction.
(2) Take $f(x)$ and $f(y)$ inside the minimum: $f(x)\min(1, f(y)/f(x)) = \min(f(x), f(y))$.
(3) The two minima are the same number. In the ratio of the two densities (9.66) the ϕ factors are equal, because ϕ is even, and only the denominators are left.

For x nearer the wall than y, $\Phi(y/s) > \Phi(x/s)$ and the ratio exceeds one. With $s = 2$, $x = 0.2$ and $y = 2$ it is $\Phi(1)/\Phi(0.1) = 0.841/0.540 \simeq 1.56$: probability leaves the neighbourhood of the wall about one and a half times as fast as it comes back. Detailed balance fails in one direction for every such pair, so the walker settles into a stationary law that is depleted near zero. That is what the run measures: it spent only 0.009384 of its time below 0.5, and its mean came out at 3.183.

C: as B, with the Hastings factor. From (9.66), the ϕ factors are equal because ϕ is even, and

$$\frac{q(x\,|\,y)}{q(y\,|\,x)} = \frac{\phi\big((x-y)/s\big)/[s\,\Phi(y/s)]}{\phi\big((y-x)/s\big)/[s\,\Phi(x/s)]} = \frac{\Phi(x/s)}{\Phi(y/s)}.$$

Multiplying the target ratio by this factor, as (9.61) instructs, restores balance: mean 2.975, fraction below 0.5 0.01402.

D: multiplicative steps with the Hastings factor. Propose $y = x\,e^{\sigma Z}$ with $\sigma = 1$ and $Z \sim \mathcal{N}(0,1)$: a Gaussian step in $\ln\lambda$, which can never cross zero. This is Lambert's log-normal proposal [Lambert §13.8]. Its density is

$$\begin{aligned} q(y\,|\,x) &\overset{(1)}{=} \frac{1}{y\,\sigma\sqrt{2\pi}}\exp\Big[-\frac{(\ln y - \ln x)^2}{2\sigma^2}\Big], \\ \frac{q(x\,|\,y)}{q(y\,|\,x)} &\overset{(2)}{=} \frac{1/x}{1/y} = \frac{y}{x}. \end{aligned} \tag{9.67}$$

(1) $\ln y = \ln x + \sigma Z$ is Gaussian; the density of y picks up the Jacobian $|\mathrm{d}\ln y/\mathrm{d}y| = 1/y$ (change of variables, Chapter 2).
(2) The exponential is symmetric in x and y and cancels; only the Jacobian factors remain.

With the factor y/x the sampler is exact: mean 3.012, fraction below 0.5 0.01366.

E: as D, without the factor. Here the walker uses $\min\big(1, f(y)/f(x)\big)$ although the proposal is not symmetric in λ. What does it sample? In the variable $u = \ln\lambda$ the proposal $u \to u + \sigma Z$ *is* symmetric, and the decision uses $f(e^{u'})/f(e^{u})$. So, as a chain in u, it is an honest Metropolis sampler of the density $h(u) \propto f(e^u)$. Back in λ,

$$p_E(\lambda) \overset{(1)}{=} h(\ln\lambda)\left|\frac{\mathrm{d}u}{\mathrm{d}\lambda}\right| \overset{(2)}{=} \frac{f(\lambda)}{\lambda} \overset{(3)}{=} \lambda\,e^{-\lambda}. \tag{9.68}$$

(1) Change of variables from u to $\lambda = e^u$.
(2) $h(\ln\lambda) \propto f(\lambda)$ and $\mathrm{d}u/\mathrm{d}\lambda = 1/\lambda$.
(3) $f(\lambda) \propto \lambda^2 e^{-\lambda}$.

The forgotten factor costs one power of λ: the walker samples a Gamma density with shape 2 and mean 2. It returned mean 1.989 and fraction 0.09093 below 0.5, against $1 - 1.5\,e^{-0.5} = 0.090$

for that wrong target. The Hastings factor is the Jacobian of the proposal's own geometry; leaving it out is the same mistake as forgetting a Jacobian when changing variables. This is also Wasserman's advice for bounded parameters, "transform to $\ln X$" [Wasserman p. 415]: sampling $u = \ln\lambda$ with symmetric steps is fine provided the target in u includes the Jacobian, $h(u) = f(e^u)\,e^u$, which is exactly the factor y/x of D.

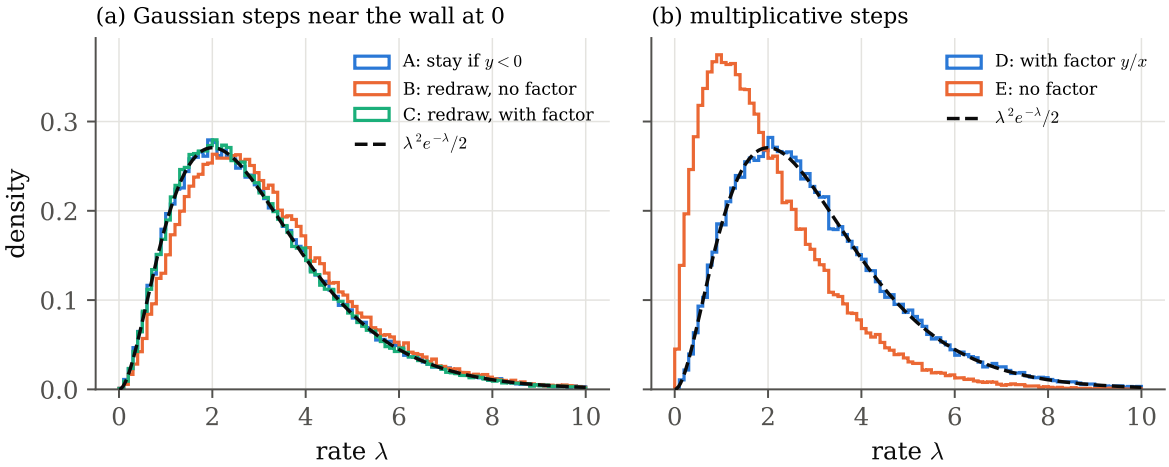


Figure 9.23: Five ways to sample a positive rate with posterior $\lambda^2e^{-\lambda}/2$ (dashed). (a) Gaussian steps: staying put when the step crosses zero (A) and redrawing with the Hastings factor (C) are exact; redrawing without it (B) pushes the walker away from the wall. (b) Multiplicative steps: with the factor y/x (D) exact; without it (E) the walker samples $\lambda e^{-\lambda}$, with one power of λ missing.

Sampler A deserves a remark because one textbook warns against it. Lambert writes that if all steps proposing a negative value are rejected, too few samples are generated near zero [Lambert §13.8].[3] What goes wrong is redrawing, not refusing.

Watch out: A refusal repeats the current point

In acceptance–rejection sampling (Chapter 6, [Cowan §3.3]) a rejected candidate is thrown away and a fresh independent one is drawn. In Metropolis–Hastings a refused proposal means the walker stays, and the current point is written into the chain *again*. The repeats are not a nuisance: they are how a high-density point earns its extra weight, by being hard to leave. Recording only the accepted moves, or redrawing until something is accepted, removes that weight and biases every average towards low-density regions [Padilla §4.1.1, step 4].

Example: the independence sampler. The proposal need not depend on x at all. Draw every candidate from a fixed density g, so $q(y \mid x) = g(y)$. Then (9.61) gives

$$R_H(x,y) = \frac{f(y)\,g(x)}{f(x)\,g(y)} = \frac{w(y)}{w(x)}, \qquad w \equiv \frac{f}{g}, \tag{9.69}$$

the ratio of the importance weights of Chapter 6 at the candidate and at the current point [Wasserman p. 415]. A candidate with a larger weight than the current point is always accepted; one with a smaller weight sometimes. When $g = f$ every weight is equal and every proposal is accepted: the chain becomes independent sampling. When g has thinner tails than f, a walker

[3] Lambert's statement and Fig. 13.15 describe what happens when the negative proposals are *discarded and redrawn*, which is our sampler B. If the walker stays where it is and records the current point again, which is how every Metropolis step treats a refusal, the region near zero is sampled correctly; sampler A agrees with the exact answer.

that reaches the tail finds $w(x)$ huge and refuses almost everything for a very long time; for exactly the reason given in Chapter 6 for importance sampling, g must have heavier tails than f.

9.10 The sampler in a dozen lines, and two textbook runs

The biased walker of Section 9.8 needs three operations per step: a Gaussian kick, one evaluation of $\ln f$, and one uniform number. That fits in a short function, and the same function will drive every chain in the rest of the book: the straight-line fit and the CMB likelihood that follow, the comparison with the Fisher matrix in Chapter 10, and the topological summaries of Chapter 12. Before trusting it with any of those, we run it on two published examples whose numbers we did not choose, Wasserman's Cauchy chains and Kruschke's coin.

Listing 9.5: Random-walk Metropolis, the core of `lib_mcmc.py` (`code/ch09/lib_mcmc.py`)

```
def metropolis(logp, x0, n, step, rng, cov=None):
    """Random-walk Metropolis with Gaussian proposal y = x + step * L z,  z ~ N(0, 1).

    L is the Cholesky factor of ``cov`` (identity if None).  The proposal is symmetric,
    q(y|x) = q(x|y), so the acceptance probability is min(1, p(y)/p(x)).
    Returns (chain, logp_chain, acceptance_fraction).
    """
    x = np.atleast_1d(np.asarray(x0, dtype=float))
    D = x.size
    L = np.eye(D) if cov is None else np.linalg.cholesky(np.atleast_2d(cov))
    chain = np.empty((n + 1, D))
    lp_chain = np.empty(n + 1)
    lp = logp(x)
    chain[0], lp_chain[0] = x, lp
    kicks = step * rng.standard_normal((n, D)) @ L.T     # all proposal increments at once
    log_u = np.log(rng.random(n))                          # all uniform draws at once
    accepted = 0
    for t in range(n):
        y = x + kicks[t]                                   # 1. propose: the unbiased walker
        lp_y = logp(y)
        if log_u[t] < lp_y - lp:                           # 2. the bias: accept with
    min(1, p(y)/p(x))
            x, lp = y, lp_y
            accepted += 1
        chain[t + 1], lp_chain[t + 1] = x, lp              # 3. a rejection repeats x
    return chain, lp_chain, accepted / n
```

The function follows the working definition of Section 9.8 line by line, with four practical choices.

- The target enters only through `logp`, the log density up to a constant. A point outside the support returns $-\infty$; then $\ln U < -\infty$ is false and the proposal is refused, which is sampler A of Section 9.9.1.
- The proposal is $y = x + s\,Lz$ with z a vector of independent standard Gaussians and L the Cholesky factor of a matrix $\mathbf{C}$, so $LL^{\mathsf{T}} = \mathbf{C}$. Its covariance is $\mathrm{Cov}(sLz) = s^2 L\,\mathrm{Cov}(z)\,L^{\mathsf{T}} = s^2 LL^{\mathsf{T}} = s^2\mathbf{C}$. With no $\mathbf{C}$ given, the steps are round; shaping them like the target is the subject of Section 9.11.
- All Gaussian kicks and all uniform numbers are drawn at once before the loop. Only the decision, which depends on where the walker is, has to be sequential.
- The value of $\ln f$ at the current point is stored, so each step costs exactly one evaluation of

the target, at the proposal. In a cosmological analysis that evaluation is a full likelihood call, and the number of steps is the cost of the analysis.

A refused step writes the current point again (the last line of the loop). The companion function `mh` in the same file takes any proposal together with its log density `log_q` and adds $\ln q(x \mid y) - \ln q(y \mid x)$ to the log ratio, which is the Hastings factor of (9.61); samplers C, D and E of Section 9.9.1 used it.

Watch out: Start where the target is positive

Follow what the function does if the starting point has $f(x_0) = 0$. Its stored log density is then $\ln f(x_0) = -\infty$. If a proposal y lands where $f(y) > 0$, the log ratio is the finite number $\ln f(y)$ minus $-\infty$, which is $+\infty$; the test $\ln U < +\infty$ is true, the move is taken, and from then on the walker is inside the support and everything is normal. But if the proposal also lands where $f = 0$, the log ratio is $(-\infty) - (-\infty)$, which floating-point arithmetic returns as `NaN` ("not a number"). Every comparison with `NaN` is false, so $\ln U <$ `NaN` is false and the proposal is refused, without any warning. A start outside the support with steps too short to reach it, say $\theta_0 = -1$ for a coin bias with steps of 0.02, therefore refuses every proposal: the trace is a flat line and the acceptance is exactly zero. The same happens inside the support if $\ln f$ underflows to $-\infty$ at the start and at all its neighbours. Always start where the posterior is positive [Kruschke §7.3]: inside the prior range, with a finite likelihood.

9.10.1 Wasserman's Cauchy chains

Wasserman's first MCMC example [Wasserman pp. 412–413, Ex. 24.10] makes one point. A Metropolis chain samples a density we cannot draw from directly, and it is correct for every proposal width; but how quickly its histogram comes to resemble the target depends on that width, and a badly chosen width wastes most of the run. He judges this by eye: a chain *mixes well* when its trace wanders over the whole range of the target without creeping or sticking. As target he takes the Cauchy density of Chapter 2,

$$f(x) = \frac{1}{\pi(1 + x^2)},$$

which is a good test for two reasons: every property of it is known exactly, so any run can be checked, and its tails are heavy, so a sampler has to reach far from the centre to get it right.

The ratio the walker uses. His proposal is a Gaussian step, $y \sim \mathcal{N}(x, b^2)$, with b the step width. It is symmetric, so the acceptance needs only the target ratio:

$$R = \frac{f(y)}{f(x)} \overset{(1)}{=} \frac{1/[\pi(1 + y^2)]}{1/[\pi(1 + x^2)]} \overset{(2)}{=} \frac{1 + x^2}{1 + y^2}.$$

(1) The Cauchy density at the proposal and at the present point.
(2) The π cancels, like every normalisation (Section 9.8.1); dividing by a fraction is multiplying by its inverse.

A proposal nearer the centre than x has $R > 1$ and is always taken; one farther out is taken with probability $(1 + x^2)/(1 + y^2)$, which falls like $1/y^2$ for a long jump.

The setup. Wasserman runs three chains that differ only in the width:

	small width	middle width	large width
proposal width b	0.1	1	10
steps N	1000	1000	1000
start	$x_0 = 0$	$x_0 = 0$	$x_0 = 0$

The start is at the centre of the target, so there is no burn-in to discard: everything that goes wrong is caused by the width. The width 0.1 is a tenth of the half-width of the Cauchy peak, and 10 is ten times it.

A number to check against. The trace plots of his Fig. 24.2 carry no numbers, so we need one that a run can be checked with. A convenient one is the fraction of time spent at $|x| < 1$, whose exact value is

$$\mathrm{P}(|X| < 1) \overset{(1)}{=} \int_{-1}^{1} \frac{\mathrm{d}x}{\pi(1+x^2)} \overset{(2)}{=} \frac{1}{\pi}\big[\arctan x\big]_{-1}^{1} = \frac{1}{\pi}\Big(\frac{\pi}{4} + \frac{\pi}{4}\Big) = \frac{1}{2}.$$

(1) Probability of an interval is the integral of the density over it.
(2) $\mathrm{d}(\arctan x)/\mathrm{d}x = 1/(1+x^2)$.

What we compute. The script `13_textbook_chains.py` calls the function `metropolis` listed above with the log target $-\ln(1+x^2)$ (the constant $-\ln\pi$ is dropped), the start [0], and each width in turn. For each width it makes two runs: one of 1000 steps, as Wasserman did, and one of 2×10^5 steps. From each it records the fraction of steps with $|x| < 1$, and from the long run also the acceptance rate and the integrated autocorrelation time τ of the indicator of $|x| < 1$ (Section 9.6.2, estimated as in Section 9.12.2). A chain of N steps with correlation time τ is worth N/τ independent draws.

Side by side. Figure 9.24 shows our three 1000-step chains, laid out as in his Fig. 24.2, and the table puts his verdicts next to our numbers.

proposal width b	0.1	1	10
Wasserman, Fig. 24.2 (by eye)	explores little	mixes well	often stuck
ours, 1000 steps: time at $\lvert x\rvert < 1$	0.8332	0.2478	0.5455
ours, 2×10^5 steps: time at $\lvert x\rvert < 1$	0.5517	0.5032	0.5002
ours, 2×10^5 steps: acceptance	0.9731	0.7723	0.2712
ours, 2×10^5 steps: τ (steps)	608	23.86	15.5

The traces have the features he describes. With $b = 0.1$ the walker takes tiny steps and is still exploring the centre after 1000 of them, so it spends far too much time at $|x| < 1$. With $b = 10$ most proposals land far out in the tail and are refused, and the trace is a staircase of long flat stretches. The $b = 1$ chain looks the liveliest. But in our run it happened to make a long excursion into a tail and spent only 0.2478 of its time at $|x| < 1$, further from the exact $\frac{1}{2}$ than the "stuck" $b = 10$ chain.

Why the verdicts differ. A 1000-step run cannot carry the weight of a ranking. The statistical error of a chain average is $\sigma\sqrt{\tau/N}$ (Section 9.12.2). For the fraction of time at $|x| < 1$, an average of an indicator that is 1 half the time, $\sigma = \frac{1}{2}$. With $N = 1000$ and the long-run τ values of the table, the error is $\frac{1}{2}\sqrt{23.86/1000} \simeq 0.08$ for $b = 1$ and $\frac{1}{2}\sqrt{15.5/1000} \simeq 0.06$ for $b = 10$: two widths whose τ differ by less than a factor of two cannot be told apart by one short run, and any single 1000-step histogram, his or ours, is a noisy copy of the target. Only $b = 0.1$, whose τ is some thirty times larger, is clearly worse in 1000 steps, and that part of his verdict stands.

The long runs are close to converged (each fraction is within two of its errors of $\frac{1}{2}$), and they rank the other two the other way round: $b = 10$ has the shorter τ, so it is at least as good as $b = 1$ although it refuses three proposals in four. The reason is the Cauchy tail: $\mathrm{P}(|X| > 10) = 1 - \frac{2}{\pi}\arctan 10 = 0.063$, so six per cent of the probability sits beyond $|x| = 10$, and only long jumps reach it in reasonable time. A walker with $b = 1$ has to diffuse out there one unit at a time and back again, which is the long excursion its short trace shows. A good step length depends on the shape of the target, not only on the width of its peak.

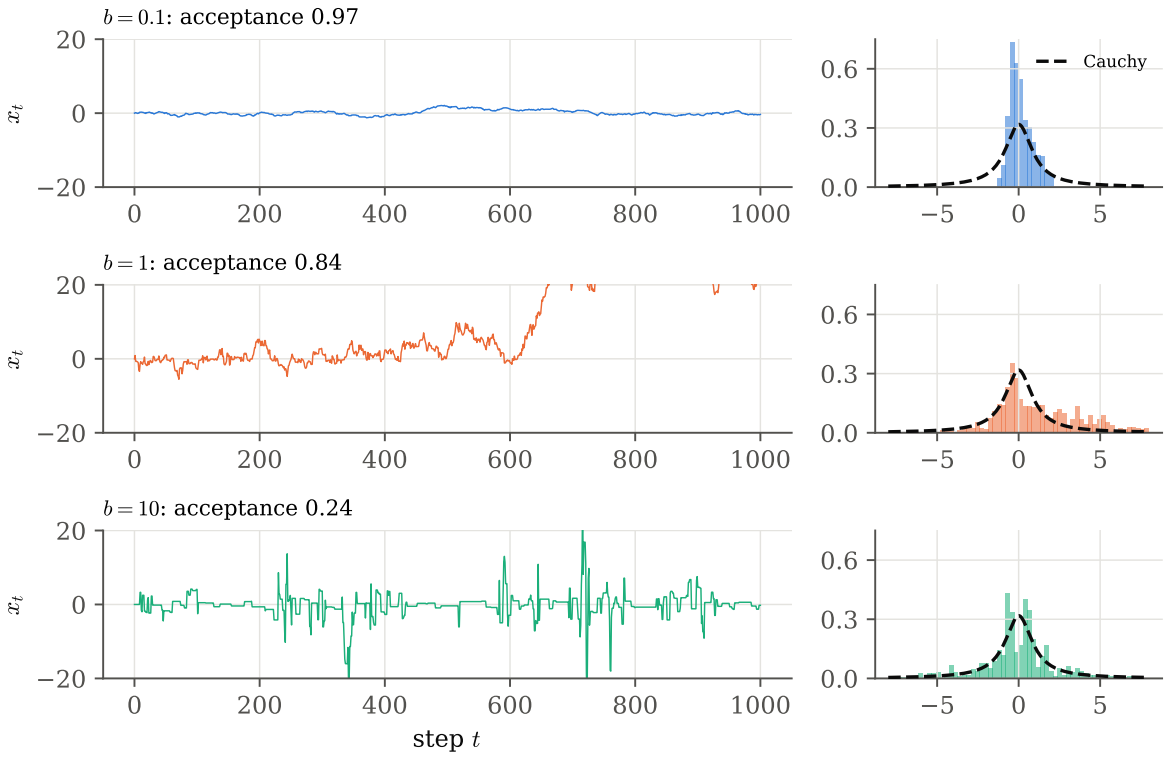


Figure 9.24: Wasserman's Cauchy example: three Metropolis chains of 1000 steps with proposal widths $b = 0.1$, 1 and 10 (left), and their histograms with the Cauchy density (right). With small proposals the chain explores the target slowly, while large proposals are often rejected and leave flat stretches in the trace. After 1000 steps all three histograms still differ visibly from the target density.

Note also what we did *not* estimate: the mean of a Cauchy variable does not exist (Chapter 2), and a chain average of x would wander forever however well the chain mixes. A sampler cannot create a moment the distribution does not have.

The lesson. A thousand steps judged by eye can tell a creeping chain from a working one, but they cannot rank two working samplers; that takes τ from a long run.

9.10.2 Kruschke's coin

Kruschke's first Metropolis run [Kruschke §7.3.1, Fig. 7.4] answers a question every user of the method has at the start. A coin was flipped $N = 20$ times and landed heads $z = 14$ times. We want the posterior for its bias θ, the probability of heads, and we do not want to integrate anything: can a walker that only ever compares two values of θ produce that posterior, and what does the choice of proposal width do to the result?

The exact answer, for checking. With the flat prior $\mathrm{Beta}(\theta; 1, 1)$ the posterior can be written down, which is why the problem makes a good test:

$$p(\theta \mid \boldsymbol{d}) \overset{(1)}{=} \frac{\theta^{14}(1-\theta)^{6} \cdot 1}{\int_0^1 t^{14}(1-t)^{6}\,\mathrm{d}t} \overset{(2)}{=} \mathrm{Beta}(\theta; 15, 7),$$

(1) Bernoulli likelihood $\theta^{z}(1-\theta)^{N-z}$ times a flat prior, normalised.

(2) The normalised density $\propto \theta^{a-1}(1-\theta)^{b-1}$ is the beta density with $a = 15$, $b = 7$ (Chapter 5); its mean is $a/(a+b) = 15/22 = 0.6818$ and its standard deviation is 0.097.

The walker does not use this formula. It receives only the log target $14\ln\theta + 6\ln(1-\theta)$ inside $(0,1)$ and $-\infty$ outside, and the integral in the denominator never appears (Section 9.8.1).

A deliberately bad start. Kruschke starts every chain at $\theta = 0.01$, far in the left tail: the posterior density there is smaller than at the mean by a factor

$$\frac{f(0.01)}{f(0.68)} = \exp\left[14\ln\frac{0.01}{0.68} + 6\ln\frac{0.99}{0.32}\right] \simeq e^{-52}.$$

A start there shows the walker finding the posterior before it explores it, the burn-in of Section 9.12.1, and how long that takes for each width.

Three widths. Each chain runs 50 000 steps, and only the width s of the Gaussian proposal changes. Compared with the posterior width of about 0.1:

- $s = 0.02$ is a fifth of the posterior width. Almost every step is a small move to a point of nearly the same density, so almost every step is accepted, but the walker covers ground only by diffusion and creeps.
- $s = 0.2$ is about twice the posterior width. A proposal lands in a region of quite different density about as often as not, so roughly half are taken, and each accepted move is a real jump across the posterior.
- $s = 2$ is twenty times the posterior width, and twice the whole allowed interval. Most proposals land outside $(0,1)$, where the target is zero, or far down a tail, and are refused; the walker sits still for long stretches.

For each chain Kruschke reports two numbers: the acceptance rate, and the **effective sample size** ESS $= N/\tau$, the number of independent draws that would give the posterior mean as accurately as the N correlated steps do, with τ the integrated autocorrelation time (defined in full, with its estimator, in Section 9.12.2).

Side by side. Our runs use the same start, widths and length, with our own random numbers:

proposal width	0.02	0.2	2
acceptance, ours	0.9334	0.4958	0.06174
acceptance, Kruschke	0.936	0.495	0.0624
effective size, ours	463	11181	2065
effective size, Kruschke	469	11724	2113

How close should they be? Two runs with different random numbers cannot agree exactly; the question is how far apart chance alone puts them. The acceptance rate is an average of N accept-or-refuse indicators, with a spread of order $\sqrt{a(1-a)/N} \simeq 0.002$ for $a = 0.5$ and $N = 50\,000$, a little more because successive decisions are correlated; the three pairs differ by a few thousandths. The estimated τ, and hence the ESS, has a relative spread of about $\sqrt{2(2M+1)/N}$ from one random-number stream to another, with $M \simeq 5\tau$ the window of the estimate (9.79). With $\tau = N/\text{ESS}$ from the table, that is about 4% for width 0.2 (an ESS spread of ±500), 10% for width 2 (±200) and 20% for width 0.02 (±100). The three ESS pairs differ by about 540, 50 and 6: all within what a second stream would give. Kruschke's ESS also comes from a different estimator (the coda package of R), which adds a few per cent of its own.

What each walker did. The 0.02 walker needs hundreds of steps to crawl from 0.01 to the bulk of the posterior (Figure 9.25, top left) and then explores it like a snake; the width-2 walker proposes mostly outside $(0,1)$ and refuses almost all of them. The middle walker, which accepts about half its proposals, is worth 11181 independent draws out of 50 000. Its posterior mean

is 0.6812, against 0.6818, and its central 95% interval is [0.4758, 0.8547] against the exact beta quantiles [0.4782, 0.8541]. (Kruschke's figure quotes the 95% highest-density interval, which for this left-skewed posterior lies slightly to the right of the central one; Chapter 5.) The same coin exercise, with five chains started between 0.1 and 0.9, opens the MCMC section of a cosmology tutorial [Padilla §4.1.2–4.1.3].

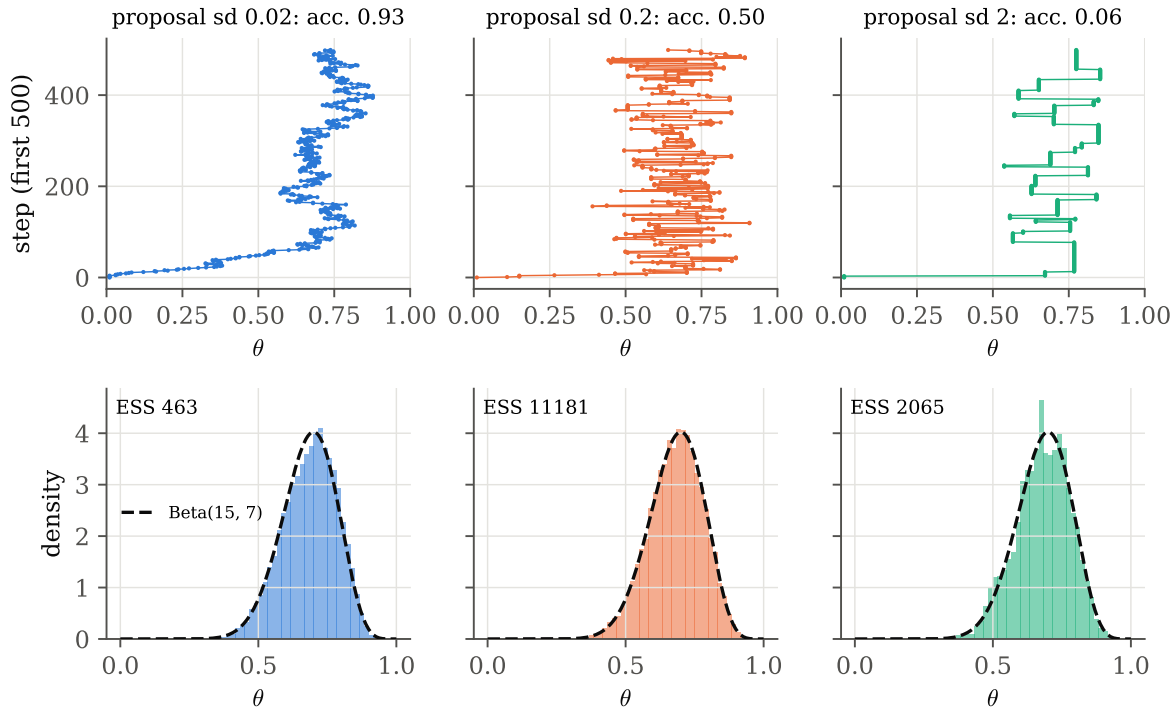


Figure 9.25: Kruschke's coin, reproduced. Top: the first 500 steps of three chains started at $\theta = 0.01$ with proposal widths 0.02, 0.2 and 2 (step number runs upwards). Bottom: histograms of the remaining steps against the exact Beta(15, 7) posterior, with the effective sample size of each chain.

All three coin histograms in Figure 9.25 are close to the beta density. Every proposal width gives the right answer eventually, because each defines a chain with the posterior as its stationary distribution. What differs is how long "eventually" is, more than an order of magnitude in effective sample size between the best and the worst width here.

The lesson. The proposal width never decides whether the answer is right, only what it costs; choosing it well is the subject of Section 9.11.

9.11 How big should the steps be?

Kruschke's three coin walkers of Section 9.10 all converged to the same beta posterior, and yet the best was worth more than ten times as many independent draws as the worst. The two bad ones failed in opposite ways. The walker with tiny steps accepted almost everything but moved like the drunkard, a distance $s\sqrt{t}$ in t steps, so crossing a posterior of width L took about $(L/s)^2$ steps, the diffusive cost met at the end of Section 9.7. The walker with huge steps proposed points far out in the tails, refused them, and sat still for long stretches. In between there must be a best step length, and since the target's width is usually unknown we would like to recognise it from something the chain itself reports.

The question. For a given target, which step length makes the walker produce independent information fastest, and is there a quantity we can monitor, such as the fraction of accepted proposals, that tells us we are there?

9.11.1 The Gaussian in many dimensions: 2.38 and 0.234

The one target where this question has a clean answer is the standard Gaussian in D dimensions, $f(x) \propto e^{-|x|^2/2}$, sampled with round Gaussian steps $y = x + (\ell/\sqrt{D})\, z$, $z \sim \mathcal{N}(0, \mathbb{1}_D)$, where ℓ is a dimensionless step scale. Why divide by $\sqrt{D}$? Because the log ratio is a sum over all D coordinates, and if each coordinate moved by a fixed amount the sum would grow with D and every proposal would be refused. Shrinking the step per coordinate as $1/\sqrt{D}$ keeps the log ratio of order one.

The log ratio becomes Gaussian. For a walker at x and a proposal $y = x + (\ell/\sqrt{D})z$,

$$\begin{aligned} \ln R &\overset{(1)}{=} -\tfrac{1}{2}\big(|y|^2 - |x|^2\big) \overset{(2)}{=} -\frac{\ell}{\sqrt{D}}\, x \cdot z - \frac{\ell^2}{2D}\, |z|^2 \\ &\overset{(3)}{=} -\ell\, \frac{|x|}{\sqrt{D}}\, \zeta - \frac{\ell^2}{2}\, \frac{|z|^2}{D} \xrightarrow{D\to\infty} \Lambda \sim \mathcal{N}\Big(-\frac{\ell^2}{2},\ \ell^2\Big). \end{aligned} \tag{9.70}$$

(1) The normalisation of f cancels; only the exponents differ.
(2) $|x+v|^2 = |x|^2 + 2x\cdot v + |v|^2$ with $v = (\ell/\sqrt{D})z$.
(3) For fixed x, $x \cdot z$ is a sum of independent Gaussians with variance $\sum_i x_i^2 = |x|^2$, so $x \cdot z = |x|\ \zeta$ with $\zeta \sim \mathcal{N}(0,1)$. A walker in equilibrium has $|x|^2/D \to 1$ and $|z|^2/D \to 1$ by the law of large numbers (both are averages of D squared standard Gaussians), so the first term tends to $-\ell\zeta$ and the second to $-\ell^2/2$.

The acceptance rate. Call the limiting log ratio of (9.70) Λ (a capital lambda, for "log ratio"). The acceptance probability of one proposal is $\min(1, R) = \min(1, e^{\Lambda})$, so the mean acceptance is $a(\ell) = \mathbb{E}\big[\min(1, e^{\Lambda})\big]$ with $\Lambda \sim \mathcal{N}(-\sigma^2/2, \sigma^2)$ and $\sigma = \ell$. Split it into the uphill part ($\Lambda > 0$, accepted for sure) and the downhill part:

$$\begin{aligned} a(\ell) &\overset{(1)}{=} \mathrm{P}(\Lambda > 0) + \mathbb{E}\big[e^{\Lambda};\, \Lambda < 0\big] \\ &\overset{(2)}{=} \Phi\Big(-\frac{\sigma}{2}\Big) + \int_{-\infty}^{0} \frac{1}{\sigma\sqrt{2\pi}}\, \exp\Big[w - \frac{(w+\sigma^2/2)^2}{2\sigma^2}\Big]\mathrm{d}w \\ &\overset{(3)}{=} \Phi\Big(-\frac{\sigma}{2}\Big) + \int_{-\infty}^{0} \frac{1}{\sigma\sqrt{2\pi}}\, \exp\Big[-\frac{(w-\sigma^2/2)^2}{2\sigma^2}\Big]\mathrm{d}w \overset{(4)}{=} 2\,\Phi\Big(-\frac{\ell}{2}\Big). \end{aligned} \tag{9.71}$$

(1) $\min(1, e^{\Lambda})$ is 1 when $\Lambda > 0$ and e^{Λ} when $\Lambda < 0$; $\mathbb{E}[X; A]$ means the expectation of X restricted to the event A.
(2) $\mathrm{P}(\Lambda > 0) = \mathrm{P}\big(\sigma\zeta - \sigma^2/2 > 0\big) = \mathrm{P}(\zeta > \sigma/2) = \Phi(-\sigma/2)$, with Φ the standard Gaussian cumulative distribution; the second term is the Gaussian density of Λ times e^{w}.
(3) Complete the square: $w - (w+\sigma^2/2)^2/2\sigma^2 = -(w^2 - \sigma^2 w + \sigma^4/4)/2\sigma^2 = -(w-\sigma^2/2)^2/2\sigma^2$. Multiplying by e^{w} has moved the Gaussian's mean from $-\sigma^2/2$ to $+\sigma^2/2$.
(4) The integral is the probability that a $\mathcal{N}(\sigma^2/2, \sigma^2)$ variable is negative, which is again $\Phi(-\sigma/2)$; and $\sigma = \ell$.

Small steps give $a \to 2\Phi(0) = 1$; large steps give $a \to 0$. Figure 9.26a shows that chains in $D = 10$ and 50 follow (9.71) closely, and that even $D = 1$ is not far from it.

The best step. *How far does one step move the walker?* When a proposal is accepted, the walker moves by $(\ell/\sqrt{D})z$; when it is refused, by nothing. Write A for the indicator of acceptance (1 if taken, 0 if refused). The mean squared jump per step, summed over coordinates, is

$$\mathbb{E}\,|x_{t+1}-x_t|^2 \overset{(1)}{=} \mathbb{E}\Big[\frac{\ell^2}{D}\,|z|^2\,A\Big] \overset{(2)}{=} \ell^2\,\mathbb{E}\Big[\frac{|z|^2}{D}\,A\Big] \overset{(3)}{\simeq} \ell^2\,\mathbb{E}[A] \overset{(4)}{=} \ell^2\,a(\ell) = 2\ell^2\,\Phi(-\ell/2). \tag{9.72}$$

(1) The squared jump is $\Big|(\ell/\sqrt{D})z\Big|^2 = (\ell^2/D)\,|z|^2$ if accepted and 0 if not.

(2) Take the constant ℓ^2 out, leaving ℓ^2/D times $|z|^2$ as ℓ^2 times the average $|z|^2/D$.

(3) $|z|^2/D$ is an average of D squared standard Gaussians, so for large D it is 1 with a spread of only $\sqrt{2/D}$; it is then a constant and comes out of the expectation. The acceptance A depends on z only through the one sum $x\cdot z$ of (9.70) and through this same average, so no single z_i matters and nothing correlates A with the size of the jump.

(4) $\mathbb{E}[A]$ is the mean acceptance (9.71).

From jump size to autocorrelation time. Follow one coordinate, x_1. Its share of the mean squared jump is $\ell^2 a(\ell)/D$ per step, and in equilibrium it spreads over a range of standard deviation 1. A coordinate that moves by diffusion forgets where it was once it has travelled about one standard deviation, and a diffusion with mean squared step v needs about $1/v$ steps for that, the cost $(L/s)^2$ of Section 9.7 with $L=1$ and $s^2=v$. So

$$\tau \;\propto\; \frac{D}{\ell^2\,a(\ell)},$$

and the step that maximises the mean squared jump minimises the autocorrelation time. The statement that this is exact as $D\to\infty$ is the theorem of Roberts, Gelman, and Gilks [173]: watched on a clock that ticks D steps at a time, each coordinate of the chain becomes a diffusion (a Langevin process) whose speed is exactly $h(\ell) = 2\ell^2\Phi(-\ell/2)$.

Maximising the speed. Set the derivative of h to zero:

$$0 = \frac{\mathrm{d}h}{\mathrm{d}\ell} \overset{(1)}{=} 4\ell\,\Phi(-\ell/2) - \ell^2\,\phi(\ell/2) \quad\Longleftrightarrow\quad 4\,\Phi(-\ell/2) = \ell\,\phi(\ell/2),$$

(1) Product rule, with $\mathrm{d}\Phi(-\ell/2)/\mathrm{d}\ell = -\frac{1}{2}\phi(\ell/2)$ and ϕ the standard Gaussian density.

whose root, found numerically, is $\ell = 2.381$. The acceptance there is

$$a = 2\,\Phi\Big(-\frac{2.381}{2}\Big) = 2\,\Phi(-1.19) = 2\times 0.117 = 0.2338,$$

so the best walker refuses $1-0.234 \simeq \frac{3}{4}$ of its proposals, three in four.

Why both extremes lose. For small ℓ the acceptance tends to 1 and $h\approx\ell^2$ is small. For large ℓ the acceptance collapses, and the rate follows from the Gaussian tail. For large u,

$$\Phi(-u) \overset{(1)}{=} \int_u^\infty \phi(t)\,\mathrm{d}t \overset{(2)}{=} \phi(u)\int_u^\infty e^{-(t^2-u^2)/2}\,\mathrm{d}t \overset{(3)}{=} \phi(u)\int_0^\infty e^{-u\,v-v^2/2}\,\mathrm{d}v \;\approx\; \frac{\phi(u)}{u} = \frac{e^{-u^2/2}}{u\sqrt{2\pi}}.$$

(1) $\Phi(-u) = \mathrm{P}(Z<-u) = \mathrm{P}(Z>u)$ for a standard Gaussian Z.

(2) $\phi(t) = \phi(u)\,e^{-(t^2-u^2)/2}$.

(3) Put $t = u+v$, so $t^2-u^2 = 2uv+v^2$. For large u the factor e^{-uv} cuts the integral off at $v\sim 1/u$, where $v^2/2$ is negligible, and $\int_0^\infty e^{-uv}\mathrm{d}v = 1/u$.

With $u=\ell/2$ this gives $a(\ell) = 2\Phi(-\ell/2) \approx (4/\ell\sqrt{2\pi})\,e^{-\ell^2/8}$: the acceptance dies like $e^{-\ell^2/8}$, faster than ℓ^2 grows, and h goes to zero. The compromise between the two sits at $\ell = 2.38$.

Key idea: The 2.38 and 0.234 rule

For a target that is roughly Gaussian in D dimensions, the most efficient round random-walk proposal has width $2.38/\sqrt{D}$ in units of the target's standard deviation, and accepts about 23% of its proposals [173]. In one dimension the best acceptance is higher, near one half. The chain then needs of order D steps per independent draw [HobsonBMC §3.2.4, p. 65].

Figure 9.26b measures the mean squared jump of real chains of 4×10^4 steps on a grid of 22 step scales. The best measured acceptances are 0.4583 in $D = 1$, 0.2816 in $D = 10$ and 0.2629 in $D = 50$, approaching 0.234 as D grows. At $\ell = 2.38$ the integrated autocorrelation time of one coordinate, divided by D, is 4.433, 3.353 and 2.98: an independent draw every three to four and a half times D steps. Dunkley et al. [171] found the same for cosmological chains, an inverse efficiency of about $3.3D$ and an optimal acceptance tending to 0.25, and recommended the width $2.4/\sqrt{D}$ that Lewis and Bridle [37] use in CosmoMC. Wasserman's rule of thumb, accept about half the proposals [Wasserman p. 415], is the one-dimensional answer; Kruschke's adaptive phase aims at "a middling value such as 0.5" [Kruschke p. 161]. In many dimensions the target is 0.2–0.3.

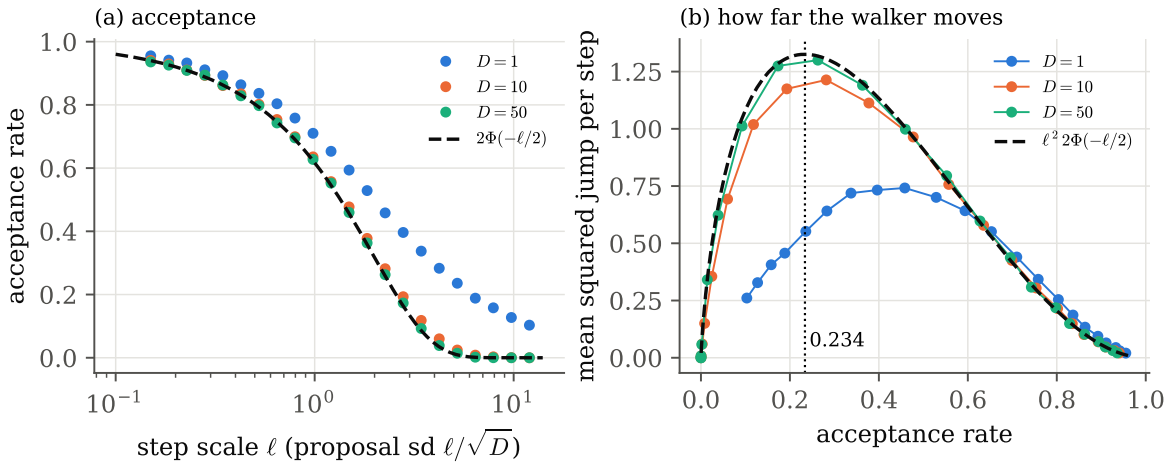


Figure 9.26: Step size on a standard Gaussian target in $D = 1$, 10 and 50 dimensions, with proposal width $\ell/\sqrt{D}$. (a) Measured acceptance against ℓ, with the large-D formula $2\Phi(-\ell/2)$ (dashed). (b) Mean squared jump per step against acceptance: it peaks near acceptance 0.23 in many dimensions (dotted line), and near 0.45 in one dimension.

9.11.2 Correlated targets: shape the steps like the posterior

Real posteriors are rarely round. Two parameters that the data constrain only in combination, such as the amplitude and tilt of a power spectrum, or the matter density and the Hubble rate, give a long thin ellipse. Take the Gaussian with unit variances and correlation $\rho = 0.95$, covariance $\mathbf{C} = \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)$. Its principal axes are the eigenvectors of $\mathbf{C}$:

$$\det(\mathbf{C} - \lambda \mathbb{1}) \overset{(1)}{=} (1-\lambda)^2 - \rho^2 = 0 \quad \Longleftrightarrow \quad \lambda_\pm \overset{(2)}{=} 1 \pm \rho,$$

(1) Determinant of a 2×2 matrix.
(2) $1 - \lambda = \mp\rho$; the eigenvectors are $(1, 1)/\sqrt{2}$ for $1 + \rho$ and $(1, -1)/\sqrt{2}$ for $1 - \rho$.

The standard deviation along the ridge is $\sqrt{1.95} = 1.40$ and across it $\sqrt{0.05} = 0.22$, a ratio of about six.

What a round step costs here. A round step cannot be tuned to both widths at once. Suppose we tune it to the narrow one, since that is the direction in which proposals are refused. For a round Gaussian in $D = 2$ the best step is $\ell/\sqrt{D} = 2.38/\sqrt{2} \simeq 1.7$ standard deviations (Section 9.11.1), so here it is about $1.7 \times 0.22 \simeq 0.37$. Along the ridge the walker then diffuses with steps of that size, and by the diffusion cost $(L/s)^2$ of Section 9.7 crossing a distance L takes about $(L/s)^2$ steps:

$$\begin{aligned} \text{ridge, step tuned to the narrow width:} \quad & \left(\frac{1.40}{1.7 \times 0.22}\right)^2 \overset{(1)}{=} \left(\frac{1.40}{0.37}\right)^2 \simeq 14, \\ \text{round target of width 1.40:} \quad & \left(\frac{1.40}{1.7 \times 1.40}\right)^2 \overset{(2)}{=} \frac{1}{1.7^2} \simeq 0.35, \\ \text{ratio:} \quad & \overset{(3)}{=} \left(\frac{1.40}{0.22}\right)^2 \simeq 40. \end{aligned}$$

(1) Ridge length $L = 1.40$, step $s = 1.7 \times 0.22$.
(2) For a round target the best step is 1.7 times its own width, so it crosses in less than one step.
(3) The factor 1.7 cancels: the penalty is the square of the ratio of the two widths.

So a round step sized for the narrow direction makes the walker some forty times slower along the ridge than the same walker on a round target. In practice the best round step is larger than this, a compromise that accepts fewer proposals and moves farther along the ridge when it does, and the penalty is smaller, as the runs below show; but no round step escapes it.

Figure 9.27 shows four walkers of 4×10^4 steps on this ellipse. Tiny round steps ($s = 0.05$) are accepted 0.9286 of the time and creep along the ridge, with integrated autocorrelation time $\tau = 2981$ steps. Huge round steps ($s = 30$) are accepted a fraction 5.750×10^{-4} of the time, $\tau = 1229$. The best round step of a scan, $s = 1.166$, gives acceptance 0.2009 and $\tau = 35.55$. A proposal with the covariance of the target itself, scaled by the 2.38 rule, gives acceptance 0.3554 and $\tau = 9.278$.

Why does shaping work so well? Write the target's covariance as $\mathbf{C} = LL^{\mathsf{T}}$ (Cholesky) and change to whitened coordinates $u = L^{-1}x$, in which the target is round. A shaped proposal $y = x + sLz$ in the original coordinates is, in the new ones,

$$u' \overset{(1)}{=} L^{-1}y = L^{-1}x + sL^{-1}Lz \overset{(2)}{=} u + s\,z,$$

(1) Apply the linear map L^{-1} to the proposal $y = x + sLz$.
(2) $L^{-1}x = u$ and $L^{-1}L = \mathbb{1}$.

a round proposal. And because the map is linear, its Jacobian is a constant that cancels from every ratio. So the shaped walker on the ellipse *is* the round walker on the round Gaussian, and inherits its efficiency: with $\ell = 2.38$ and $D = 2$ the measured $\tau = 9.278$ steps is a few times D, whatever ρ is. In practice the covariance is unknown, so one runs a short pilot chain, estimates the covariance from it, and uses that for the proposal of the real run, repeating if necessary [Padilla §4.1.4], [HobsonBMC §3.2.4, p. 65]. The adaptation must stop before the samples that are kept: a proposal that keeps changing with the chain's own history is no longer a Markov chain with f as its stationary law [HobsonBMC p. 63].

Watch out: A good acceptance rate is not a good chain

The best round walker above accepted 0.2009 of its proposals, right in the recommended range, and was still several times slower than the shaped walker. The acceptance rate measures only whether the step length suits the narrowest direction. It knows nothing about whether the walker moves along the long ones. Judge a chain by its autocorrelation

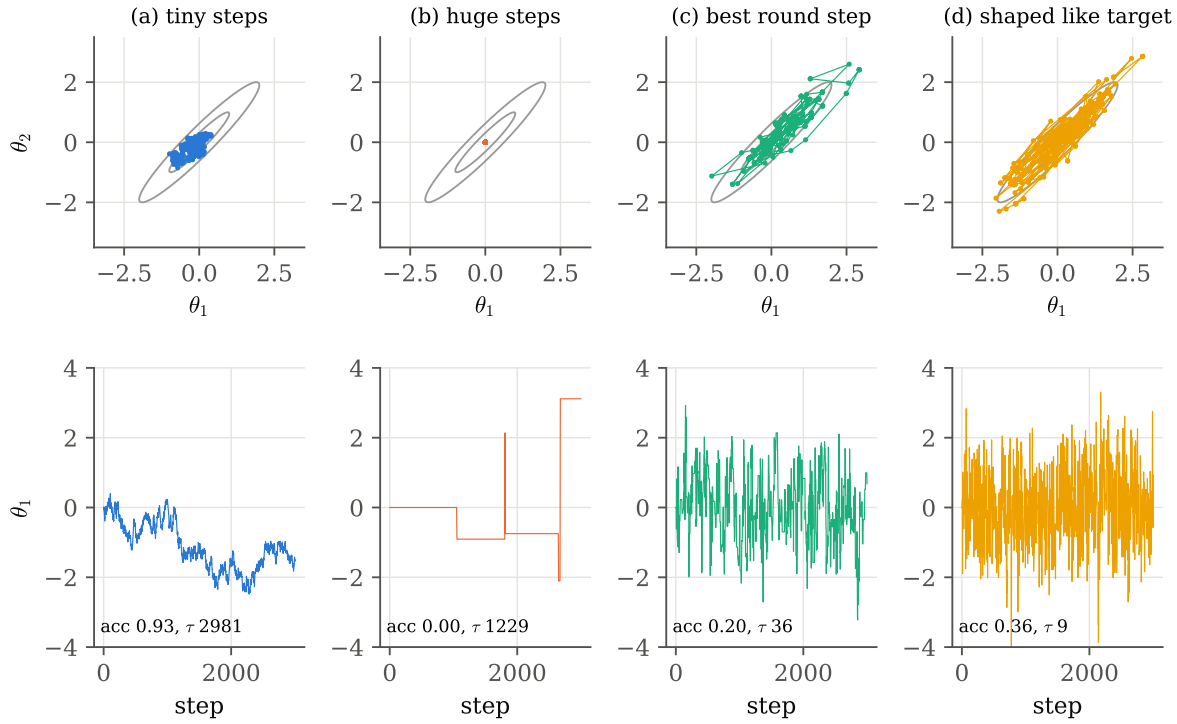


Figure 9.27: Four walkers on a Gaussian with correlation 0.95 (68% and 95% contours in grey). Top: the first 400 steps. Bottom: 3000 steps of θ_1, with acceptance rate and integrated autocorrelation time. (a) Tiny round steps creep. (b) Huge round steps are almost always refused. (c) The best round step still has to diffuse along the ridge. (d) Steps shaped like the target cover it in a few moves.

time or effective sample size (Section 9.12), never by its acceptance rate alone.

9.11.3 Curved targets: change variables

A covariance matrix describes a straight ellipse. Many posteriors are curved, because the data constrain a nonlinear combination of the parameters. The standard test case is the banana

$$p(\theta_1, \theta_2) \propto \exp\Big[-\frac{\theta_1^2}{2} - \frac{(\theta_2 - \theta_1^2)^2}{2\sigma^2}\Big], \qquad \sigma = 0.3. \tag{9.73}$$

Read it as a recipe: $\theta_1 \sim \mathcal{N}(0,1)$, and given θ_1, θ_2 is Gaussian with mean θ_1^2 and standard deviation σ. The exact moments follow from this reading:

$$\begin{aligned}
\mathbb{E}\,\theta_2 &\overset{(1)}{=} \mathbb{E}\big[\mathbb{E}(\theta_2 \mid \theta_1)\big] = \mathbb{E}\,\theta_1^2 = 1, \\
\operatorname{Var}\theta_2 &\overset{(2)}{=} \operatorname{Var}\big(\mathbb{E}(\theta_2 \mid \theta_1)\big) + \mathbb{E}\big[\operatorname{Var}(\theta_2 \mid \theta_1)\big] = \operatorname{Var}(\theta_1^2) + \sigma^2 \overset{(3)}{=} (3-1) + \sigma^2 = 2 + \sigma^2, \\
\operatorname{Cov}(\theta_1, \theta_2) &\overset{(4)}{=} \mathbb{E}\big[\theta_1\ \mathbb{E}(\theta_2 \mid \theta_1)\big] - \mathbb{E}\,\theta_1\ \mathbb{E}\,\theta_2 = \mathbb{E}\,\theta_1^3 = 0.
\end{aligned}$$

(1) Tower rule: average the conditional mean over θ_1.
(2) Law of total variance (Chapter 1).
(3) $\mathbb{E}\,\theta_1^4 = 3$ and $\mathbb{E}\,\theta_1^2 = 1$ for a standard Gaussian.
(4) Odd moments of a standard Gaussian vanish.

The last line is the trap: θ_1 and θ_2 are strongly dependent, yet their linear correlation is exactly zero. A covariance matrix cannot see a symmetric banana.

Figure 9.28 confirms it with 200 000 steps per walker. The best round walker has $\tau = 265.8$. A pilot run's covariance has correlation 0.169, which is only noise, since the true correlation is zero. So the "shaped" proposal does little more than rescale the two axes; it has acceptance 0.07981 and $\tau = 89.19$. Both walkers struggle to get into the arms of the banana. In this run their estimates of Var θ_2 are 2.323 and 1.887, against the exact 2.09. How much of that is luck? Run each walker again 12 times with fresh random numbers: the estimate moves by about 0.28 from run to run for the round walker and 0.31 for the shaped one. A slow chain sees only a few excursions into the arms, and the variance of a long-tailed quantity depends on exactly those.

The cure is to remove the curvature by a change of variables. With $u = \theta_2 - \theta_1^2$ the target becomes $\exp[-\theta_1^2/2 - u^2/2\sigma^2]$, a product of two independent Gaussians. The map $(\theta_1, \theta_2) \mapsto (\theta_1, u)$ has Jacobian determinant $\det\left(\begin{smallmatrix} 1 & 0 \\ -2\theta_1 & 1 \end{smallmatrix}\right) = 1$, so no extra factor appears in the density. A round walker in (θ_1, u), mapped back with $\theta_2 = u + \theta_1^2$, has acceptance 0.3576, $\tau = 7.486$, and Var $\theta_2 = 2.091$, which moves by only 0.05 from run to run. Lewis and Bridle give the same advice for cosmology: when a degeneracy is known, sample in the combination the data constrain well, for example the angular size of the sound horizon instead of the Hubble rate, and remember the Jacobian in the prior [HobsonBMC §3.2.4, p. 66].

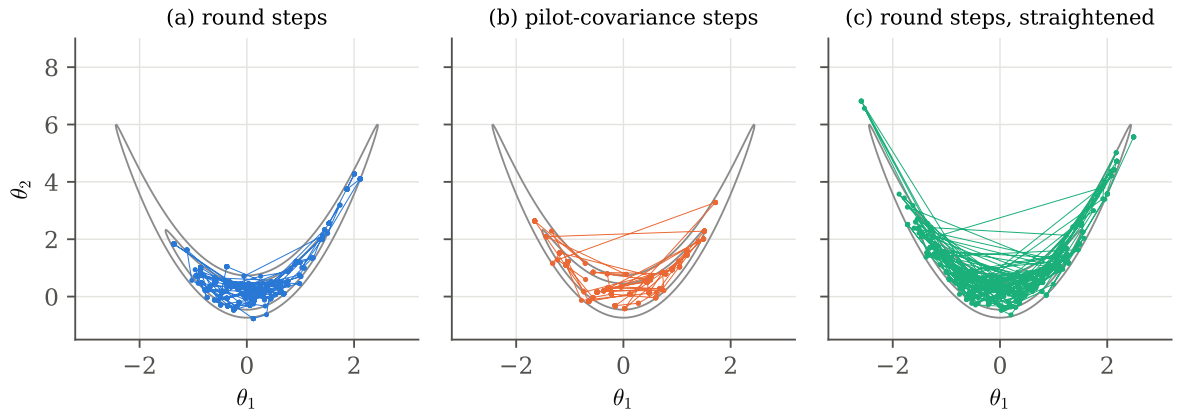


Figure 9.28: Six hundred steps of three walkers on the banana (9.73). (a) Round steps and (b) steps shaped by a pilot covariance both crawl along the curved ridge. (c) Round steps in the straightened variables $(\theta_1, \theta_2 - \theta_1^2)$, mapped back, jump across the whole banana.

Cross-check against a standard library. The affine-invariant ensemble sampler `emcee` [174], described in Section 9.16, runs many walkers that propose moves along the lines joining them. With 32 walkers and 20000 steps each on the banana it accepts 0.5143 of its moves and reports an autocorrelation time of 192.9 steps, far longer than the 7.486 of the straightened walker: an affine-invariant method is blind to linear correlations, but a banana is not linear. Its Var θ_2 is 1.861 in this run and moves by about 0.20 from run to run. The triangle plot of Figure 9.29, drawn with our own `corner` function, overlays the straightened Metropolis chain and the `emcee` chain; the cores agree, and the tips of the arms, where few samples land, are noisy in both.

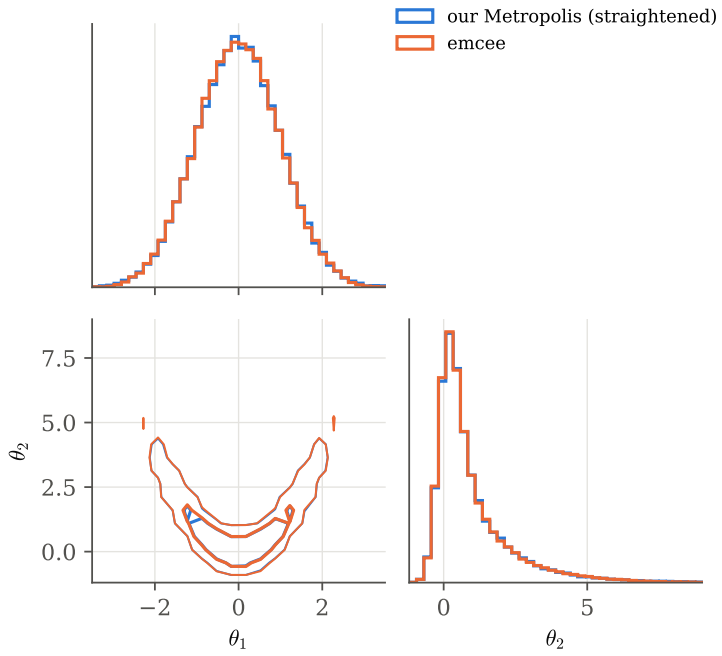


Figure 9.29: Triangle plot of the banana: one-dimensional marginal histograms on the diagonal, 68% and 95% regions of the two-dimensional histogram below it. Blue: our Metropolis sampler in the straightened variables. Orange: `emcee` in the original variables.

9.12 Reading a chain: burn-in, autocorrelation and effective sample size

A finished run is a list of numbers, one row per step. Before quoting any average from it we have to answer two questions [Kruschke §7.5]. Which part of the list is still remembering the arbitrary starting point, and must be thrown away? And how much information is in the rest: how many independent draws is it worth, and how accurate are the averages computed from it? The chain itself has to answer both, because in a real problem nothing else knows the answer.

9.12.1 Burn-in: where does equilibrium start?

Start a Metropolis walker for the standard Gaussian in $D = 10$ dimensions at the point $(10, 10, \ldots, 10)$, ten standard deviations out in every coordinate, with the 2.38 step of Section 9.11.1. The quantity that shows the approach to equilibrium most clearly is the log density along the chain, $\ln f(\theta_t)$, here $-|\theta_t|^2/2$ up to a constant. It starts at -500 and climbs (Figure 9.30b).

The question. When has the walker arrived? The tempting answer, "when it reaches the peak", is wrong, and seeing why is the most useful fact about high-dimensional posteriors.

In equilibrium $\theta \sim \mathcal{N}(0, \mathbb{1}_D)$, so $|\theta|^2 = \sum_i \theta_i^2$ is a sum of D squared standard Gaussians, a χ^2_D variable (Chapter 2). Therefore

$$\mathbb{E}\big[\ln f(\theta)\big] \overset{(1)}{=} -\tfrac{1}{2}\,\mathbb{E}\,\chi^2_D = -\frac{D}{2}, \qquad \mathrm{sd}\big[\ln f(\theta)\big] \overset{(2)}{=} \tfrac{1}{2}\sqrt{\mathrm{Var}\,\chi^2_D} = \tfrac{1}{2}\sqrt{2D} = \sqrt{D/2}. \tag{9.74}$$

(1) $\mathbb{E}\,\theta_i^2 = 1$ for each of the D terms.

(2) $\mathrm{Var}\,\theta_i^2 = \mathbb{E}\,\theta_i^4 - 1 = 3 - 1 = 2$, and the D terms are independent.

For $D = 10$ the equilibrium walker lives at $\ln f = -5 \pm 2.2$, never at the peak value 0. The chain agrees: after step 10 000 its $\ln f$ averages -4.995 with standard deviation 2.29, and the highest value it ever reached was -0.6265.

Why the walker avoids the peak. The reason is volume. Ask how probable it is to find the walker at a distance between r and $r + \mathrm{d}r$ from the centre. The density there is the same at every point, $e^{-r^2/2}$ up to a constant, and the points form a thin spherical shell whose volume is its area times its thickness. The area of a sphere of radius r in D dimensions grows as r^{D-1} (a circle's length as r, a sphere's area as r^2, and so on). So the radius has the density

$$p(r) \overset{(1)}{=} (\text{area} \propto r^{D-1}) \times (\text{density } e^{-r^2/2}) \propto r^{D-1} e^{-r^2/2},$$
$$\frac{\mathrm{d} \ln p}{\mathrm{d}r} \overset{(2)}{=} \frac{D-1}{r} - r = 0 \quad \Longrightarrow \quad r^2_{\text{peak}} \overset{(3)}{=} D - 1 \approx D. \tag{9.75}$$

(1) Probability in a thin shell is density times shell volume, and the shell volume is area times $\mathrm{d}r$.
(2) $\ln p = (D-1)\ln r - r^2/2 + \text{const}$; set its derivative to zero to find the most probable radius.
(3) Multiply by r and solve.

The first factor is zero at the centre and grows; the second is largest at the centre and dies quickly; their product peaks at $r \approx \sqrt{D}$. At that radius $\ln f = -r^2/2 \approx -D/2$, the equilibrium mean of (9.74). For $D = 10$ the most probable radius is $\sqrt{9} = 3$, where $\ln f = -4.5$, against the chain's average -4.995; the peak of f at $r = 0$ has no volume around it at all. The bulk of the probability sits on a thin shell away from the mode. Lambert calls this region the **typical set**, and describes MCMC as two phases, first finding the typical set and then exploring it [Lambert §13.7.1].

Burn-in therefore ends when $\ln f$ stops climbing and starts fluctuating around a steady level, not when it reaches the maximum. Here the trace first crossed its equilibrium median at step 147 (Figure 9.30b). A popular rule in cosmology discards the start of a chain while $p/p_{\max} < 0.1$ [171]. Read literally with the true peak, it must be scaled with dimension. For a Gaussian posterior, $-2\ln(p/p_{\max})$ is exactly the $|\theta|^2$ above, a χ^2_D variable, so the fraction of equilibrium draws that satisfy $p/p_{\max} \geq 0.1$ is $\mathrm{P}(\chi^2_D \leq 2\ln 10)$: 0.9 for $D = 2$, 0.08405 for $D = 10$, and 8.033×10^{-18} for $D = 50$. In fifty dimensions an equilibrium chain essentially never comes within a factor of ten of the peak. Any threshold on $\ln p$ has to be judged against the spread (9.74) of $\ln p$ in equilibrium. Two practical rules are common: look at the trace of $\ln p$ and of each parameter and cut where they become stationary; or, with several chains, discard the first half of each by default [Lambert §13.9.5], [Kruschke §7.5.1]. Multiple chains make the cut checkable (Section 9.13).

9.12.2 Autocorrelation and the integrated autocorrelation time

Once the burn-in is gone, the remaining N steps $x_1, \ldots, x_N$ of one parameter are a stationary but correlated series: every x_t has the same variance σ^2, and the covariance $C_k = \mathrm{Cov}(x_t, x_{t+k}) = \sigma^2 \rho_k$ depends only on the lag k, with ρ_k the autocorrelation. How accurate is the chain average $\bar{x} = \frac{1}{N}\sum_t x_t$? The derivation of Section 9.6.2, repeated in three lines:

$$\mathrm{Var}(\bar{x}) \overset{(1)}{=} \frac{1}{N^2} \sum_{s=1}^{N} \sum_{t=1}^{N} C_{|t-s|} \overset{(2)}{=} \frac{\sigma^2}{N} \left[1 + 2 \sum_{k=1}^{N-1} \left(1 - \frac{k}{N}\right) \rho_k \right] \overset{(3)}{=} \frac{\sigma^2}{N}\, \tau, \qquad \tau = 1 + 2\sum_{k=1}^{\infty} \rho_k. \tag{9.76}$$

(1) The variance of a sum is the sum of all the covariances of its terms.
(2) The N pairs with $s = t$ give $N\sigma^2$; the $2(N-k)$ ordered pairs at lag $k \geq 1$ give $2(N-k)\sigma^2\rho_k$; divide by N^2.

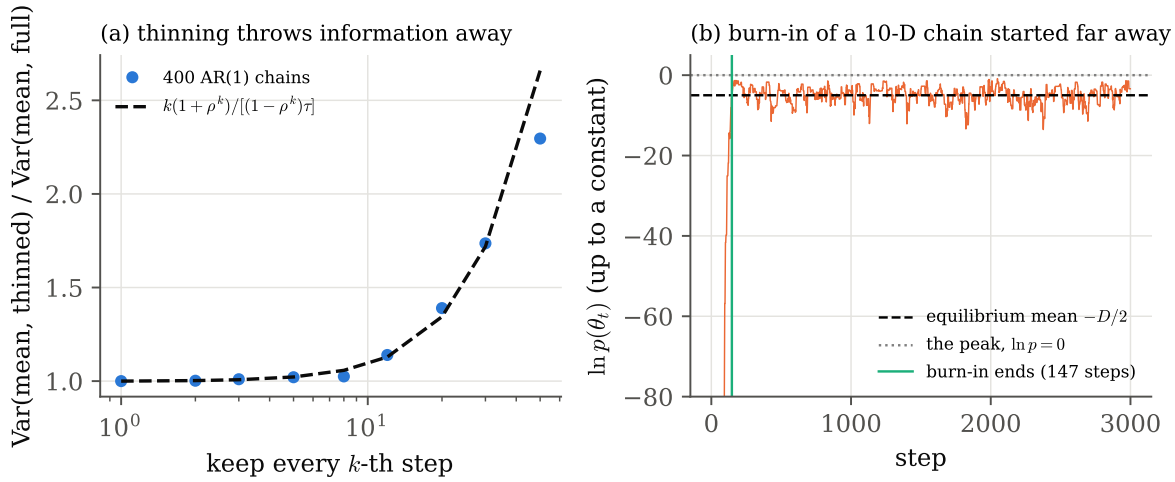


Figure 9.30: (a) Thinning an AR(1) chain with $\rho = 0.9$ by keeping every kth step: the variance of the mean of the thinned chain, relative to the full chain, measured over 400 chains (dots) and predicted (dashed). It is never below one. (b) $\ln p$ along a 10-dimensional chain started far out: it climbs from -500 and settles around $-D/2$ (dashed), well below the peak value 0 (dotted).

(3) For N much longer than the range of the correlations, k/N is negligible wherever ρ_k is not, and the sum can run to infinity.

With independent draws $\tau = 1$ and this is the familiar σ^2/N; correlation replaces N by N/τ, so the chain is worth N/τ independent draws. In Section 9.6 we computed τ exactly from the transition matrix; now we must estimate it from the series alone.

Estimating the autocorrelation. The natural estimator of the lag-k covariance averages the products of deviations k steps apart,

$$\hat{C}_k = \frac{1}{N}\sum_{t=1}^{N-k}(x_t - \bar{x})(x_{t+k} - \bar{x}), \qquad \hat{\rho}_k = \hat{C}_k/\hat{C}_0, \tag{9.77}$$

the **autocorrelation function** (ACF) of the chain [Kruschke §7.5.2]. It is computed for all lags at once with a fast Fourier transform (`acf` in `lib_mcmc.py`). One might now plug every $\hat{\rho}_k$ into (9.76). That gives exactly zero, for any data:

$$N\left[\hat{C}_0 + 2\sum_{k=1}^{N-1}\hat{C}_k\right] \overset{(1)}{=} \sum_{s=1}^{N}\sum_{t=1}^{N}(x_s - \bar{x})(x_t - \bar{x}) \overset{(2)}{=} \left[\sum_{t=1}^{N}(x_t - \bar{x})\right]^2 \overset{(3)}{=} 0. \tag{9.78}$$

(1) Group the N^2 ordered pairs (s,t) by their lag $|t-s| = k$: lag 0 gives $N\hat{C}_0$ and each lag $k \geq 1$ appears twice, giving $2N\hat{C}_k$.
(2) A double sum of products of the same terms is the square of the single sum.
(3) Deviations from the sample mean sum to zero.

So $1 + 2\sum_{k\geq 1}\hat{\rho}_k = 0$ identically; our script finds it to be zero to within 1.4×10^{-14}, rounding error. The estimated autocorrelations at large lags are almost pure noise, and they are forced to cancel the real correlation at small lags. The sum must be cut off at a window M beyond which the true ρ_k is negligible but the noise is not yet large.

Choosing the window. Kruschke stops the sum at the first lag with $\hat{\rho}_k < 0.05$ [Kruschke eq. 7.11]. Stan, following Geyer, stops when the sum of two consecutive estimates first turns negative

[Lambert §13.10]. Our library uses Sokal's self-consistent window: the smallest M with $M \geq c\,\hat{\tau}(M)$, where $\hat{\tau}(M) = 1 + 2\sum_{k=1}^{M} \hat{\rho}_k$ and $c \simeq 5$ [175]. The window then grows with the correlation time itself, so that slowly decaying correlations are not cut short.

Why the window scales with τ, and why $c \simeq 5$. Cutting the sum at M makes two errors that pull in opposite directions. The first is the correlation left out. If the true autocorrelation decays as $\rho_k \simeq e^{-k/T}$ over some number of steps T, the sum gives $\tau \simeq 2T$ and the omitted tail is $2\sum_{k>M} e^{-k/T} \simeq 2T\, e^{-M/T}$, a fraction $e^{-M/T}$ of τ. The second is noise. Each $\hat{\rho}_k$ fluctuates by about $1/\sqrt{N}$ (a little more for a correlated series), and adding M of them adds noise that grows like $\sqrt{M}$; the careful calculation [175] gives

$$\frac{\operatorname{Var}(\hat{\tau})}{\tau^2} \simeq \frac{2(2M+1)}{N}. \tag{9.79}$$

The omitted part shrinks exponentially in M/T, the noise grows only like $\sqrt{M}$, so the right window is a few times the correlation length, and since T is not known the estimate $\hat{\tau}$ itself is used as the ruler: $M = c\hat{\tau}$. With $c = 5$ the omitted fraction is e^{-10} for a single exponential and stays below $e^{-5} \simeq 0.7\%$ even if a slow component makes T as long as τ, while the relative noise is $\sqrt{2(10\tau+1)/N}$, still small once N is a few hundred times τ. For the AR(1) test below, $\tau = 19$ and $N = 10^4$ give $M \simeq 95$ and a predicted spread of $\sqrt{191 \times 2/10^4} \times 19 \simeq 3.7$ in $\hat{\tau}$.

Listing 9.6: Integrated autocorrelation time with Sokal's window (`code/ch09/lib_mcmc.py`)

```
def tau_int(x, c=5.0):
    """Integrated autocorrelation time tau = 1 + 2 sum_{k=1}^{M} rho_k with Sokal's window:
    the smallest M such that M >= c * tau(M).  Returns (tau, M)."""
    rho = acf(x)
    taus = 2.0 * np.cumsum(rho) - 1.0                    # taus[M] = 1 + 2 sum_{k=1}^{M} rho_k
    ok = np.arange(rho.size) >= c * taus
    M = int(np.argmax(ok)) if ok.any() else rho.size - 1
    return float(taus[M]), M
```

Example: calibrating on a chain with known τ. The AR(1) chain $x_{t+1} = \rho x_t + \sqrt{1-\rho^2}\, z_t$ of Section 9.7.4 has $\rho_k = \rho^k$ exactly, so

$$\tau = 1 + 2\sum_{k\geq 1} \rho^k = 1 + \frac{2\rho}{1-\rho} = \frac{1+\rho}{1-\rho},$$

which is 19 for $\rho = 0.9$. Over 300 chains of 10 000 steps, Kruschke's rule gives $\hat{\tau} = 18.33 \pm 1.779$ (mean and spread over the chains), slightly low because it drops the tail of the correlation; Sokal's window gives 18.95 ± 3.412, unbiased but noisier (Figure 9.31a), with a spread close to the 3.7 that (9.79) predicts. The prediction (9.76) itself checks out: across the same kind of chains, $N\operatorname{Var}(\bar{x})/\sigma^2 = 18.84$ against $\tau = 19$. As a cross-check against a standard library, `emcee`'s estimator `integrated_time`, applied to a Metropolis chain of 2×10^5 steps on a standard Gaussian, returns $\tau = 4.336$, and ours returns 4.336.

Working definition 9.10 (effective sample size and Monte Carlo standard error)

For a chain of N steps of one quantity with integrated autocorrelation time τ, the **effective sample size** is

$$\mathrm{ESS} = \frac{N}{\tau} = \frac{N}{1 + 2\sum_{k\geq 1} \rho_k}$$

[the number of independent draws that would give the mean with the same accuracy]
[Kruschke eq. 7.11], [Lambert §13.10], and the **Monte Carlo standard error** of the chain average is $\mathrm{MCSE} = \mathrm{SD}/\sqrt{\mathrm{ESS}} = \mathrm{SD}\sqrt{\tau/N}$ [Kruschke eq. 7.12], with SD the standard deviation of the chain. *Operational recipe:* discard the burn-in, estimate τ with a window (`tau_int`), and quote ESS for every parameter; the smallest one counts.

Note the difference between the two numbers in the box. The MCSE is the error of our *computation* of the posterior mean, the error that a longer run would reduce. The SD is the width of the posterior itself, the uncertainty about the parameter that no amount of computing removes. A run is long enough when the first is small compared with the second.

Example: Kruschke's coin chains. Figure 9.31b shows the ACFs of the three coin chains of Section 9.10. With proposal width 0.02 the ACF is still high after 60 steps; with width 2 it decays slowly too, because the chain repeats the same value for long stretches; with width 0.2 it falls to near zero within ten lags. These shapes are why the three effective sizes differed so much. For the middle chain, ESS ≈ 11181 and the posterior SD is about 0.1, so the posterior mean is computed to MCSE $\approx 0.1/\sqrt{11181} \approx 0.001$, far below the SD. Kruschke recommends an ESS of about 10 000 when the limits of a 95% interval are wanted, since tails are sampled more sparsely than the centre [Kruschke §7.5.2].

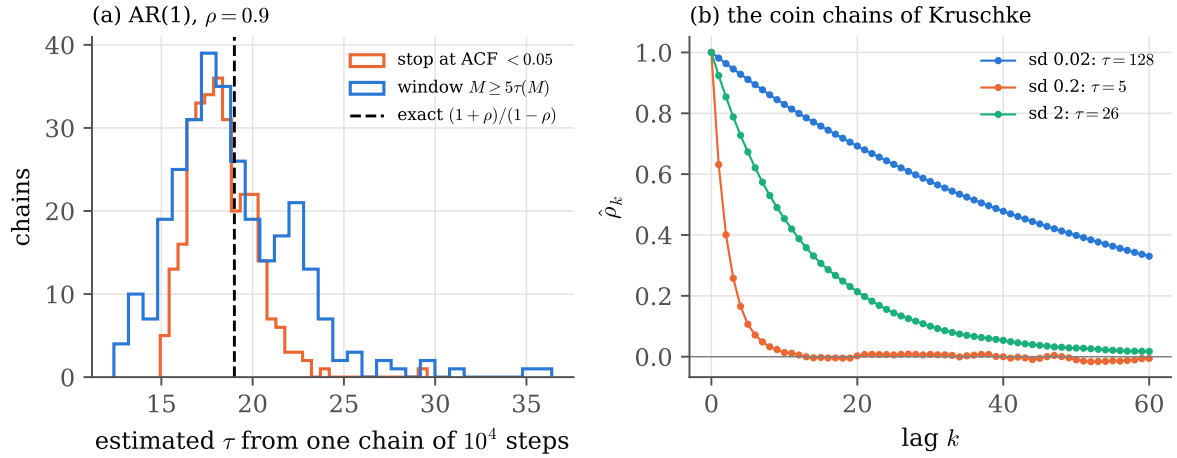


Figure 9.31: (a) Estimates of τ from single AR(1) chains of 10^4 steps with exact $\tau = 19$ (dashed): stopping at the first $\hat{\rho}_k < 0.05$ is slightly biased low, Sokal's self-consistent window is centred on the truth. (b) Autocorrelation functions of Kruschke's three coin chains, with their estimated τ.

9.12.3 Thinning

Keeping only every kth step makes the stored points look less correlated, and is often done. Does it make the averages better? For the AR(1) chain $x_{t+1} = \rho x_t + \sqrt{1-\rho^2}\, z_t$ (Section 9.12.2, with z_t independent standard Gaussians) the answer can be computed exactly.

The thinned chain is again AR(1). Apply the recursion k times:

$$x_{t+k} \overset{(1)}{=} \rho^k x_t + \sqrt{1-\rho^2} \sum_{j=0}^{k-1} \rho^j z_{t+k-1-j},$$

$$\operatorname{Var}\Big(\sqrt{1-\rho^2}\sum_{j=0}^{k-1}\rho^j z_{t+k-1-j}\Big) \overset{(2)}{=} (1-\rho^2)\frac{1-\rho^{2k}}{1-\rho^2} = 1-\rho^{2k}.$$

(1) Substitute $x_{t+k-1} = \rho x_{t+k-2} + \dots$ into itself k times; each earlier kick arrives multiplied by one more factor ρ.
(2) The kicks are independent with variance 1, so the variances add: $(1-\rho^2)\sum_{j=0}^{k-1}\rho^{2j}$, a geometric sum.

So the kept points obey $x_{t+k} = \rho^k x_t + \sqrt{1-\rho^{2k}}\,z'$ with a fresh standard Gaussian z', built from kicks that no other kept step uses: the same AR(1) form with coefficient ρ^k. Its stationary variance is still $\sigma^2 = 1$, as it must be, since every x_t in equilibrium has the same distribution whichever steps we keep. The thinned chain has N/k points. By (9.76) with the AR(1) τ,

$$\frac{\operatorname{Var}(\bar{x}_{\text{thinned}})}{\operatorname{Var}(\bar{x}_{\text{full}})} \overset{(1)}{=} \frac{\dfrac{\sigma^2 k}{N}\dfrac{1+\rho^k}{1-\rho^k}}{\dfrac{\sigma^2}{N}\dfrac{1+\rho}{1-\rho}} \overset{(2)}{=} \frac{k\coth(ka/2)}{\coth(a/2)} \overset{(3)}{\geq} 1, \qquad \rho = e^{-a}. \tag{9.80}$$

(1) Numerator: N/k points with correlation time $(1+\rho^k)/(1-\rho^k)$. Denominator: N points with $\tau = (1+\rho)/(1-\rho)$.
(2) $(1+e^{-b})/(1-e^{-b}) = \coth(b/2)$, used with $b = ka$ and $b = a$.
(3) The function $u\coth u$ increases for $u > 0$: its derivative is $\coth u - u/\sinh^2 u = (\tfrac{1}{2}\sinh 2u - u)/\sinh^2 u > 0$, because $\sinh v > v$ for $v > 0$. Hence $(ka/2)\coth(ka/2) \geq (a/2)\coth(a/2)$ for $k \geq 1$; multiply by $2/a$.

Thinning never helps the average; it can only lose information. For $\rho = 0.9$ and $k = 20$ the predicted ratio is 1.344 and 400 simulated chains give 1.39 (Figure 9.30a). The variance grows by a third although only one point in twenty was kept; when the steps are strongly correlated the discarded ones carried little, but never nothing.

Watch out: Thinning is for storage, not accuracy

Thinning a chain does not make it "more independent" in any useful sense: the thinned average is never more accurate than the full one (9.80). Thin only to save memory or disk, for example when each step stores a whole map [Lambert §13.10.1], [Padilla §4.1.4], and report the ESS of what you kept.

9.13 Many chains from far apart: the Gelman–Rubin statistic

A single chain can look perfectly healthy and be wrong. If the posterior has a second region of high probability that the walker never found, the trace is stationary, the ACF decays, the ESS is large, and every number in Section 9.12 is computed from the wrong distribution. Nothing inside one chain can reveal a region the chain has never visited. The remedy is to send out several walkers from starting points spread far wider than the posterior is expected to be, and to see whether they end up telling the same story [Padilla §4.1.3], [Kruschke §7.5.1].

The question. This was the question Gelman and Rubin [38] set out to answer: given M chains of N steps each, started far apart, how can we tell from the chains alone that none of them still remembers where it started? Put differently, how do we turn "they agree" into a number, and how close to its ideal value must the number be?

Example: four walkers on the two bumps. Take again the two-bump density (9.57), with bumps at $x = -2$ and $x = +2$, and start four Metropolis walkers (step 1.5) at $x_0 = -30$, -10, $+10$ and $+30$, ten to fifteen posterior widths away. Figure 9.32a shows that within a hundred steps all four have fallen into the region of the bumps and are hopping between them. The histograms of the second halves of the four chains, 20 000 steps each, lie on top of one another and of f (Figure 9.32b). Their four means range from 0.7282 to 0.8603, around the exact 0.8.

Within and between. Index the chains by $j = 1, \dots, M$ and the steps by $i = 1, \dots, N$, so θ_{ij} is step i of chain j; let $\bar\theta_j$ be the mean of chain j and $\bar\theta$ the mean of everything. Two quantities summarise the spread [Padilla eqs. 57–60], [Lambert §13.9.3]:

$$W = \frac{1}{M}\sum_{j=1}^{M} s_j^2, \quad s_j^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\theta_{ij}-\bar\theta_j)^2, \qquad \frac{B}{N} = \frac{1}{M-1}\sum_{j=1}^{M}(\bar\theta_j-\bar\theta)^2. \tag{9.81}$$

W is the average variance *within* a chain. B/N is the variance *between* the chain means. The sources write these with slightly different letters:

	here	Gelman and Rubin [38], Lambert	Padilla, eq. 59
variance of the chain means	B/N	B/N (their B is N times it)	B
average within-chain variance	W	W	W

Gelman and Rubin multiply the variance of the means by N so that their B is comparable with the variance of single draws; we keep their B and always write it as B/N. If the chains have not mixed, each one has explored only part of the posterior: W is too small, because no chain has seen the full spread, and B/N is too large, because the chains sit in different places.

How are W and B/N related when the chains *are* all in equilibrium? Write σ^2 for the posterior variance and $v = \mathrm{Var}(\bar\theta_j)$ for the variance of a chain mean, which by (9.76) is $\sigma^2\tau/N$. Then

$$\mathbb{E}\Big[\frac{N-1}{N}W\Big] \overset{(1)}{=} \mathbb{E}\Big[\frac{1}{N}\sum_{i=1}^{N}(\theta_{ij}-\bar\theta_j)^2\Big] \overset{(2)}{=} \mathbb{E}\Big[\frac{1}{N}\sum_{i}(\theta_{ij}-\mu)^2\Big] - \mathbb{E}(\bar\theta_j-\mu)^2 \overset{(3)}{=} \sigma^2 - v,$$
$$\mathbb{E}\Big[\frac{B}{N}\Big] \overset{(4)}{=} v, \qquad \text{so} \qquad \mathbb{E}\Big[\frac{N-1}{N}W + \frac{B}{N}\Big] = \sigma^2. \tag{9.82}$$

(1) Definition of s_j^2; all chains have the same expectation.

(2) With μ the posterior mean, $\sum_i(\theta_{ij}-\bar\theta_j)^2 = \sum_i(\theta_{ij}-\mu)^2 - N(\bar\theta_j-\mu)^2$, the parallel-axis theorem for the sum of squares.

(3) Each θ_{ij} has variance σ^2; $\mathbb{E}(\bar\theta_j-\mu)^2 = v$. This holds for correlated draws as well: correlation only changes v from σ^2/N to $\sigma^2\tau/N$.

(4) Independent chains: the chain means are M independent draws with variance v, and (9.81) is their unbiased sample variance.

So the pooled estimate $\hat\sigma_+^2 = \frac{N-1}{N}W + \frac{B}{N}$ is unbiased for the posterior variance once the chains are in equilibrium. From here Gelman and Rubin's statistic follows in three moves.

Before equilibrium, the two estimates err in opposite directions. Started overdispersed, the chain means are spread not only by the noise v of (9.82) but also by the memory of their starting points, which were chosen wider than the posterior. That extra spread enters B/N, so $\hat\sigma_+^2$ is too *large*. At the same time each chain has explored only the part of the posterior near where it is, and the sample variance s_j^2 of a chain that has not yet seen the whole posterior is too *small*; so is

their average W. The true σ^2 lies between the two:

$$\mathbb{E}\,W \;\lesssim\; \sigma^2 \;\lesssim\; \mathbb{E}\,\hat{\sigma}_+^2,$$

and both inequalities become equalities as the chains forget their starts.

The uncertainty of the centre. The pooled estimate describes the spread of draws around the true posterior mean μ. But μ is not known; it is estimated by the grand mean $\bar{\theta}$, the average of M chain means, each with variance B/N. Its variance is therefore $B/(MN)$, and the spread of a draw around the *estimated* centre is

$$\hat{V} \overset{(1)}{=} \hat{\sigma}_+^2 + \frac{B}{MN} \overset{(2)}{=} \frac{N-1}{N}\,W + \Big(1+\frac{1}{M}\Big)\frac{B}{N}.$$

(1) The variance of a draw around $\bar{\theta}$ is the variance around μ plus the variance of $\bar{\theta}$ itself: variances of independent pieces add, and one draw out of MN barely moves $\bar{\theta}$, so the two are very nearly independent.
(2) Collect the two terms in B/N.

With few chains the centre is poorly known and the correction matters: for $M = 4$ it adds a quarter to the between-chain term.

The ratio, in units of a standard deviation. $\hat{V}$ errs high and W errs low, so $\hat{V}/W$ is above 1 while the chains still remember their starts and tends to 1 as they mix. Its square root compares two standard deviations instead of two variances: it is the factor by which the present estimate of the posterior's *scale* could still shrink if the chains ran forever, hence the name [38].

Working definition 9.11 (potential scale reduction factor $\hat{R}$)

For M chains of N post-burn-in steps of one parameter, with W and B/N as in (9.81),

$$\hat{V} = \frac{N-1}{N}\,W + \Big(1+\frac{1}{M}\Big)\frac{B}{N}, \qquad \hat{R} = \sqrt{\hat{V}/W} \tag{9.83}$$

[by how much the estimated posterior spread could still shrink if the chains ran forever]. $\hat{R} \to 1$ from above as the chains mix. *Operational test:* compute $\hat{R}$ for every parameter, on the second halves of the chains, and keep running until all are below a threshold. Kruschke worries above 1.1 [Kruschke §7.5.1]; current practice often asks for 1.01. Splitting every chain into two halves before computing $\hat{R}$ ("split-$\hat{R}$") also catches a single chain that is still drifting [Lambert §13.9.4].

Padilla's version [Padilla eq. 61] is the ratio $\hat{V}/W$ without the square root, with the criterion $0.97 < \hat{R} < 1.03$; since $\sqrt{1+\epsilon} \simeq 1+\epsilon/2$, a threshold of 1.03 on $\hat{V}/W$ is roughly 1.015 on $\sqrt{\hat{V}/W}$. When reading a paper, check which is meant.

How close to 1 is close enough? Even perfectly converged chains give $\hat{R} > 1$, because W and B are estimated from finite chains. Using (9.82) with $\mathbb{E}\,W = \frac{N}{N-1}(\sigma^2 - v)$,

$$\hat{R}^2 \overset{(1)}{=} \frac{\hat{V}}{W} \simeq \frac{N-1}{N}\,\frac{\sigma^2 + v/M}{\sigma^2 - v} \overset{(2)}{=} \Big(1-\frac{1}{N}\Big)\frac{1+\tau/(MN)}{1-\tau/N} \overset{(3)}{=} 1 + \Big[\Big(1+\frac{1}{M}\Big)\tau - 1\Big]\frac{1}{N} + O\big((\tau/N)^2\big),$$

$$\hat{R} - 1 \overset{(4)}{=} \frac{(1+1/M)\,\tau - 1}{2N} \simeq \Big(1+\frac{1}{M}\Big)\frac{1}{2\,\mathrm{ESS}_{\mathrm{chain}}}. \tag{9.84}$$

(1) Replace each estimate by its expectation: $\hat{V} \simeq \sigma^2 + v/M$ and $W \simeq \frac{N}{N-1}(\sigma^2 - v)$.
(2) $v = \sigma^2\tau/N$; divide numerator and denominator by σ^2.

(3) $1/(1-\epsilon) \simeq 1+\epsilon$ for small $\epsilon = \tau/N$, and multiply out, keeping terms of order $1/N$. The $-1/N$ comes from the factor $(N-1)/N$; it is as large as the other terms when τ is a few, so it cannot be dropped.

(4) $\sqrt{1+\epsilon} \simeq 1+\epsilon/2$. The last form holds when $\tau \gg 1$; $\mathrm{ESS}_{\mathrm{chain}} = N/\tau$ is the effective sample size of one chain.

In equilibrium $\hat{R}-1$ is about half the inverse effective size per chain. A threshold of 1.01 with four chains therefore asks for at least $1.25/(2\times 0.01) \simeq 60$ effective draws per chain; a threshold of 1.1 asks for only about six, which is why it is too lax for quoting tails. A calibration with 200 replicate sets of $M = 4$ chains of $N = 2000$ steps on a standard Gaussian, all started in equilibrium, with mean $\tau = 4.338$, gives an average $\hat{R}-1 = 0.00111 \pm 7.644\times 10^{-5}$, against 0.001106 from the first form of (9.84).

Back to the four walkers on the two bumps. Figure 9.32c shows $\hat{R}-1$ computed from steps $n/2$ to n as the chains grow. After 100 steps $\hat{R}$ is still 1.618, dominated by the starting points; $\hat{R}-1$ falls below 0.01 after 847 steps and ends at $\hat{R} = 1.001$ (split-$\hat{R} = 1.002$).

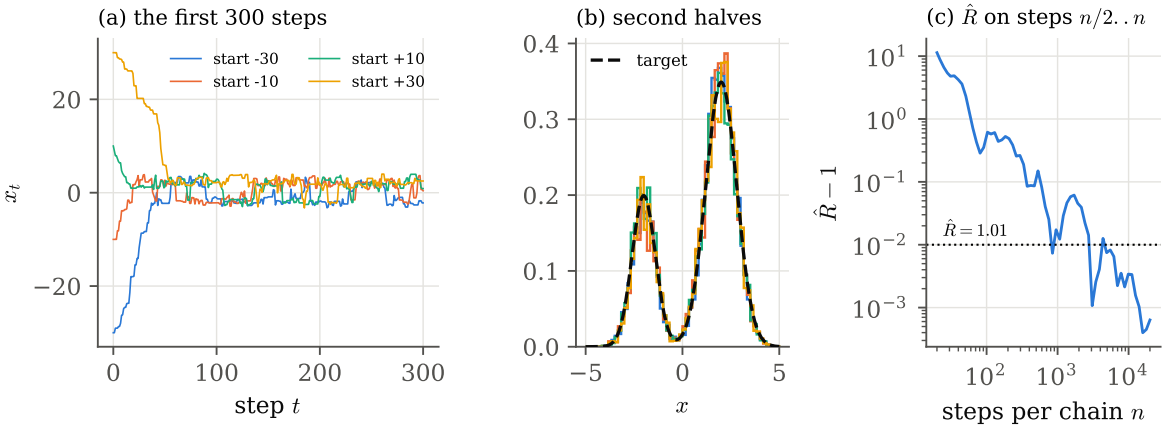


Figure 9.32: Four Metropolis chains on the two-bump density, started at −30, −10, +10 and +30. (a) The first 300 steps: every chain falls into the bumps and starts hopping between them. (b) Histograms of the second halves of the four chains, with the target (dashed). (c) $\hat{R}-1$ on the second half of the steps so far, falling through the 1.01 line.

Example: what $\hat{R}$ sees, and what it cannot see. Now move the bumps to $x = \pm 6$, keep their widths, and use short steps of 0.5. The valley between the bumps is now so deep that a walker essentially never crosses it. Four walkers from −30, −10, +10 and +30 settle into the bump on their own side and stay there (Figure 9.33a). Each chain on its own looks excellent: its trace is stationary, and its autocorrelation time is only about 19.55 steps. Only the comparison between chains reveals the problem: $\hat{R} = 11.1$.

Then start all four walkers on the right, at 8, 12, 20 and 30 (Figure 9.33b). They all fall into the right bump and agree perfectly: $\hat{R} = 1.000$. Their common mean is 6.008, while the true mean of the target is $0.3\times(-6) + 0.7\times 6 = 2.4$. Thirty per cent of the probability is in a bump that none of the walkers has seen, and no statistic computed from them can know.

Watch out: $\hat{R} \simeq 1$ is necessary, not sufficient

$\hat{R}$ compares chains with each other. It detects chains that disagree; it cannot detect chains that agree on the wrong answer because none of them reached part of the posterior [Kruschke §7.5.1]. Start the chains genuinely overdispersed, wider than any plausible

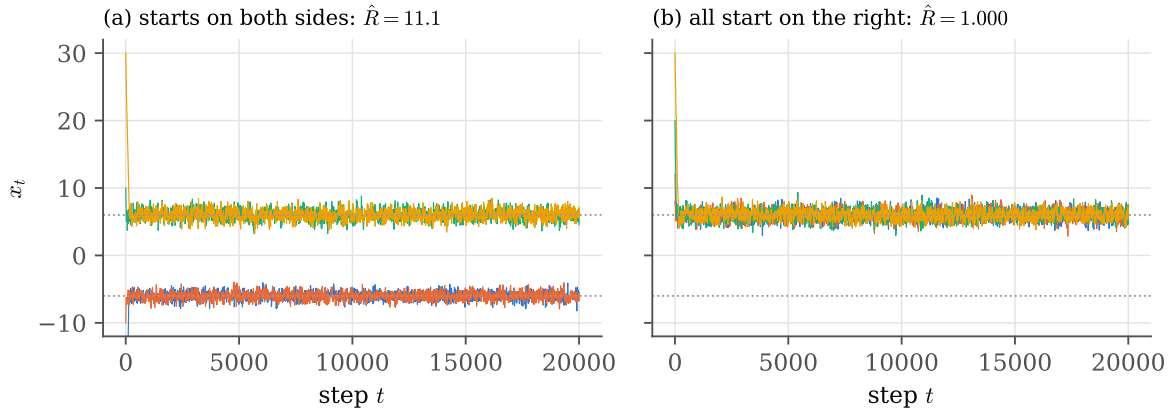


Figure 9.33: The same test on a target with bumps at ±6 (dotted), sampled with short steps. (a) Chains started on both sides stay on their own side; $\hat{R}$ is far above one and exposes the problem. (b) Chains all started on the right agree with each other perfectly and $\hat{R}$ is one, yet all of them miss the left bump.

posterior and on all sides of it, and be most suspicious when the posterior may have several separated modes. The same trap reappears in physics as broken ergodicity: a double well at low temperature (Section 9.15).

Lambert distinguishes the two ways chains can fail to converge [Lambert §13.9.4]: poor mixing *between* chains, as in Figure 9.33a, which ordinary $\hat{R}$ catches; and poor mixing *within* a chain, a chain whose first half and second half disagree because it is still drifting, which only split-$\hat{R}$ catches. Brooks and Gelman [176] extended the diagnostic to many parameters at once; modern practice computes split-$\hat{R}$ for every parameter, together with the ESS of Section 9.12.2.

9.14 Forgetting the start is not forgetting the prior

Two things in an MCMC analysis are chosen before any step is taken, and both are sometimes loosely called "initial conditions": the point where the walker starts, and the prior. It is easy to think that a long enough chain forgets both. It forgets only one of them.

The starting point is forgotten by the chain. That is ergodicity (Section 9.5): for an irreducible aperiodic chain the distribution of X_t approaches the stationary distribution geometrically fast from any start (the Doeblin contraction of Section 9.5.3), and the three crowds of the AR(1) chain, started twenty standard deviations apart, ended with the same histogram (Section 9.7.4). The four walkers of Section 9.13 did the same on the two bumps.

The prior is not part of the walk; it is part of the target. The walker samples $f(\boldsymbol{\theta}) \propto \mathcal{L}(\boldsymbol{\theta})\, p(\boldsymbol{\theta})$, and a different prior is a different f. A chain run with a wrong prior converges, faithfully and from any start, to the posterior of that wrong prior. Running it longer only makes the wrong answer more precise. What forgets the prior is more *data*: the likelihood is a product over data points and grows sharper as their number N grows, while the prior stays as it is, so for large N the posterior is governed by the likelihood. That is the Bernstein–von Mises statement of Chapter 5.

A physicist has met this distinction before. A gas in a box forgets its initial microstate as it thermalises. It does not forget its Hamiltonian: an external potential added to the Hamiltonian changes the equilibrium state, and waiting longer only brings the gas closer to the equilibrium of that Hamiltonian. In the language of Section 9.8.1, $-\ln f = -\ln \mathcal{L} - \ln p$ plays the role of the

energy, and the prior is a fixed external potential $-\ln p$ that the data's potential $-\ln \mathcal{L}$, growing in proportion to N, eventually overwhelms.

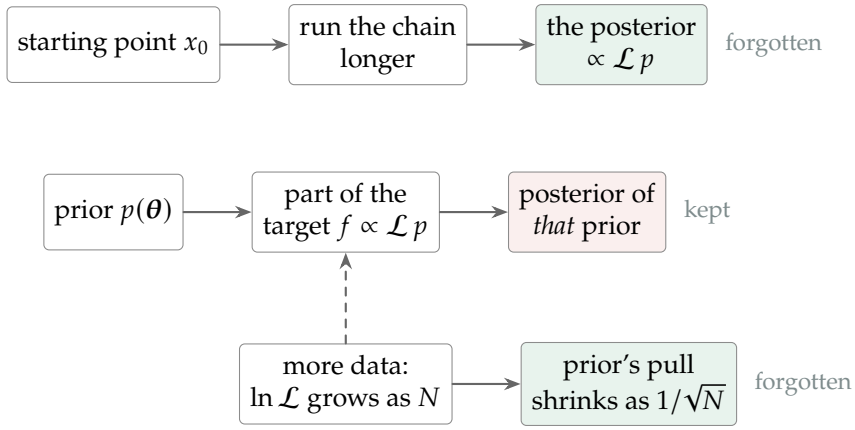


Figure 9.34: Two different kinds of forgetting. Top: the starting point is erased by running the chain (ergodicity). Middle: the prior is built into the target, so running the chain longer leads to the posterior of that prior. Bottom: only more data, through a likelihood that sharpens with N, reduce the prior's influence.

The question. Can we see both statements in one experiment? Same data; two priors, one sensible and one confidently wrong; several chains from wildly different starting points; few data and many data.

The experiment. A detector counts events from a Poisson process with true rate $\lambda = 4$ per unit exposure; this is the counting situation of a collider bin or a photon detector (Chapter 2). We observe N exposures with counts $n_1, \dots, n_N$ and total $S = \sum_i n_i$. Two priors for λ:

A: flat on $\lambda > 0$;

B: confident and wrong, $\ln \lambda \sim \mathcal{N}(\ln 10, 0.2^2)$, that is, "$\lambda$ is 10, give or take 20%".

The log posterior, up to a constant, is

$$\ln f(\lambda) = \underbrace{S \ln \lambda - N\lambda}_{\ln \mathcal{L}} + \ln p(\lambda), \qquad \ln p_B(\lambda) = -\ln \lambda - \frac{(\ln \lambda - \ln 10)^2}{2 \times 0.2^2}, \tag{9.85}$$

where the Poisson likelihood $\prod_i e^{-\lambda} \lambda^{n_i}/n_i!$ contributes $S \ln \lambda - N\lambda$ after dropping the $n_i!$, which do not depend on λ, and $\ln p_B$ includes the Jacobian $1/\lambda$ of the log-normal density. Under prior A the posterior is known exactly:

$$p_A(\lambda \mid \boldsymbol{d}) \overset{(1)}{=} \frac{\lambda^S e^{-N\lambda}}{\int_0^\infty t^S e^{-Nt}\,\mathrm{d}t} \overset{(2)}{=} \frac{N^{S+1}}{S!}\,\lambda^S e^{-N\lambda}, \qquad \mathbb{E}_A(\lambda \mid \boldsymbol{d}) \overset{(3)}{=} \frac{S+1}{N}.$$

(1) Likelihood times a flat prior, normalised.

(2) $\int_0^\infty t^S e^{-Nt}\,\mathrm{d}t = S!/N^{S+1}$ (substitute $u = Nt$; Gamma integral).

(3) The mean of a Gamma density with shape $S+1$ and rate N; the same integral with $S+1$ in place of S.

Under prior B the posterior is computed by one-dimensional quadrature on a fine grid, as the reference. For each prior and each data set we run four Metropolis–Hastings chains of 20 000 steps with multiplicative steps $\lambda \to \lambda e^{sZ}$ and the Hastings factor y/x of Section 9.9.1, started at $\lambda_0 = 0.05$, 1, 40 and 200, four decades apart.

Few data: the start is forgotten, the prior is not. With $N = 3$ exposures the counts were $(3, 3, 1)$, $S = 7$; these happen to be low for a true rate of 4, as three Poisson counts easily are. Within some tens of steps the four chains from four decades apart have merged (Figure 9.35a), and their second-phase means differ by only 0.03322; $\hat{R} = 1.000$ for prior B and 1.000 for prior A. The starting points are gone. The priors are not. The posterior mean is 2.669 under A (exact $(S + 1)/N = 2.667$) and 6.297 under B (quadrature: 6.291), 3.83 posterior standard deviations apart (Figure 9.35c). Running prior B ten times longer, 200 000 steps, gives 6.303: no movement at all towards the truth, because there is nothing for it to move towards. The chain has already found its target.

Many data: the likelihood takes over. With $N = 300$ exposures, $S = 1183$. The chains again forget their starts, more slowly now because the posterior is narrow and the steps small (Figure 9.35b). And now the two priors nearly agree: 3.947 under A (exact 3.947) and 4.018 under B, only 0.621 posterior standard deviations apart, both close to the true 4 (Figure 9.35d). The data have outweighed even a prior that claimed 10 within 20%.

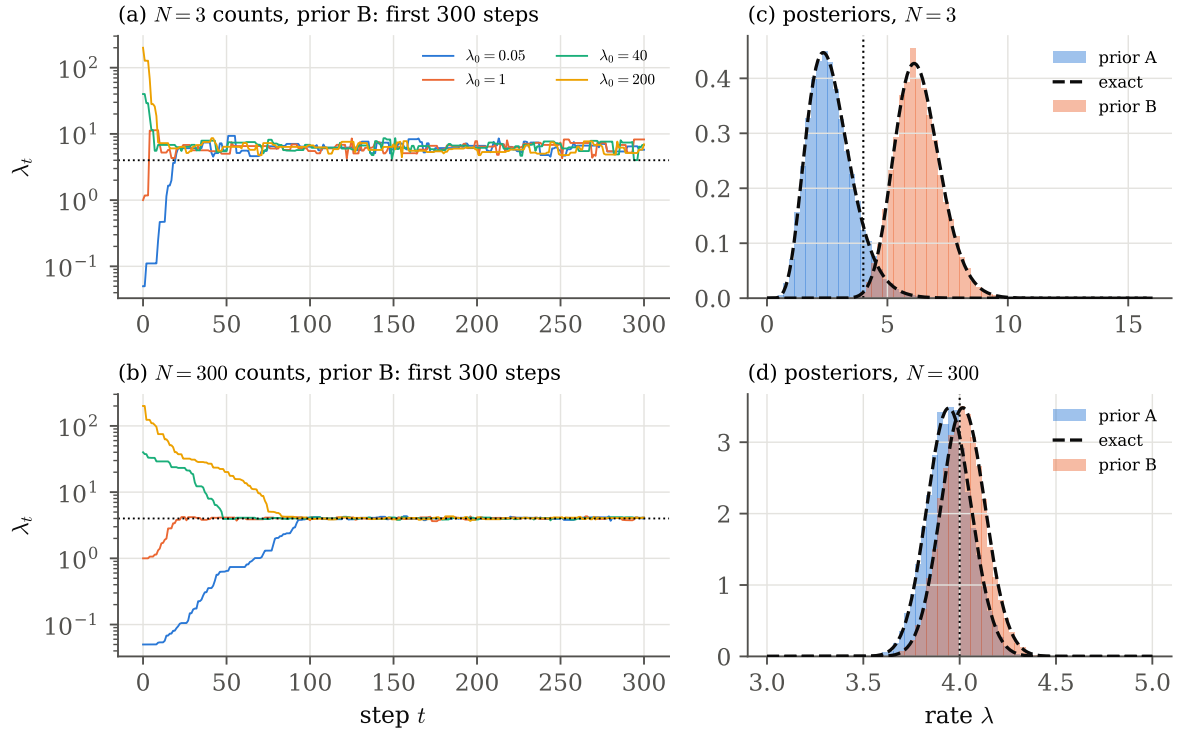


Figure 9.35: Starting point versus prior for a Poisson rate (true value 4, dotted). Left: the first 300 steps of four chains under the wrong prior B, started at $\lambda_0 = 0.05$, 1, 40 and 200 (logarithmic axis); with 3 counts (a) and 300 counts (b) they merge, so the start is forgotten. Right: posteriors from the chains (histograms) and exact (dashed) under the flat prior A and the wrong prior B. With 3 counts (c) the two priors give different answers; with 300 counts (d) they nearly agree.

How fast is the prior forgotten? The rate can be predicted by expanding the log posterior around the peak of the likelihood. The log likelihood $S \ln \lambda - N\lambda$ peaks at $\hat{\lambda} = S/N$ with curvature $-S/\hat{\lambda}^2 = -N/\hat{\lambda}$, so near the peak it is a Gaussian of variance $\sigma_L^2 = \hat{\lambda}/N$. Add the log prior, approximated by its tangent line at $\hat{\lambda}$:

$$\ln f(\lambda) \overset{(1)}{=} \text{const} - \frac{(\lambda - \hat{\lambda})^2}{2\sigma_L^2} + (\ln p)'(\hat{\lambda})\,(\lambda - \hat{\lambda})$$

$$\overset{(2)}{=} \text{const}' - \frac{\left[\lambda - \hat{\lambda} - \sigma_L^2 (\ln p)'(\hat{\lambda})\right]^2}{2\sigma_L^2},$$

$$\frac{\text{shift}}{\text{width}} \overset{(3)}{=} \frac{\sigma_L^2 \, (\ln p)'(\hat{\lambda})}{\sigma_L} = \sqrt{\frac{\hat{\lambda}}{N}} \, (\ln p)'(\hat{\lambda}). \tag{9.86}$$

(1) Second-order Taylor expansion of $\ln \mathcal{L}$ at its maximum, first-order expansion of $\ln p$, valid when the posterior is narrow compared with the scale on which the prior changes.
(2) Complete the square; the constant absorbs everything independent of λ.
(3) The posterior peak moves by $\sigma_L^2 (\ln p)'$ while its width stays σ_L.

The prior's pull, measured in posterior widths, shrinks as $1/\sqrt{N}$.

The slope of the wrong prior. To use (9.86) for prior B we need the derivative of its log density, $\ln p_B(\lambda) = -\ln \lambda - (\ln \lambda - \ln 10)^2/(2 \times 0.2^2)$, at the true rate $\lambda \simeq 4$:

$$(\ln p_B)'(\lambda) \overset{(1)}{=} -\frac{1}{\lambda} - \frac{2(\ln \lambda - \ln 10)}{2 \times 0.2^2} \cdot \frac{1}{\lambda} \overset{(2)}{=} -\frac{1}{\lambda} - \frac{\ln \lambda - \ln 10}{0.04 \, \lambda},$$

$$(\ln p_B)'(4) \overset{(3)}{=} -0.25 - \frac{-0.916}{0.04 \times 4} = -0.25 + 5.73 \simeq 5.5.$$

(1) The first term is the Jacobian $1/\lambda$ of the log-normal density, whose log has derivative $-1/\lambda$. The second is a square: chain rule, with $\mathrm{d}(\ln \lambda)/\mathrm{d}\lambda = 1/\lambda$.
(2) $2 \times 0.2^2/2 = 0.2^2 = 0.04$, the prior's variance in $\ln \lambda$.
(3) $\ln 4 - \ln 10 = \ln 0.4 = -0.916$.

The Jacobian term is small; the second term dominates because $\lambda = 4$ lies $0.916/0.2 \simeq 4.6$ prior standard deviations below the prior's centre at 10. The slope is positive: the prior pulls the posterior peak upwards, towards 10. With $N = 300$, $\hat{\lambda} \simeq 3.94$, (9.86) predicts a shift of $\sqrt{3.94/300} \times 5.5 \simeq 0.63$ widths, matching the 0.621 found by the chains. For $N = 3$ the expansion is not valid (the posterior is as wide as the prior), and the chains are the only honest answer. The result is the one-parameter form of the Bernstein–von Mises theorem of Chapter 5.

Key idea: Two kinds of forgetting

A Markov chain forgets its starting point because it is ergodic, and running it longer completes the forgetting. It does not forget the prior, because the prior is part of the distribution it samples; running longer only converges more precisely to the posterior of that prior. The prior's influence fades only as the data accumulate, as $1/\sqrt{N}$ in units of the posterior width.

Two practical consequences follow. First, a convergence diagnostic such as $\hat{R}$ tests the chain, never the prior: in the experiment above $\hat{R}$ was 1.000 for the wrong prior with three counts. Checking a prior is a modelling question, answered by trying alternatives and seeing how much the posterior moves (a prior sensitivity analysis), or by importance-reweighting the existing chain to the new prior (Chapter 6) when the two posteriors overlap. Second, the experiment explains why cosmological analyses report which priors they used on poorly constrained parameters, such as the optical depth to reionisation in a CMB analysis without large-scale polarisation: there the data are "few" for that parameter, and the prior still shows in the answer.

Watch out: "Burn-in will take care of the prior"

Burn-in removes the memory of the starting point, nothing else. If two analyses of the same data with different priors disagree, no amount of running will reconcile them; only

more informative data will, and only to the extent (9.86) allows.

9.15 Back to statistical mechanics: sampling the canonical ensemble

Recall the oscillator of fixed energy (Section 9.1): a mass on a spring, $x(t) = A\cos(\omega t + \varphi)$, looked at at random instants. Its own motion did the sampling. Every point of the energy ellipse in phase space is visited once per period, the time spent near x is proportional to $1/|v|$, and the histogram of x is the arcsine law $1/(\pi\sqrt{A^2 - x^2})$ of (9.1). That is the **microcanonical** ensemble: the energy is fixed, and the states on the energy shell are weighted by the time the motion spends in them.

Now put the same oscillator in contact with a heat bath at temperature T, a large reservoir it can exchange energy with. Its energy is no longer fixed. Statistical mechanics says that the probability density of a phase-space point (x, p) is then the **canonical**, or Boltzmann, density

$$\rho(x,p) = \frac{e^{-H(x,p)/kT}}{Z}, \qquad Z = \int e^{-H(x,p)/kT}\,\mathrm{d}x\,\mathrm{d}p, \tag{9.87}$$

with H the Hamiltonian, k Boltzmann's constant and Z the partition function. In this section ρ is this Boltzmann density (not an autocorrelation) and p is the momentum (not a prior or posterior). The oscillator's own Hamiltonian motion cannot produce this density. Hamilton's equations conserve H, so a trajectory started on one energy shell stays on it forever and never visits the others. In nature the energy changes through collisions with the bath. On a computer something else must change it, and that something is the Metropolis walker. This is the problem for which Metropolis et al. [36] invented the algorithm: to compute canonical averages, an equation of state, for a system of interacting particles, where neither the partition function nor a direct way to draw configurations is available.

	microcanonical	canonical
what is fixed	energy E	temperature T (energy exchanged with a bath)
distribution	uniform on the shell $H = E$, weighted by time	$e^{-H/kT}/Z$ over all of phase space
who samples it	the system's own dynamics (if ergodic)	a Metropolis walker; physically, the bath
oscillator's x	arcsine law on $\lvert x\rvert < A$	Gaussian, variance $kT/m\omega^2$

The question. Run Metropolis on $e^{-H/kT}$. Does it reproduce what statistical mechanics predicts where we can calculate (the oscillator), and how does it behave where we cannot (a double well, and a lattice of interacting spins)?

The decision ratio of the walker is

$$R = \frac{\rho(y)}{\rho(x)} = \frac{e^{-H(y)/kT}/Z}{e^{-H(x)/kT}/Z} = e^{-[H(y)-H(x)]/kT}\,: \tag{9.88}$$

a move that lowers the energy is always taken, and a move that raises it by ΔH is taken with probability $e^{-\Delta H/kT}$. The partition function cancels exactly as the evidence did in (9.58). The two problems are the same problem in two vocabularies:

Bayesian inference	statistical mechanics
posterior $p(\boldsymbol{\theta} \mid \boldsymbol{d})$	Boltzmann density ρ
$-\ln\big[\mathcal{L}(\boldsymbol{\theta})\,p(\boldsymbol{\theta})\big]$	energy in units of kT, H/kT
evidence $Z = \int \mathcal{L}\,p\,\mathrm{d}\boldsymbol{\theta}$	partition function $Z = \int e^{-H/kT}$
tempered target f^{β} (Section 9.9)	Boltzmann weight at temperature $1/\beta$
maximum a posteriori point	ground state
posterior mean of g	thermal average $\langle g \rangle$

9.15.1 The harmonic oscillator in a heat bath

For $H = p^2/2m + m\omega^2 x^2/2$ the canonical density factorises into two Gaussians, so every prediction can be written down. (We write the spring as $m\omega^2$ to keep the letter k for Boltzmann's constant.)

$$\rho(x,p) \overset{(1)}{=} \frac{1}{Z}\exp\!\Big(-\frac{p^2}{2mkT}\Big)\exp\!\Big(-\frac{x^2}{2kT/m\omega^2}\Big),$$
$$\Big\langle \frac{p^2}{2m}\Big\rangle \overset{(2)}{=} \frac{mkT}{2m} = \frac{kT}{2}, \qquad \Big\langle \frac{m\omega^2 x^2}{2}\Big\rangle \overset{(3)}{=} \frac{m\omega^2}{2}\,\frac{kT}{m\omega^2} = \frac{kT}{2}, \qquad \langle H\rangle = kT. \tag{9.89}$$

(1) $e^{-H/kT} = e^{-p^2/2mkT}\,e^{-m\omega^2x^2/2kT}$: p and x are independent Gaussians with variances mkT and $kT/m\omega^2$.
(2) $\langle p^2\rangle$ is the variance of p, since its mean is zero.
(3) Likewise for x. Each quadratic term in H carries $kT/2$: equipartition.

The distribution of the energy itself follows from the area of phase space at each energy. The region $H \le E$ is an ellipse with semi-axes $\sqrt{2E/m\omega^2}$ in x and $\sqrt{2mE}$ in p, so its area is $\mathcal{A}(E) = \pi\sqrt{2E/m\omega^2}\sqrt{2mE} = 2\pi E/\omega$, and the area between E and $E + \mathrm{d}E$ is $\mathcal{A}'(E)\,\mathrm{d}E = (2\pi/\omega)\,\mathrm{d}E$, the same for every E: the one-dimensional oscillator has a constant density of states. Hence

$$p(E) \propto e^{-E/kT}\,\frac{2\pi}{\omega} \quad\Longrightarrow\quad p(E) = \frac{1}{kT}\,e^{-E/kT}, \qquad \mathrm{P}(E > 2kT) = e^{-2} = 0.135. \tag{9.90}$$

Canonical from microcanonical. The two ensembles are not unrelated: the canonical one is a mixture of microcanonical shells, each weighted by its Boltzmann probability (9.90). On the shell of energy E the position follows the arcsine law with amplitude $A = \sqrt{2E/m\omega^2}$. Averaging it over E should give back the Gaussian of (9.89). In units with $m\omega^2 = 1$ (so $A^2 = 2E$):

$$p(x) \overset{(1)}{=} \int_{x^2/2}^{\infty} \frac{e^{-E/kT}}{kT}\,\frac{1}{\pi\sqrt{2E - x^2}}\,\mathrm{d}E \overset{(2)}{=} \frac{e^{-x^2/2kT}}{2\pi kT}\int_0^{\infty} e^{-u/2kT}\,u^{-1/2}\,\mathrm{d}u$$
$$\overset{(3)}{=} \frac{e^{-x^2/2kT}}{2\pi kT}\,\Gamma\!\big(\tfrac12\big)\sqrt{2kT} \overset{(4)}{=} \frac{e^{-x^2/2kT}}{\sqrt{2\pi kT}}. \tag{9.91}$$

(1) Law of total probability over the energy: $p(x) = \int p(E)\,p(x \mid E)\,\mathrm{d}E$; only shells with $A > |x|$, that is $E > x^2/2$, reach x.
(2) Substitute $u = 2E - x^2$, so $E = (u + x^2)/2$ and $\mathrm{d}E = \mathrm{d}u/2$; the factor $e^{-x^2/2kT}$ comes out of the integral.
(3) $\int_0^{\infty} e^{-u/c}u^{-1/2}\mathrm{d}u = \Gamma(\tfrac12)\sqrt{c}$ with $c = 2kT$.
(4) $\Gamma(\tfrac12) = \sqrt{\pi}$, and $\sqrt{\pi}\sqrt{2kT}/(2\pi kT) = 1/\sqrt{2\pi kT}$.

The arcsine laws, each piled up at its own turning points, add up to a Gaussian peaked at the centre. The same Hamiltonian gives two different distributions of x because the sampling prescription is different: a random instant at fixed energy, or a random state at fixed temperature.

The walker in phase space. A Metropolis walker in the plane (x, p), with $m = \omega = 1$, $kT = 1$, round steps of width $2.38/\sqrt{2}$, started at $(3, -3)$ and run for 200 000 steps, accepted 0.355 of its proposals (Figure 9.36). It never uses Z and knows nothing about shells or equipartition. It gives $\mathrm{Var}\, x = 1.008$ and $\mathrm{Var}\, p = 1.008$ (exact: $kT = 1$), kinetic and potential energies $\langle p^2/2 \rangle = 0.5039$ and $\langle x^2/2 \rangle = 0.5039$ (exact: $1/2$ each), $\langle H \rangle = 1.008$, a fraction 0.1394 of states with $E > 2kT$ (exact $e^{-2} = 0.1353$), and a correlation of -0.004609 between x and p (exact 0). Drawing instead an energy from $p(E)$ and a uniform phase on that shell, as in (9.91), gives $\mathrm{Var}\, x = 0.9943$: the mixture of microcanonical ensembles is the canonical one.

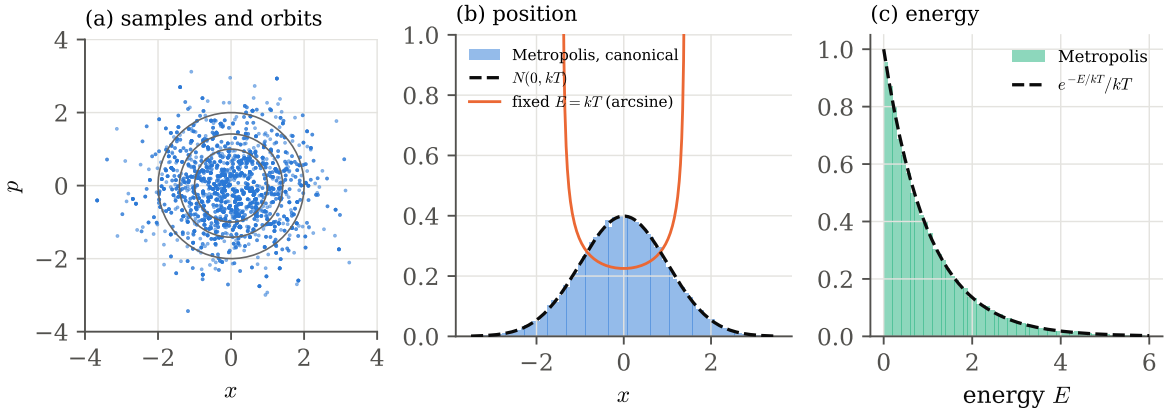


Figure 9.36: Metropolis for the oscillator at $kT = 1$. (a) The first 3000 states in phase space, with three energy shells: the walker crosses shells freely, which the Hamiltonian motion never does. (b) Position: the canonical histogram is the Gaussian $\mathcal{N}(0, kT)$ (dashed), not the arcsine law of a single shell at $E = kT$ (orange). (c) Energy: exponential, $e^{-E/kT}/kT$.

Key idea: Dynamics samples the shell, Metropolis samples the ensemble

Hamiltonian motion conserves energy, so at best it samples one microcanonical shell, uniformly in time. The canonical density $e^{-H/kT}/Z$ is a mixture of all shells weighted by $e^{-E/kT}$ times the density of states. A Metropolis walker samples it directly, because its acceptance rule $\min(1, e^{-\Delta H/kT})$ plays the part of the heat bath: it lets the energy change, and makes rises in energy exponentially rare. The partition function, like the evidence, never appears.

9.15.2 A double well: no formula, and a barrier

Replace the spring by the potential $V(x) = \Delta V\,(x^2 - 1)^2$, with minima at $x = \pm 1$ and a barrier of height $\Delta V = 1$ at $x = 0$: a particle in a symmetric double well, the shape of a molecule with two conformations or of an order parameter below a phase transition. The momentum part of $e^{-H/kT}$ is the same Gaussian as before and factorises away, so the question is the position density $\rho(x) \propto e^{-V(x)/kT}$. Its normalisation $\int e^{-V/kT} \mathrm{d}x$ has no elementary closed form. In one dimension we can still compute it by quadrature as a check; in a thousand dimensions we could not, but the walker would not notice the difference.

With 400 000 steps of width 0.15 the walker reproduces $e^{-V/kT}$ at all temperatures (Figure 9.37a): $\langle x^2 \rangle = 0.8294$ at $kT = 1$ (quadrature 0.8327) and 0.919 at $kT = 0.25$ (0.9177), with fractions 0.4953 and 0.5109 of the time in the right well (exact $\frac{1}{2}$ by symmetry). But the way it gets there changes with temperature. At $kT = 0.25$ the barrier is four times kT, and the walker spends long stretches in one well before hopping to the other (Figure 9.37b): 56.5 hops per 10^5 steps, against 395.8 at $kT = 1$.

How rare are the hops? A walker that moves in small steps behaves like a diffusing particle, and the rate at which it crosses the barrier follows the same reasoning as a thermally activated chemical reaction. To hop, the walker must first be near the top of the barrier, and must then diffuse across it. We estimate the two pieces separately and multiply.

How often is the walker at the top? Call the strip of width w around the top of the barrier, centred on $x = 0$, the top region. In equilibrium the probability of being in it is the Boltzmann weight integrated over the strip, divided by the integral over everything. The denominator is dominated by the two wells. Near a minimum, $V(x) \simeq V''(1)(x-1)^2/2$ (the minimum value is 0), so each well contributes a Gaussian integral:

$$\mathrm{P(top)} \overset{(1)}{=} \frac{\int_{\text{strip}} e^{-V(x)/kT}\,\mathrm{d}x}{\int e^{-V(x)/kT}\,\mathrm{d}x} \overset{(2)}{=} \frac{e^{-\Delta V/kT}\,w}{2Z_{\text{well}}}, \qquad Z_{\text{well}} \overset{(3)}{=} \int e^{-V''(1)(x-1)^2/2kT}\,\mathrm{d}x = \sqrt{\frac{2\pi kT}{V''(1)}}.$$

(1) The fraction of time in a region is the integral of the normalised density $\rho(x)$ over it.
(2) Inside the strip V is within about kT of its maximum ΔV, so $e^{-V/kT} \simeq e^{-\Delta V/kT}$ up to a factor of order one, and the integral is that value times the width. The two wells each contribute Z_{well}.
(3) The Gaussian integral $\int e^{-u^2/2\sigma^2}\mathrm{d}u = \sqrt{2\pi}\,\sigma$ with $\sigma^2 = kT/V''(1)$.

How wide is the top? Near $x = 0$ the potential is a downward parabola, $V(x) \simeq \Delta V - |V''(0)|\,x^2/2$. The top region is where the walker is within kT of the summit, where nothing pushes it strongly either way:

$$\frac{|V''(0)|\,x^2}{2} \le kT \quad\Longleftrightarrow\quad |x| \le \sqrt{\frac{2kT}{|V''(0)|}}, \qquad w = 2\sqrt{\frac{2kT}{|V''(0)|}}.$$

A hotter walker has a wider top, $w \propto \sqrt{kT}$.

How long does it take to cross? A Metropolis walker with steps s much smaller than the scale of V accepts nearly every step, and so moves like the drunkard of Section 9.7: its mean squared displacement after t steps is $s^2 t$. Writing that as $2Dt$ defines its diffusion constant $D = s^2/2$ per step (here D is a diffusion constant, not a dimension). Crossing a strip of width w by diffusion takes about $t_{\text{cross}} = w^2/2D$ steps, the diffusive cost $(L/s)^2$ again.

Multiply. The hop rate per step is the probability of being at the top divided by the time it takes to cross it:

$$\begin{aligned}\text{rate} &\overset{(1)}{=} \frac{\mathrm{P(top)}}{t_{\text{cross}}} = \frac{e^{-\Delta V/kT}\,w}{2Z_{\text{well}}}\cdot\frac{2D}{w^2} = \frac{D\,e^{-\Delta V/kT}}{w\,Z_{\text{well}}} \\ &\overset{(2)}{=} \frac{D\,e^{-\Delta V/kT}}{2\sqrt{2kT/|V''(0)|}\,\sqrt{2\pi kT/V''(1)}} \overset{(3)}{=} \frac{\sqrt{|V''(0)|\,V''(1)}}{4\sqrt{\pi}}\,\frac{D}{kT}\,e^{-\Delta V/kT}.\end{aligned} \tag{9.92}$$

(1) A walker in the strip leaves it, to one side or the other, after about t_{cross} steps, so the strip is crossed about $\mathrm{P(top)}/t_{\text{cross}}$ times per step.
(2) Insert w and Z_{well}.
(3) $\sqrt{kT}\times\sqrt{kT} = kT$: the two square roots, one from the width of the top and one from the width of a well, combine into a single $1/kT$; everything else is a constant of the potential and of the step size.

The prefactor $1/(4\sqrt{\pi}) \simeq 0.14$ comes from our rough choice of the strip; a careful treatment of the same problem replaces it by $1/(2\pi) \simeq 0.16$, and leaves the dependence on temperature unchanged. The result is the form of Kramers' rate for an overdamped particle [177]. Taking logarithms, $\ln(\text{rate}\times kT) = \text{const} - \Delta V/kT$, so plotted against $\Delta V/kT$ it is a straight line of slope

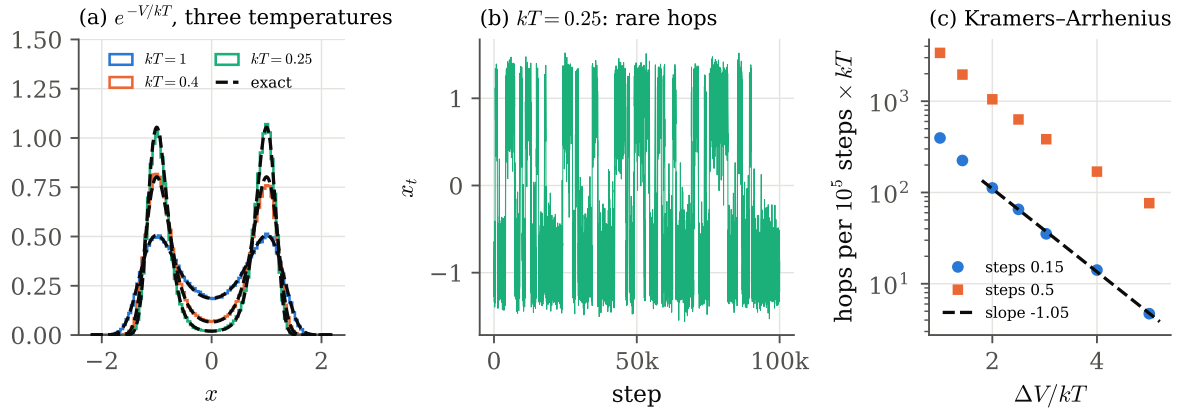


Figure 9.37: Metropolis on the double well $V = (x^2 - 1)^2$. (a) Histograms at three temperatures against $e^{-V/kT}$ normalised by quadrature (dashed). (b) At $kT = 0.25$ the walker stays in one well for long stretches. (c) Hop rate times kT against $\Delta V/kT$: small steps (dots) follow the Kramers–Arrhenius line of slope -1 (dashed); large steps (squares) hop far more often.

-1: every additional kT of barrier height makes hops $e \simeq 2.7$ times rarer. The walker gives slope -1.047 for $kT \leq 0.5$ (Figure 9.37c): Arrhenius's law, recovered from a sampling algorithm.

The squares in Figure 9.37c are walkers with steps of 0.5 instead of 0.15. They sample the same $e^{-V/kT}$ but hop far more often: 677.2 hops per 10^5 steps at $kT = 0.25$, with a fitted slope of only -0.8675. A step of that size can jump from $x \simeq -0.6$ straight to $x \simeq +0.6$, where $V = 0.4096$, without ever visiting the top of the barrier. A Metropolis walker is not a physical particle, and its steps are not physical time. That is a strength for sampling, since moves that no real particle could make may cross barriers cheaply (parallel tempering and cluster moves exploit this), and a warning for dynamics: rates and time correlations read off an MCMC run are properties of the algorithm, not of the system.

Watch out: MCMC time is not physical time

The stationary distribution of a Metropolis chain is the Boltzmann (or posterior) density; its *path* is not a trajectory of the physical system. Equilibrium averages from the chain are physical; hop rates, relaxation times and autocorrelations depend on the proposal and are diagnostics of the sampler. At low temperature the slow hops are also a sampling hazard: a chain that has not hopped often enough is the broken-ergodicity trap of Figure 9.33.

9.15.3 The Ising model: a distribution nobody can sample directly

An $L \times L$ square lattice carries a spin $s_i = \pm 1$ at each site, with periodic boundaries. Neighbouring spins prefer to align; the energy of a configuration $s = (s_1, \ldots, s_{L^2})$ is

$$E(s) = -J \sum_{\langle ij \rangle} s_i s_j, \tag{9.93}$$

where $\langle ij \rangle$ runs over nearest-neighbour pairs and $J > 0$. For $L = 32$ there are $2^{1024} \simeq 10^{308.3}$ configurations. The partition function is a sum over all of them, and no method of drawing configurations directly from $e^{-E/kT}/Z$ is known for general temperatures. A grid in configuration space is not even conceivable.

The exact answer we compare with. Yet the model has been solved exactly, in a limit: an infinite lattice in zero external field. Onsager [178] computed its partition function per spin, by writing

the sum over configurations row by row as a product of transfer matrices and finding their largest eigenvalue; from the free energy follows the energy per spin $u = \langle E\rangle/L^2$, measured in units of J. Yang [179] computed the spontaneous magnetisation, the mean spin $M = \langle s_i\rangle$ that the lattice keeps on its own when no field is applied. Both calculations run to many pages and we do not repeat them: we use their results as targets. They say that below the critical temperature

$$kT_c = \frac{2J}{\ln(1+\sqrt{2})} = 2.269\,J \tag{9.94}$$

the spins order, with magnetisation per spin $M(T) = \left[1-\sinh^{-4}(2J/kT)\right]^{1/8}$, and above it $M = 0$. This is a perfect test: a distribution with no direct sampler and an exact answer.

The single-spin-flip walker. The walker proposes to flip one spin, $s_i \to -s_i$; the proposal is symmetric (flipping it back is equally likely). Only the four bonds that touch site i change. Before the flip they contribute $-Js_ih_i$, with $h_i = \sum_{j\,\mathrm{nb}\,i} s_j$ the sum of the four neighbours; after it, $+Js_ih_i$. So

$$\Delta E = 2Js_ih_i \in \{-8J,-4J,0,4J,8J\}, \qquad a = \min\!\left(1, e^{-\Delta E/kT}\right), \tag{9.95}$$

a local computation that never needs Z. The code visits the sites in order:

Listing 9.7: One Metropolis sweep of the Ising lattice (`code/ch09/20_ising.py`)

```
def sweep(s, beta, u):
    """One sweep: visit every site once (in order) and propose to flip its spin.
    dE = 2 J s_i (sum of the 4 neighbours), J = 1.  Accept with min(1, exp(-beta dE)).
    Each single-site update satisfies detailed balance, so each preserves the Boltzmann
    law."""
    L = s.shape[0]
    acc = 0
    k = 0
    for i in range(L):
        for j in range(L):
            nb = s[(i + 1) % L, j] + s[(i - 1) % L, j] + s[i, (j + 1) % L] + s[i, (j - 1) %
    L]
            dE = 2.0 * s[i, j] * nb
            if dE <= 0.0 or u[k] < np.exp(-beta * dE):
                s[i, j] = -s[i, j]
                acc += 1
            k += 1
    return acc
```

A systematic sweep is not one reversible chain, but it does not need to be. Each single-site update satisfies detailed balance with $e^{-E/kT}$ and therefore leaves it stationary, and a sequence of updates that each leave a distribution unchanged leaves it unchanged as a whole: if $\pi P_1 = \pi$ and $\pi P_2 = \pi$, then $\pi P_1 P_2 = \pi P_2 = \pi$.

The run. On a finite lattice the sign of the magnetisation flips now and then, so we record the mean of its absolute value, $\langle|m|\rangle$ with $m = \frac{1}{L^2}\sum_i s_i$, and the mean energy per spin. The settings are

lattice	$L = 32$, 1024 spins, periodic boundaries
start	all spins up, at the lowest temperature
temperatures	twelve, from $kT = 1.5J$ to $3.5J$, denser near T_c; each starts from the last configuration of the one before (heating)
burn-in	2000 sweeps at each temperature
measurement	6000 sweeps at each temperature

A sweep is one pass of the code above, one proposed flip per site.

Side by side. Figure 9.38 shows the whole scan; at one temperature on each side of T_c the numbers are

	$kT = 1.8J$ (ordered)		$kT = 3J$ (disordered)	
	exact	ours	exact	ours
$\langle\lvert m\rvert\rangle$	0.9569	0.9566	0	0.08438
u/J	−1.859	−1.859	−0.8173	−0.8158
accepted flips		0.04152		0.4615

The snapshots show what the numbers mean: an ordered sea with a few flipped spins at $1.8J$, domains of every size at $2.3J \approx T_c$, and short-range disorder at $3J$. At $1.8J$ almost every proposed flip breaks four aligned bonds, $\Delta E = 8J$ in (9.95), and is refused.

Why they differ. *Statistical noise.* At $1.8J$ the magnetisations differ by 3×10^{-4}, and at $3J$ the energies by 1.5×10^{-3}. The energy per spin of a lattice of a thousand spins fluctuates by a few hundredths from sweep to sweep, and 6000 correlated sweeps average that down to about 10^{-3}: both differences are of the size of the run's own noise.

A magnetisation above T_c. The exact $M = 0$ is for an infinite lattice. A finite one has a net magnetisation m that fluctuates about zero, and the mean of its absolute value is positive. If the L^2 spins were independent, m would have standard deviation $1/L$ and $\langle\lvert m\rvert\rangle = \sqrt{2/\pi}/L \simeq 0.025$ for $L = 32$. At $3J$ each spin is correlated with its neighbours over a few lattice spacings, which enlarges the fluctuation to the measured 0.08438. Either way it falls like $1/L$: a lattice twice as wide would give about half.

Rounding near T_c. At T_c the infinite lattice has domains of every size; ours cannot hold a domain wider than $L = 32$. The sharp corner of $M(T)$ is therefore smeared over a range of temperatures in which the correlation length of the infinite lattice exceeds L, a few per cent of T_c for $L = 32$, visible in Figure 9.38a as the dots leaving the exact curve before T_c and not reaching zero after it.

The ordered start. A random start at low temperature would not reproduce Yang's curve. The lattice orders locally into domains of both signs, and a band of down spins that wraps around the periodic lattice can only be removed by moving its walls one spin at a time, which at $1.8J$ may not happen within the run; $\langle\lvert m\rvert\rangle$ would then come out far below 0.96. Starting ordered and heating keeps the lattice in the equilibrium state at every step of the scan.

The lesson. A walker that knows only the energy change of a single flip reproduces Onsager's and Yang's infinite-lattice answers to the noise of the run, as long as the temperature is not so close to T_c that the lattice size itself matters.

The whole scan takes seconds on a laptop: about 10^8 spin updates, compiled with `numba`. Research studies of the same model use lattices of 10^3–10^4 sites per side, many more sweeps near T_c, where the single-flip walker becomes very slow because domains of all sizes must be rearranged one spin at a time, and cluster algorithms that flip whole domains at once; they extract critical exponents from how the results change with L. Underneath, the method is the same: a biased random walk through a space far too large to enumerate, guided only by energy differences.

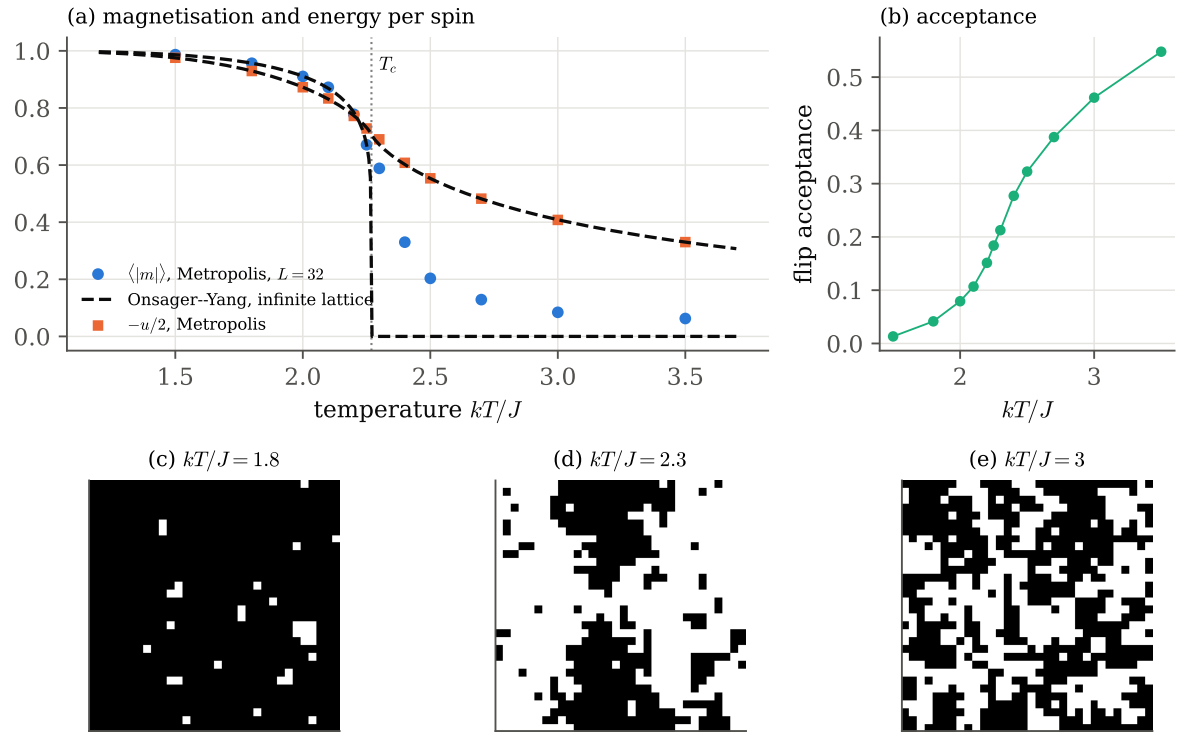


Figure 9.38: The two-dimensional Ising model on a 32×32 lattice by single-spin-flip Metropolis. (a) Mean absolute magnetisation per spin (dots) and minus half the energy per spin (squares) against temperature, with the exact infinite-lattice results of Onsager and Yang (dashed); the dotted line is T_c. (b) Fraction of accepted flips. (c)–(e) Configurations at three temperatures (black and white are the two spin values).

9.16 Relatives of the biased walker: Gibbs, Hamiltonian and ensemble samplers

Every walker so far proposed a random step and corrected it with the Metropolis–Hastings decision. Its weakness was visible on the correlated ellipse of Section 9.11.2: the step must be small enough for the narrowest direction, and along the long direction the walker then diffuses, paying the $(L/s)^2$ cost of the free walk. Three relatives keep the decision rule of (9.61), which guarantees the right stationary distribution, and change only the proposal, so as to be accepted more often or to travel further. A fourth method addresses the one quantity MCMC was designed to avoid, the evidence.

The question. Can a proposal be accepted every time? Can it travel far along the posterior and still be accepted? Can it adapt to the posterior's shape without being tuned?

9.16.1 Gibbs sampling: proposals that are always accepted

Suppose that, although we cannot draw from the joint density $f(\theta_1,\theta_2)$, we can draw each coordinate from its conditional given the other. Gibbs sampling updates the coordinates in turn, drawing θ_1 from $f(\theta_1\mid\theta_2)$ with θ_2 held at its current value, then θ_2 from $f(\theta_2\mid\theta_1)$ with the new θ_1, and repeats [Wasserman p. 416], [Kruschke §7.4.4]. It was introduced by Geman and Geman [180] for restoring images, whose pixels have simple conditionals given their neighbours, and is named after the Gibbs distributions of statistical mechanics.

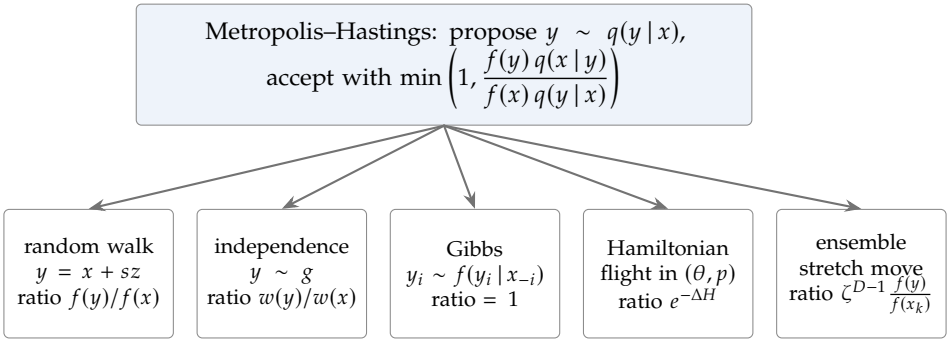


Figure 9.39: One decision rule, several proposals. Each sampler is Metropolis–Hastings with a particular way of proposing y; the ratio shown is what its Hastings ratio reduces to. $w = f/g$ is the importance weight, x_{-i} means all coordinates except the ith, ΔH is the energy error of the numerical flight, and ζ is the stretch factor of the ensemble move.

Why does it need no accept–reject step? Treat the update of θ_1 as a Metropolis–Hastings proposal $y = (y_1, x_2)$ from $x = (x_1, x_2)$ with $q(y \mid x) = f(y_1 \mid x_2)$. Its Hastings ratio is

$$R_H \overset{(1)}{=} \frac{f(y_1, x_2)\; f(x_1 \mid x_2)}{f(x_1, x_2)\; f(y_1 \mid x_2)} \overset{(2)}{=} \frac{f(y_1 \mid x_2)\, f_2(x_2)\; f(x_1 \mid x_2)}{f(x_1 \mid x_2)\, f_2(x_2)\; f(y_1 \mid x_2)} = 1. \tag{9.96}$$

(1) Definition (9.61); the reverse proposal draws x_1 from the same conditional, because the second coordinate is the same x_2 before and after.
(2) Write each joint density as conditional times marginal, $f(\theta_1, x_2) = f(\theta_1 \mid x_2)\, f_2(x_2)$; every factor cancels.

Every Gibbs proposal is accepted [HobsonBMC §3.2.5, p. 66]. When a conditional cannot be sampled directly, a single Metropolis step for that coordinate, with the others held fixed, takes its place (slice sampling, which draws a point uniformly under the graph of the conditional, is another substitute [HobsonBMC §3.2.5]); Wasserman calls this Metropolis within Gibbs [Wasserman p. 419].

Example: the correlated Gaussian. For the Gaussian with unit variances and correlation ρ of Section 9.11.2, the exponent $-(\theta_1^2 - 2\rho\theta_1\theta_2 + \theta_2^2)/2(1-\rho^2)$, viewed as a function of θ_1 alone, is $-(\theta_1 - \rho\theta_2)^2/2(1-\rho^2)$ plus terms without θ_1 (complete the square). So $\theta_1 \mid \theta_2 \sim \mathcal{N}(\rho\theta_2, 1-\rho^2)$, and symmetrically for θ_2. One sweep from $(\theta_1^{(t)}, \theta_2^{(t)})$ is

$$\begin{aligned} \theta_1^{(t+1)} &\overset{(1)}{=} \rho\, \theta_2^{(t)} + \sqrt{1-\rho^2}\; z_t \overset{(2)}{=} \rho\left[\rho\, \theta_1^{(t)} + \sqrt{1-\rho^2}\; z'_{t-1}\right] + \sqrt{1-\rho^2}\; z_t \\ &= \rho^2\, \theta_1^{(t)} + (\text{fresh noise}), \qquad \tau_{\text{Gibbs}} \overset{(3)}{=} \frac{1+\rho^2}{1-\rho^2}. \end{aligned} \tag{9.97}$$

(1) Draw from the conditional of θ_1 given the current θ_2.
(2) The current $\theta_2^{(t)}$ was itself drawn in the previous sweep from its conditional given $\theta_1^{(t)}$, with a noise z'_{t-1} independent of z_t.
(3) The sequence $\theta_1^{(t)}$ is an AR(1) chain with coefficient ρ^2, whose τ was found in Section 9.12.2.

For $\rho = 0.95$ that is 19.51 sweeps; a chain of 50 000 sweeps gives 18.73. Every proposal was accepted, and yet the chain is no faster than the best round Metropolis walker of Section 9.11.2 and slower than the shaped one. The reason is visible in Figure 9.40a: Gibbs moves only parallel to the axes, and a ridge at 45° must be climbed as a staircase of short steps. Gibbs is excellent when the coordinates are nearly independent given each other, and as slow as a random walk when they are strongly correlated [Lambert §14.6].

Gibbs shines in hierarchical models, where the conditionals have known forms even though the joint does not. Wasserman's example [Wasserman pp. 416–418, Ex. 24.11] has disease rates in k cities. The log-odds ψ_i of city i is drawn from a population $\mathcal{N}(\mu, 1)$, and what we observe is an estimate $Z_i \sim \mathcal{N}(\psi_i, \sigma_i^2)$ of it, with a known standard error σ_i; the population mean μ has a flat prior. We keep his letters; in this paragraph Z_i is the estimate for city i, not an evidence. The joint posterior of $(\mu, \psi_1, \dots, \psi_k)$ is proportional to $\prod_i \mathcal{N}(Z_i; \psi_i, \sigma_i^2)\,\mathcal{N}(\psi_i; \mu, 1)$, and Gibbs needs its conditionals.

The conditional of one city. Hold μ and the other cities fixed. Only two factors contain ψ_i, so, writing ψ for ψ_i,

$$\ln p(\psi \mid \text{rest}) \overset{(1)}{=} -\frac{(Z_i-\psi)^2}{2\sigma_i^2} - \frac{(\psi-\mu)^2}{2} + \text{const}$$
$$\overset{(2)}{=} -\frac{1}{2}\Big[\Big(\frac{1}{\sigma_i^2}+1\Big)\psi^2 - 2\Big(\frac{Z_i}{\sigma_i^2}+\mu\Big)\psi\Big] + \text{const} \overset{(3)}{=} -\frac{(\psi-e_i)^2}{2d_i^2} + \text{const},$$

(1) The log of the two Gaussian factors that contain ψ; all others are constants here.
(2) Expand the squares and keep only the terms with ψ.
(3) Complete the square. A log density $-\frac{1}{2}(A\psi^2 - 2B\psi)$ is a Gaussian with precision A and mean B/A; here $A = 1/\sigma_i^2 + 1$ and $B = Z_i/\sigma_i^2 + \mu$.

so $\psi_i \mid \text{rest} \sim \mathcal{N}(e_i, d_i^2)$ with

$$e_i = \frac{Z_i/\sigma_i^2 + \mu}{1/\sigma_i^2 + 1}, \qquad d_i^2 = \frac{1}{1/\sigma_i^2 + 1}.$$

Each city's rate is a precision-weighted average of its own data (precision $1/\sigma_i^2$) and the population mean (precision 1), the shrinkage of Chapter 5: a city with a poor estimate (large σ_i) is pulled most of the way to μ.

The conditional of the population mean. Hold all ψ_i fixed. With the flat prior, μ appears only in the k population factors:

$$\ln p(\mu \mid \text{rest}) \overset{(1)}{=} -\frac{1}{2}\sum_{i=1}^{k}(\psi_i-\mu)^2 + \text{const} \overset{(2)}{=} -\frac{1}{2}\big[k\mu^2 - 2\mu\textstyle\sum_i \psi_i\big] + \text{const} \overset{(3)}{=} -\tfrac{k}{2}(\mu-\bar\psi)^2 + \text{const},$$

(1) The k factors $\mathcal{N}(\psi_i; \mu, 1)$.
(2) Expand and keep the terms with μ.
(3) Complete the square with $A = k$ and $B = \sum_i \psi_i = k\bar\psi$, where $\bar\psi$ is the mean of the ψ_i.

so $\mu \mid \text{rest} \sim \mathcal{N}(\bar\psi, 1/k)$: the k city log-odds act as k independent measurements of μ, each with unit variance. Gibbs sampling is just cycling through these $k+1$ Gaussian draws.

9.16.2 Hamiltonian Monte Carlo: the oscillator's dynamics as a proposal

The canonical oscillator of Section 9.15 suggests a better proposal. Treat $V(\theta) = -\ln f(\theta)$ as a potential energy, as V was for the double well of Section 9.15.2, give the walker a momentum p of the same dimension (in this subsection p is a momentum, not a prior), and define $H(\theta, p) = V(\theta) + |p|^2/2$. The joint density e^{-H} is $f(\theta)$ times a standard Gaussian in p, so its θ-marginal is the target. One iteration of **Hamiltonian Monte Carlo** [181, 182] has three parts:

1. *Kick.* Draw a fresh momentum $p \sim \mathcal{N}(0, \mathbb{1})$. This is an exact draw from the conditional of p given θ (a Gibbs step), the heat bath's contribution.

2. *Flight.* Follow Hamilton's equations $\dot\theta = p$, $\dot p = -\nabla V$ for a time $t_f = n_{\text{leap}}\varepsilon$, using n_{leap} steps of size ε of the leapfrog integrator: $p \leftarrow p - \frac{\varepsilon}{2}\nabla V(\theta)$, $\theta \leftarrow \theta + \varepsilon p$, $p \leftarrow p - \frac{\varepsilon}{2}\nabla V(\theta)$. This is microcanonical motion on an energy shell, exactly what the oscillator of Section 9.1 did.
3. *Decide.* Accept the end point with probability $\min(1, e^{-\Delta H})$, where ΔH is the change of H along the numerical flight.

Exact Hamiltonian motion conserves H, so it would always be accepted; the decision only corrects the small errors of the integrator.

Why the decision is this simple. The proposal has no randomness in it once p is drawn. Write S for the whole flight, the map that takes the phase-space point $x = (\theta, p)$ to the end point after n_{leap} leapfrog steps, and $F(\theta, p) = (\theta, -p)$ for a flip of the momentum. The proposal is $y = G(x)$ with $G = F \circ S$: fly, then reverse the momentum. Two properties of G do all the work.

- *It undoes itself.* Starting from y, flying, and flipping returns to x: $G(G(x)) = x$. Each leapfrog step run with reversed momentum retraces the step it came from, so flying from the flipped end point retraces the whole path and arrives at x with reversed momentum, which the final flip restores. So the reverse proposal, from y back to x, is the same map G.
- *It keeps volumes.* Each sub-step changes p by an amount that depends only on θ, or θ by an amount that depends only on p. Such a shear has a Jacobian matrix with ones on the diagonal and zeros on one side of it, so its determinant is one; the flip has determinant ± 1. The determinant of G is a product of these, so $|\det \partial G/\partial x| = 1$.

For a deterministic proposal, detailed balance is a statement about small regions. Let A be a small region around x; the walker in A moves to $G(A)$, and the walker in $G(A)$ moves back to A. The two flows must be equal:

$$\int_A e^{-H(x)} a(x)\,\mathrm{d}x \overset{(1)}{=} \int_{G(A)} e^{-H(y)} a(y)\,\mathrm{d}y \overset{(2)}{=} \int_A e^{-H(G(x))} a(G(x)) \left|\det \frac{\partial G}{\partial x}\right| \mathrm{d}x \overset{(3)}{=} \int_A e^{-H(G(x))} a(G(x))\,\mathrm{d}x,$$

(1) Probability that leaves A for $G(A)$ equals probability that leaves $G(A)$ for A; a is the acceptance probability of the move from that point.
(2) Change variables $y = G(x)$.
(3) The flight keeps volumes.

This holds for every small A if $e^{-H(x)}a(x) = e^{-H(y)}a(y)$, which is the condition (9.63) with ratio $R = e^{-H(y)}/e^{-H(x)} = e^{-\Delta H}$, and the Metropolis choice $a = \min(1, e^{-\Delta H})$ satisfies it. No Hastings factor is left over, because the proposal is the same map in both directions and does not stretch volumes [182]. The final flip is harmless: H depends on p only through $|p|^2$, so flipping does not change e^{-H}, and the next kick draws a fresh momentum anyway.

On the correlated Gaussian with $\varepsilon = 0.15$ and $n_{\text{leap}} = 15$, the flights travel an average distance 1.646 per iteration, the energy error averages $|\Delta H| = 0.05734$, and 0.9737 of the proposals are accepted. The integrated autocorrelation time is 0.9427 iterations, below one: successive samples are slightly *anti*correlated, as for the two-state Metropolis chain of Section 9.9 (Figure 9.40b,c). The momentum carries the walker along the ridge in a straight sweep instead of a diffusion, which is what HMC buys in many dimensions. A flight of n_{leap} steps moves a distance of order $n_{\text{leap}}\varepsilon$, where a random walk of as many steps would move only $\sqrt{n_{\text{leap}}}$ times its step. For a target of D independent Gaussians, Neal [182] shows that the step size needed to keep the acceptance fixed shrinks only as $D^{-1/4}$, so an independent sample costs of order $D^{1/4}$ gradient

evaluations, against of order D target evaluations for random-walk Metropolis (Section 9.11.1). The price is the gradient $\nabla \ln f$ at every leapfrog step, and two tuning parameters, ε and n_{leap}. [HobsonBMC §3.2.5, p. 67] describes the method in the cosmological context; Stan's default sampler is an adaptive version of it.

The oscillator thus closes its loop. Its own motion sampled one energy shell; a heat bath moves it between shells with Boltzmann weights; HMC alternates the two: deterministic flights on a shell, and random momentum kicks between shells.

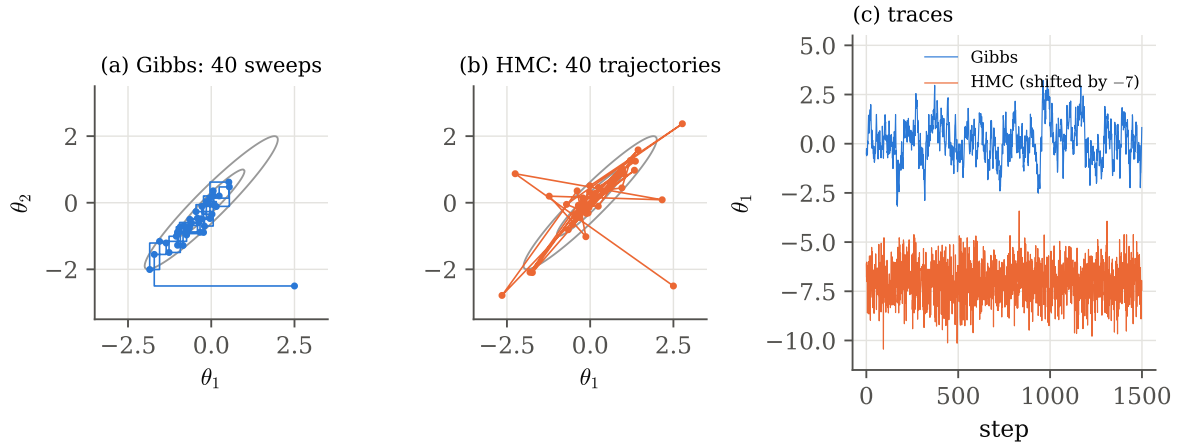


Figure 9.40: Gibbs and Hamiltonian Monte Carlo on the Gaussian with correlation 0.95, both started at $(2.5, -2.5)$. (a) Forty Gibbs sweeps: every move is parallel to an axis, so the ridge is climbed as a staircase. (b) Forty HMC iterations (end points of the flights): each jumps across a large part of the ellipse. (c) Traces of θ_1: Gibbs (top) wanders slowly; HMC (bottom, shifted down by 7) decorrelates in about one iteration.

9.16.3 Ensembles of walkers, and the evidence

A different way to avoid tuning a proposal covariance is to let many walkers propose moves for each other. In the affine-invariant ensemble sampler of Goodman and Weare [183], implemented in `emcee` [174], walker k at x_k picks another walker x_j at random and proposes the *stretch move*

$$y = x_j + \zeta\,(x_k - x_j), \qquad g(\zeta) \propto \frac{1}{\sqrt{\zeta}} \text{ on } [1/a, a], \qquad \text{accept with } \min\!\Big(1, \zeta^{D-1}\frac{f(y)}{f(x_k)}\Big), \tag{9.98}$$

with $a = 2$ by default. The candidate lies on the line through the two walkers, and the **stretch factor** ζ says how far along it: $\zeta > 1$ moves walker k away from walker j, $\zeta < 1$ towards it. (Goodman and Weare call it Z; we write ζ because Z is the evidence and the partition function elsewhere in this chapter.)

Where ζ^{D-1} comes from. Hold every other walker fixed, and measure positions from x_j. Write the present position of walker k in polar form, $x_k - x_j = r\,\omega$, with $r > 0$ its distance from x_j and ω a unit vector. The stretch keeps ω and replaces r by $r' = \zeta r$: it is a one-dimensional move in r. In polar coordinates the volume element is $r^{D-1}\mathrm{d}r\,\mathrm{d}\omega$ (the area of a sphere of radius r times its thickness, as in (9.75)), so the target, as a density in (r, ω), is $f\,r^{D-1}$. Two things are then needed. First, the proposal density of r' given r, and the reverse, which must stretch by $1/\zeta$ to come back:

$$\frac{q(r\,|\,r')}{q(r'\,|\,r)} \overset{(1)}{=} \frac{g(1/\zeta)/r'}{g(\zeta)/r} \overset{(2)}{=} \frac{g(1/\zeta)}{\zeta\,g(\zeta)} \overset{(3)}{=} \frac{\sqrt{\zeta}}{\zeta/\sqrt{\zeta}} = 1.$$

(1) $r' = \zeta r$ with $\zeta \sim g$ has density $g(r'/r)/r$ in r' (change of variable); the reverse move needs the factor $r/r' = 1/\zeta$.
(2) $r'/r = \zeta$.
(3) With $g(\zeta) \propto 1/\sqrt{\zeta}$: $g(1/\zeta) \propto \sqrt{\zeta}$ and $\zeta g(\zeta) \propto \sqrt{\zeta}$. The range $[1/a, a]$ is unchanged by $\zeta \to 1/\zeta$, so the constants agree too.

This is why g is chosen $\propto 1/\sqrt{\zeta}$: it is the choice that makes the stretch symmetric in r, so no proposal factor is left. Second, the target ratio, in the same coordinates:

$$\frac{f(y)\,(\zeta r)^{D-1}}{f(x_k)\,r^{D-1}} = \zeta^{D-1}\frac{f(y)}{f(x_k)},$$

which is the acceptance ratio of (9.98). The factor ζ^{D-1} is the volume of the larger sphere on which the stretched point sits.

Why linear correlations do not matter. Apply a linear change of variables $x \to Ax + b$ to every walker. Then $y - x_j = \zeta(x_k - x_j)$ becomes $A(y - x_j) = \zeta\, A(x_k - x_j)$: the same ζ produces the image of the same move. The transformed target differs from the original by the constant factor $1/|\det A|$, which cancels from the ratio. So every run in the new variables is the image of a run in the old ones, with the same acceptances: a correlated Gaussian is sampled exactly as efficiently as a round one, with no covariance to estimate. That is why `emcee` is so popular in astrophysics. A curved banana is not a linear transformation of anything round, and Section 9.11.3 showed `emcee` struggling on it, just like our walkers.

All these samplers share the property of Section 9.8.1: they never compute the normalisation Z of the target. That is their strength for parameter estimation and their blind spot for model comparison, where the evidence is exactly what a Bayes factor needs (Chapter 5). **Nested sampling** [39] attacks Z directly. It rewrites $Z = \int \mathcal{L}\, p\, \mathrm{d}\boldsymbol{\theta}$ as a one-dimensional integral $Z = \int_0^1 \mathcal{L}(X)\, \mathrm{d}X$ over the prior mass X enclosed by each likelihood contour, and estimates it with a set of "live" points that are repeatedly replaced by new points of higher likelihood, so that the enclosed prior mass shrinks at each replacement by a factor whose distribution is known [HobsonBMC §1.4.1, pp. 21–27]. Drawing the replacement point from the prior inside the current contour is itself usually done with a short Markov chain. Chapter 5 works through the method; the posterior samples come as a by-product.

Taxonomy: Which sampler when

random-walk Metropolis	any target you can evaluate; few parameters, or a good proposal covariance from a pilot run; no gradients needed
Metropolis–Hastings with asymmetric q	bounded parameters (log or logit steps with the Hastings factor); independence proposals from a good approximation
Gibbs	conditionals easy to sample (conjugate hierarchical models); weakly correlated blocks
Hamiltonian Monte Carlo	many parameters, smooth target, gradients available (automatic differentiation)
ensemble (`emcee`)	moderate dimension, linear degeneracies, no gradients, little tuning
tempering / nested sampling	several separated modes; the evidence Z is wanted

Every example since Section 9.8 was chosen so that the answer was known in advance: a beta posterior, a Gamma posterior, Gaussians, Onsager's solution. That made them verifications of the machinery. The machinery earns its keep where no answer is known.

How this is used in real research: MCMC in cosmological parameter estimation

Verification versus research. Every chain above checked the sampler against an exact answer. In research the posterior is unknown, and the sampler, with its diagnostics, is the instrument that produces the published constraints.
What is run. The workhorse for CMB and large-scale-structure likelihoods is still Metropolis–Hastings, as in CosmoMC [37] and its successor Cobaya: a Gaussian proposal with the covariance of a previous run, scaled by about $2.4/\sqrt{D}$; "fast" nuisance parameters (foreground amplitudes, calibrations) updated more often than "slow" cosmological ones, because changing them does not require a new Boltzmann calculation; several chains run in parallel, with $\hat{R} - 1$ of a few hundredths or less as the stopping rule. `emcee` is common in astrophysics for models with up to a few tens of parameters; HMC (through Stan, NumPyro or similar) is used when the model is differentiable; nested samplers are run when the evidence is needed. Lattice QCD runs on HMC, where it was invented [181].
Laptop versus cluster. Our chains are 10^4–10^5 steps of targets that cost microseconds. A Planck-like run evaluates a likelihood that needs a Boltzmann-code calculation of order a second per step, for 10^5 or more steps per chain and several chains: days on a cluster. The CMB analysis that follows keeps the laptop scale by replacing the Boltzmann code with a fast emulator; the statistics, the sampler and the diagnostics stay the same.

Whichever sampler produced the chain, the same three questions are put to its output. Has it forgotten where it started (burn-in and $\hat{R}$, Sections 9.12.1 and 9.13)? How many independent draws does it hold (τ and the effective sample size, Section 9.12.2)? And is the density it samples the one we meant, prior included (Section 9.14)? The samplers differ only in how they propose; none of them changes the answers to these questions.

Recap

- The unbiased random walker samples nothing; Metropolis–Hastings adds the bias. Propose $y \sim q(y \mid x)$, accept with $\min(1, R_H)$, $R_H = f(y)q(x \mid y)/[f(x)q(y \mid x)]$; a refusal repeats the current point. Without the accept–reject step it is a random walk.
- Only ratios of f enter, so the evidence of a posterior and the partition function of a Boltzmann weight cancel. Work with $\ln f$.
- Detailed balance fixes only the ratio of forward and backward acceptances; $\min(1, R_H)$ is the largest admissible choice (Peskun). Accepting with $\min(1, R^\beta)$ samples f^β. Asymmetric proposals need the Hastings factor, which is the Jacobian of the proposal (for $y = xe^{sZ}$ it is y/x); redrawing illegal proposals without it biases the chain.
- Step size: in D dimensions the best round step is $2.38/\sqrt{D}$ target widths, with acceptance ≈ 0.23 (about 0.45 in one dimension), and an independent draw every few D steps. Correlated targets need a proposal shaped by the target's covariance; curved ones need a change of variables. A good acceptance rate does not imply a good chain.
- Burn-in ends when $\ln f$ fluctuates about its equilibrium level, $\simeq \ln f_{\max} - D/2$ for a Gaussian, not at the peak. $\tau = 1 + 2\sum \rho_k$ is estimated with a window (the full sum is identically zero); ESS $= N/\tau$, MCSE $= \mathrm{SD}\sqrt{\tau/N}$; thinning never improves an average.
- Several chains from overdispersed starts: $\hat{R} = \sqrt{\hat{V}/W} \to 1$; in equilibrium $\hat{R} - 1 \simeq (1 + 1/M)\tau/2N$. $\hat{R}$ catches chains that disagree, never a mode that no chain found.

- The chain forgets its start (ergodicity); it never forgets the prior, which is part of the target. The prior's pull fades only with more data, as $1/\sqrt{N}$ posterior widths.
- Metropolis was invented for the canonical ensemble $e^{-H/kT}/Z$: it reproduces equipartition and the exponential energy law of the oscillator (a Boltzmann mixture of arcsine shells), Kramers' barrier-crossing rate in a double well, and Onsager's Ising solution on a lattice with 10^{308} states. MCMC time is not physical time.
- Relatives with the same decision rule: Gibbs (always accepted, slow on correlations), HMC (Hamiltonian flights plus momentum kicks, accepted with $e^{-\Delta H}$), affine-invariant ensembles (`emcee`); nested sampling computes the evidence that MCMC avoids.

9.17 A straight line: the chain against the exact answer

The biased random walker of Section 9.8 has so far sampled densities that we wrote down for the purpose: a Gaussian, a banana, a Boltzmann weight. Now it meets data. The first data set is the simplest one a physicist ever fits, points scattered about a straight line, and it is chosen for one reason: here the posterior is known exactly, so every number the chain produces can be checked. Once the chain has passed that test on a line, we can trust the same code on a posterior nobody can write down, the one of cosmological parameters from a CMB spectrum.

The data. We repeat the first example of Padilla et al. [Padilla §5.1][40]. A theory says that a measured quantity y depends on a control variable x through a straight line, $y = a + bx$. Two experiments measure it at 16 values of x each. Experiment D_1 has Gaussian errors of standard deviation $\sigma_1 = 0.3$, experiment D_2 has $\sigma_2 = 0.2$. The data are simulated from the line with $a = 3$, $b = 2$, at x values drawn as $3 \times \mathcal{N}(0, 1)$, which is how the analysis notebook of the repository the paper cites generates them. Our realisation is in Figure 9.41a.

The question in words. The event is the whole set of $N = 32$ measured y_i. The candidate explanations are the straight lines, one for every pair (a, b). For each candidate we ask: if this line were the truth, how probable is it that the experiments would have returned exactly these y_i? That probability, read as a function of (a, b), is the likelihood; multiplied by the prior and normalised it is the posterior, and the posterior is what the chain will sample. This is the reasoning of Chapter 5 applied to a two-dimensional space of hypotheses.

9.17.1 The likelihood from its inputs

The inputs are of two kinds. The *physical* ones: (i) each measurement is the true value of the line plus an error, $y_i = a + bx_i + \epsilon_i$; (ii) the errors of different points are independent (no common calibration error); (iii) each error is Gaussian with mean zero and a known standard deviation σ_i (0.3 or 0.2); (iv) the x_i are known exactly. The *mathematical* one is only that the prior density on (a, b) exists. Everything else follows. Given (a, b), the only random thing in y_i is ϵ_i, so y_i is Gaussian with mean $a + bx_i$ and variance σ_i^2. Then

$$p(\boldsymbol{d} \mid a, b) \overset{(1)}{=} \prod_{i=1}^{N} p(y_i \mid a, b) \overset{(2)}{=} \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\Big[-\frac{(y_i - a - bx_i)^2}{2\sigma_i^2}\Big]$$

$$\overset{(3)}{=} \left(\prod_i \frac{1}{\sqrt{2\pi}\,\sigma_i} \right) e^{-\chi^2(a,b)/2}, \qquad \chi^2(a,b) = \sum_{i=1}^{N} \frac{(y_i - a - b x_i)^2}{\sigma_i^2}. \tag{9.99}$$

(1) Input (ii): independent errors make the joint density a product.
(2) Inputs (i), (iii), (iv): y_i is the fixed number $a + bx_i$ plus a $\mathcal{N}(0, \sigma_i^2)$ error.
(3) The product of exponentials is the exponential of the sum; the prefactor does not depend on (a, b).

This is Padilla's eq. (70). In the words of Chapter 3, $e^{-\chi^2/2}$ measures how plausible is the particular noise realisation $\epsilon_i = y_i - a - bx_i$ that the candidate line needs in order to turn its prediction into the data we saw. A line that misses the points by many error bars needs an improbable noise realisation and gets a small likelihood.

The prior follows Padilla's eq. (69): flat in the box $0 < a < 5$, $0 < b < 3$, zero outside, chosen because a glance at the data excludes anything outside it. Inside the box the posterior is therefore

$$\ln p(a, b \mid \boldsymbol{d}) = -\tfrac{1}{2}\chi^2(a, b) + \text{const}, \tag{9.100}$$

and outside it $\ln p = -\infty$. This one line is all the sampler needs.

9.17.2 The exact posterior: complete the square

Because the model is linear in the parameters, χ^2 is a quadratic function of (a, b), and the exponential of minus a quadratic is a Gaussian. To find which Gaussian, write the model in matrix form. Collect the parameters in $\boldsymbol{\theta} = (a, b)^{\mathsf{T}}$, the data in $\boldsymbol{y}$, and build the $N \times 2$ **design matrix** $\mathbf{A}$ [row i is $(1, x_i)$, so $\mathbf{A}\boldsymbol{\theta}$ is the vector of the line's predictions] and the weight matrix $\mathbf{W} = \mathrm{diag}(1/\sigma_i^2)$. Then $\chi^2 = (\boldsymbol{y} - \mathbf{A}\boldsymbol{\theta})^{\mathsf{T}}\mathbf{W}(\boldsymbol{y} - \mathbf{A}\boldsymbol{\theta})$, and with $\mathbf{F} \equiv \mathbf{A}^{\mathsf{T}}\mathbf{W}\mathbf{A}$ and $\hat{\boldsymbol{\theta}} \equiv \mathbf{F}^{-1}\mathbf{A}^{\mathsf{T}}\mathbf{W}\boldsymbol{y}$,

$$\begin{aligned}
\chi^2(\boldsymbol{\theta}) &\overset{(1)}{=} \boldsymbol{y}^{\mathsf{T}}\mathbf{W}\boldsymbol{y} - 2\boldsymbol{\theta}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}}\mathbf{W}\boldsymbol{y} + \boldsymbol{\theta}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}}\mathbf{W}\mathbf{A}\boldsymbol{\theta} \\
&\overset{(2)}{=} \boldsymbol{y}^{\mathsf{T}}\mathbf{W}\boldsymbol{y} - 2\boldsymbol{\theta}^{\mathsf{T}}\mathbf{F}\hat{\boldsymbol{\theta}} + \boldsymbol{\theta}^{\mathsf{T}}\mathbf{F}\boldsymbol{\theta} \\
&\overset{(3)}{=} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^{\mathsf{T}}\mathbf{F}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \left[\boldsymbol{y}^{\mathsf{T}}\mathbf{W}\boldsymbol{y} - \hat{\boldsymbol{\theta}}^{\mathsf{T}}\mathbf{F}\hat{\boldsymbol{\theta}}\right] \overset{(4)}{=} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^{\mathsf{T}}\mathbf{F}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \chi^2_{\min}.
\end{aligned} \tag{9.101}$$

(1) Multiply out; $\boldsymbol{y}^{\mathsf{T}}\mathbf{W}\mathbf{A}\boldsymbol{\theta}$ is a number and equals its transpose $\boldsymbol{\theta}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}}\mathbf{W}\boldsymbol{y}$, so the two cross terms combine.
(2) By the definitions, $\mathbf{A}^{\mathsf{T}}\mathbf{W}\boldsymbol{y} = \mathbf{F}\hat{\boldsymbol{\theta}}$ and $\mathbf{A}^{\mathsf{T}}\mathbf{W}\mathbf{A} = \mathbf{F}$.
(3) Add and subtract $\hat{\boldsymbol{\theta}}^{\mathsf{T}}\mathbf{F}\hat{\boldsymbol{\theta}}$; $\mathbf{F}$ is symmetric, so $(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^{\mathsf{T}}\mathbf{F}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}^{\mathsf{T}}\mathbf{F}\boldsymbol{\theta} - 2\boldsymbol{\theta}^{\mathsf{T}}\mathbf{F}\hat{\boldsymbol{\theta}} + \hat{\boldsymbol{\theta}}^{\mathsf{T}}\mathbf{F}\hat{\boldsymbol{\theta}}$.
(4) The bracket does not depend on $\boldsymbol{\theta}$; setting $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ shows it is the minimum of χ^2.

So $p(\boldsymbol{\theta} \mid \boldsymbol{d}) \propto e^{-(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^{\mathsf{T}}\mathbf{F}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})/2}$: a Gaussian with mean $\hat{\boldsymbol{\theta}}$, the weighted least-squares line, and covariance $\mathbf{F}^{-1}$. (The flat prior box cuts it off, but the box edges lie dozens of standard deviations away, so the truncation is invisible.) The matrix $\mathbf{F}$ is the Fisher matrix of Chapter 10; for a linear model with Gaussian errors the Fisher forecast is not an approximation but the exact posterior covariance.

Example: the 2×2 matrix written out. Name the weighted sums $S = \sum_i \sigma_i^{-2}$, $S_x = \sum_i x_i \sigma_i^{-2}$, $S_{xx} = \sum_i x_i^2 \sigma_i^{-2}$, and $\Delta = S S_{xx} - S_x^2$. Then

$$\begin{aligned}
\mathbf{F} &= \begin{pmatrix} S & S_x \\ S_x & S_{xx} \end{pmatrix}, & \mathbf{F}^{-1} &\overset{(1)}{=} \frac{1}{\Delta}\begin{pmatrix} S_{xx} & -S_x \\ -S_x & S \end{pmatrix}, \\
\sigma_a^2 &= \frac{S_{xx}}{\Delta} \overset{(2)}{=} \frac{1}{S}\,\frac{1}{1 - \bar{x}_w^2/\overline{x^2}_w}, & \rho_{ab} &\overset{(3)}{=} \frac{-S_x}{\sqrt{S S_{xx}}} = -\frac{\bar{x}_w}{\sqrt{\overline{x^2}_w}},
\end{aligned} \tag{9.102}$$

where $\bar{x}_w = S_x/S$ and $\overline{x^2}_w = S_{xx}/S$ are the error-weighted means of x and x^2.

(1) The inverse of $\left(\begin{smallmatrix} p & q \\ q & r \end{smallmatrix}\right)$ is $\left(\begin{smallmatrix} r & -q \\ -q & p \end{smallmatrix}\right)/(pr - q^2)$.
(2) Divide numerator and denominator by SS_{xx}: $\Delta/S_{xx} = S(1 - S_x^2/SS_{xx})$.
(3) $\rho = \mathrm{Cov}(a,b)/(\sigma_a\sigma_b) = (-S_x/\Delta)/\sqrt{S_{xx}S/\Delta^2}$.

Two facts drop out. The intercept can never be known better than $1/\sqrt{S}$, the error of a weighted mean of the y_i, and it reaches that floor when the weighted mean of x is zero. Here $S = 16/0.3^2 + 16/0.2^2$, so $1/\sqrt{S} = 0.0416$, and because our x values are scattered about zero the exact width is $\sigma_a = 0.0417$, essentially the floor. And the correlation of slope and intercept is set by where the data sit relative to $x = 0$ (Section 9.17.6). For our data

$$a = 2.999 \pm 0.0417, \qquad b = 1.9810 \pm 0.0134, \qquad \rho_{ab} = -0.06,$$

and the best line has $\chi^2_{\min} = 33.0$ for $32 - 2$ degrees of freedom, a perfectly ordinary value.

9.17.3 The chain

The sampler is the random-walk Metropolis algorithm of Section 9.8, in the compact form imported by every script of the analyses that follow:

Listing 9.8: A compact random-walk Metropolis sampler (`code/ch09/lib_mcmc9c.py`)

```
def mh(logpost, x0, cov, n, rng, scale=None):
    """Random-walk Metropolis: propose y = x + L z (z ~ N(0,1), L L^T = scale^2 cov),
    accept with probability min(1, p(y)/p(x)).  The proposal is symmetric, so q cancels.

    Returns (chain, logp, acceptance fraction).  scale defaults to 2.38/sqrt(d).
    """
    x = np.array(x0, dtype=float)
    d = x.size
    s = 2.38 / np.sqrt(d) if scale is None else scale
    L = np.linalg.cholesky(np.atleast_2d(cov)) * s
    chain = np.empty((n, d))
    lp = np.empty(n)
    lx = logpost(x)
    z = rng.standard_normal((n, d)) @ L.T          # all proposal steps drawn at once
    logu = np.log(rng.random(n))                   # all accept-reject coins drawn at once
    acc = 0
    for t in range(n):
        y = x + z[t]
        ly = logpost(y)
        if logu[t] < ly - lx:                      # the biasing step: uphill always,
    downhill sometimes
            x, lx = y, ly
            acc += 1
        chain[t] = x
        lp[t] = lx
    return chain, lp, acc / n
```

It proposes $y = x + Lz$ with z a vector of independent standard Gaussians and LL^T equal to the proposal covariance times $(2.38/\sqrt{d})^2$, the optimal scaling of Section 9.11.1. Since the proposal is symmetric the Hastings factor is one, and the move is accepted when $\ln u < \ln p(y) - \ln p(x)$ for a uniform u: uphill always, downhill with probability $p(y)/p(x)$. A refusal writes the current point again. All the proposals and all the uniforms are drawn before the loop, which changes nothing in the statistics and makes the loop faster. The target is (9.100):

Listing 9.9: The log posterior of the straight line (`code/ch09/40_line_fit.py`)

```python
def logpost_K(th):
    """Known errors: ln p(a, b | d) = -chi^2/2 + const inside the flat prior box."""
    a, b = th
    if not inside(a, b):
        return -np.inf
    r = (y - a - b * x) / sig
    return -0.5 * r @ r
```

Listing 9.10: Five chains, as in Padilla et al. (`code/ch09/40_line_fit.py`)

```python
NCH, NST, BURN = 5, 10000, 1000
starts = np.c_[rng.uniform(*PRIOR["a"], NCH), rng.uniform(*PRIOR["b"], NCH)]
prop = np.diag([0.05 ** 2, 0.015 ** 2])            # a rough guess of the widths, not the answer
ch, lp, acc = mc.run_chains(logpost_K, starts, prop, NST, rng)
post = ch[:, BURN:, :].reshape(-1, 2)
m_mc, s_mc, R_mc = mc.summary(post)
R = mc.rhat(ch[:, BURN:, :])
taus = [mc.tau_int(ch[0, BURN:, k]) for k in range(2)]
ess_tot = [sum(mc.ess(ch[c, BURN:, k]) for c in range(NCH)) for k in range(2)]
```

As in the paper we run five chains of 10000 steps, started at random points of the prior box. The proposal covariance is a rough guess, $\mathrm{diag}(0.05^2, 0.015^2)$, not the answer; in a real problem it would come from a short exploratory run (Section 9.11.2). We drop the first 1000 steps of each chain as burn-in, leaving 45000 samples.

The diagnostics of Section 9.12 come first, before any number is quoted. The acceptance rate is 0.31; the Gelman–Rubin statistic is $\hat{R} = 1.0013$ for a and 1.0012 for b; the integrated autocorrelation times are $\tau = 6.7$ and 8.1 steps, so the five chains hold about 6326 and 5658 effective samples. The Monte Carlo error of the posterior mean of b is then $\sigma_b\sqrt{\tau/N} = 0.00018$, a hundred times smaller than the posterior width. The chain can be trusted, and it reports

$$a = 2.999 \pm 0.0416, \qquad b = 1.9811 \pm 0.0136, \qquad \rho_{ab} = -0.05,$$

against the exact 2.999 ± 0.0417, 1.9810 ± 0.0134, -0.06. The library sampler `emcee` [174], run on the same log posterior, gives $a = 2.998 \pm 0.0422$ and $b = 1.9804 \pm 0.0129$. In Figure 9.41b the chain's 68 and 95 per cent contours lie on top of the exact ellipses.

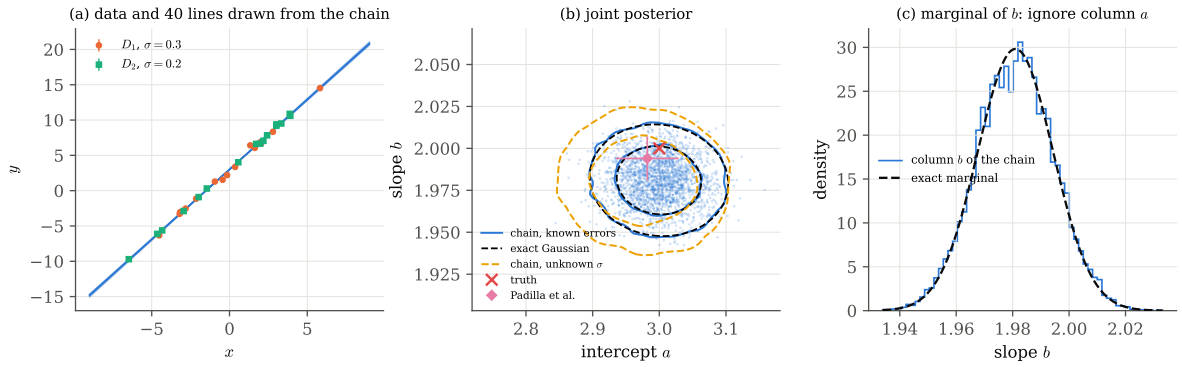


Figure 9.41: The straight line. (a) The two data sets and 40 lines whose (a, b) were picked at random from the chain; they all pass through the bulk of the points. (b) The joint posterior: chain samples (dots) with their 68 and 95 per cent contours (solid), the exact Gaussian (black dashed), the posterior when the noise level is an unknown parameter (orange dashed), the truth (cross) and the result of Padilla et al. (diamond with error bars). (c) The histogram of the b column of the chain against the exact marginal density of b.

9.17.4 Marginalising by ignoring a column

The chain is a table with one row per step and one column per parameter. Suppose we care only about the slope. Its marginal posterior is $p(b \mid \boldsymbol{d}) = \int p(a, b \mid \boldsymbol{d})\, \mathrm{d}a$: the joint density with the nuisance parameter a averaged out, exactly the operation that Chapter 1 introduced to isolate one variable of a joint distribution. With a chain we never do that integral. We take the b column, throw the a column away, and histogram what is left (Figure 9.41c, Figure 9.42).

Key idea: A column of the chain is a sample of the marginal

If the rows (a_t, b_t) are draws from $p(a, b \mid \boldsymbol{d})$, the numbers b_t alone are draws from $p(b \mid \boldsymbol{d})$. Marginalising over any nuisance parameter costs nothing: drop its column.

Why is this true? Take any interval B of slopes and ask how often a row of the chain has its slope in B, whatever its intercept:

$$\mathrm{P}(b_t \in B) \overset{(1)}{=} \int\!\!\int \mathbb{1}_B(b)\, p(a, b \mid \boldsymbol{d})\, \mathrm{d}a\, \mathrm{d}b \overset{(2)}{=} \int_B \Big[\int p(a, b \mid \boldsymbol{d})\, \mathrm{d}a \Big] \mathrm{d}b \overset{(3)}{=} \int_B p(b \mid \boldsymbol{d})\, \mathrm{d}b. \tag{9.103}$$

(1) The row (a_t, b_t) is distributed as the joint posterior, and the event "$b_t \in B$" puts no condition on a_t: its indicator $\mathbb{1}_B(b)$ is one for every a when $b \in B$.
(2) The indicator restricts the b integral to B; the a integral is done first.
(3) The inner integral is the marginal density, by definition.

A histogram counts rows with b_t in each bin, so by (9.103) and the ergodic theorem (Section 9.6) it converges to the marginal. The column does the integral for us because the rows with large $|a - \hat{a}|$ already appear less often, in proportion to their posterior weight. The same holds in any dimension: from a chain in 6 or 30 parameters, every one- and two-dimensional marginal is a histogram of one or two columns. This is how the marginals of Figure 9.41, and all the triangle plots that follow, are made [HobsonBMC p. 71].

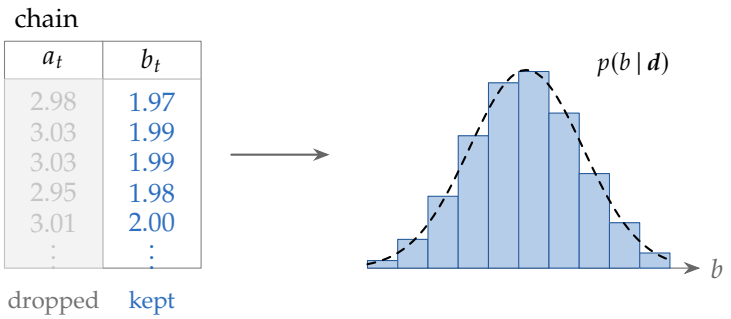


Figure 9.42: Marginalising with a chain. Each row of the chain is a draw from the joint posterior. Dropping the column of the nuisance parameter a and histogramming the column b gives the marginal density of b (dashed), with a integrated out.

Example: marginal is not conditional. It is tempting to report the width of b "at the best a", that is, the width of the slice $p(b \mid a = \hat{a}, \boldsymbol{d})$. Is that the same number? (9.101) answers it. Hold a fixed in the quadratic form $(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^{\mathsf{T}} \mathbf{F} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$; what is left is a quadratic in b whose b^2 coefficient is $F_{bb} = S_{xx}$, and completing the square in b once more turns it into $S_{xx}(b - b_0)^2$ plus terms free of b. So the slice is a Gaussian of variance $1/S_{xx}$, whatever the value at which a was fixed. The marginal variance is the bb element of $\mathbf{F}^{-1}$ in (9.102), $\sigma_b^2 = S/\Delta$. To compare the two,

multiply the marginal by $1-\rho^2$:

$$\sigma_b^2(1-\rho^2) \overset{(1)}{=} \frac{S}{\Delta}\Big(1-\frac{S_x^2}{SS_{xx}}\Big) \overset{(2)}{=} \frac{S}{\Delta}\,\frac{SS_{xx}-S_x^2}{SS_{xx}} \overset{(3)}{=} \frac{1}{S_{xx}}. \tag{9.104}$$

(1) $\sigma_b^2 = S/\Delta$ and $\rho^2 = S_x^2/(SS_{xx})$, both from (9.102).
(2) Common denominator.
(3) The numerator is Δ by its definition, and it cancels.

So the slice has standard deviation $\sigma_b\sqrt{1-\rho^2}$, while the marginal has σ_b: fixing a forgets the uncertainty in a, which, through the correlation, would have spread b. With our data $\rho = -0.06$ and the two agree to a fraction of a per cent. With the same data plotted against $x+10$ (next subsection) the correlation becomes $\rho = -0.955$, so $1-\rho^2 \approx 0.088$ and the slice is $\sqrt{1-\rho^2} \approx 0.30$ times the marginal width: reporting it would claim the slope three times better than the data know it. The column histogram cannot make this mistake, because it never fixes a.

Example: the lines the posterior believes. Each row of the chain is a whole line, so a random selection of rows draws the family of lines compatible with the data (Figure 9.41a; Kruschke calls these the credible regression lines [Kruschke §17.2.4]). At a new x the predicted mean $a+bx$ has variance

$$\mathrm{Var}(a+bx) = \sigma_a^2 + 2x\ \mathrm{Cov}(a,b) + x^2\sigma_b^2,$$

smallest near the weighted centre of the data and growing quadratically away from it. With our numbers the band is ± 0.0417 wide at $x=0$ and about $\sqrt{0.042^2 + 10^2 \times 0.013^2} \approx 0.14$ at $x=10$, outside the measured range: the lines fan out where there are no data to pin them. The chain gives this without any error propagation: compute $a_t + b_t x$ row by row and take the spread.

9.17.5 Reproducing Padilla et al.: why their error bars are wider

Their question. Padilla et al. use this straight line as the first test of a sampler written for cosmology [Padilla §5.1]. They wanted to know whether a short Metropolis run recovers the two parameters of a line measured by two experiments, and what the error bars it reports mean. The first half of the question has the answer "yes" in their paper and in ours. The second half is the interesting one, because their error bars turn out to answer a different question from ours.

Their setup, one item at a time. *The data:* two sets of 16 points from the line $y = 3+2x$, at x values drawn as $3 \times \mathcal{N}(0,1)$, one set with Gaussian errors of 0.3 and one with 0.2. This is our experiment exactly, but a different random realisation of it. *The prior:* flat in the box $0 < a < 5$, $0 < b < 3$, as ours. *The noise:* here the two analyses part. The paper's text gives the two error bars, but the analysis notebook in the repository the paper cites does not use them. It describes the noise of all 32 points by one common standard deviation σ, which it treats as unknown, with a flat prior on $0 < \sigma < 1.5$. So σ is a third parameter: the chain samples (a, b, σ), and the numbers reported for a and b come from the first two columns.

Their key equation: the line with the noise level integrated out. In our notation the change is one substitution. Set every σ_i of (9.99) equal to the one unknown σ. The exponent becomes $-\mathrm{RSS}(a,b)/2\sigma^2$, with the **residual sum of squares** $\mathrm{RSS} = \sum_i (y_i - a - bx_i)^2$ [the χ^2 of the line with all errors set to one], and the prefactor $\prod_i (\sqrt{2\pi}\,\sigma_i)^{-1}$ becomes $(2\pi)^{-N/2}\sigma^{-N}$. That prefactor could be dropped when the σ_i were known; now it depends on a parameter and must stay. With the flat priors, $\ln p(a,b,\sigma \mid \boldsymbol{d}) = -N\ln\sigma - \mathrm{RSS}(a,b)/2\sigma^2 + \text{const}$. The noise level is a nuisance

parameter, and integrating it out (the upper prior limit is far out in the tail and can be sent to infinity) gives

$$
\begin{aligned}
p(a,b\mid \boldsymbol{d}) &\overset{(1)}{=} \int_0^\infty \sigma^{-N} e^{-\mathrm{RSS}/2\sigma^2}\,\mathrm{d}\sigma \overset{(2)}{=} \frac{1}{2}\Big(\frac{2}{\mathrm{RSS}}\Big)^{(N-1)/2} \int_0^\infty u^{(N-3)/2} e^{-u}\,\mathrm{d}u \\
&\overset{(3)}{=} \frac{1}{2}\,\Gamma\Big(\frac{N-1}{2}\Big)\Big(\frac{2}{\mathrm{RSS}}\Big)^{(N-1)/2} \;\propto\; \mathrm{RSS}(a,b)^{-(N-1)/2} \\
&\overset{(4)}{=} \Big[1+\frac{(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_0)^{\mathsf{T}}\mathbf{A}^{\mathsf{T}}\mathbf{A}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_0)}{\mathrm{RSS}_{\min}}\Big]^{-(\nu+2)/2}, \qquad \nu = N-3.
\end{aligned}
\tag{9.105}
$$

(1) Flat prior in σ; dropped constants. Marginalising σ is integrating it out.

(2) Substitute $u = \mathrm{RSS}/2\sigma^2$: $\sigma = (\mathrm{RSS}/2u)^{1/2}$, $\mathrm{d}\sigma = -\frac{1}{2}(\mathrm{RSS}/2)^{1/2}u^{-3/2}\mathrm{d}u$, and $\sigma^{-N} = (2u/\mathrm{RSS})^{N/2}$; the powers of u combine to $u^{(N-3)/2}$, and the minus sign is absorbed by swapping the limits.

(3) $\int_0^\infty u^{s-1}e^{-u}\mathrm{d}u = \Gamma(s)$ with $s = (N-1)/2$.

(4) (9.101) with $\mathbf{W} = \mathbb{1}$: $\mathrm{RSS} = \mathrm{RSS}_{\min} + (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_0)^{\mathsf{T}}\mathbf{A}^{\mathsf{T}}\mathbf{A}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_0)$, $\hat{\boldsymbol{\theta}}_0$ the ordinary least-squares line; divide by $\mathrm{RSS}_{\min}$ and write $(N-1)/2 = (\nu+2)/2$.

The last line is a two-dimensional Student t density with $\nu = 29$ degrees of freedom (Chapter 2), centred on the ordinary least-squares line, with covariance $\mathrm{RSS}_{\min}(\mathbf{A}^{\mathsf{T}}\mathbf{A})^{-1}/(\nu-2)$. Set it beside the known-error posterior of Section 9.17.2. There the covariance was $(\mathbf{A}^{\mathsf{T}}\mathbf{W}\mathbf{A})^{-1}$, which for one known noise level σ ($\mathbf{W} = \mathbb{1}/\sigma^2$) is $\sigma^2(\mathbf{A}^{\mathsf{T}}\mathbf{A})^{-1}$. The symbol map is therefore short: their σ^2 is not given but estimated, and in the covariance it is replaced by $\mathrm{RSS}_{\min}/(\nu-2)$, the scatter the data actually show about the best line; the Gaussian becomes a Student t, with heavier tails, because that estimate is itself uncertain. The widths now depend on the data through $\mathrm{RSS}_{\min}$. The common σ must absorb the larger scatter of D_1, and the data put it at $\sqrt{\mathrm{RSS}_{\min}/(N-2)} = 0.276$, above the 0.2 of the better experiment.

What we compute. Two posteriors for our realisation: the known-error one of Section 9.17.3, and the unknown-σ one, sampled by a chain in (a,b,σ) whose σ column is then ignored (Section 9.17.4) and checked against the exact Student t of (9.105). The chain gives $a = 2.976 \pm 0.0516$, $b = 1.9805 \pm 0.0171$ (acceptance 0.37, largest $\hat{R} = 1.0013$), against the exact 2.977 ± 0.0514 and 1.9805 ± 0.0172; the data say $\sigma = 0.287 \pm 0.038$. Because their realisation differs from ours, one data set cannot settle which model they used, so we also simulate 20000 replications of the whole design, new x_i and new y_i each time, and record the widths each model would report.

Side by side. Padilla et al. report $a = 2.982 \pm 0.047$ and $b = 1.994 \pm 0.013$ [Padilla §5.1].

	a	b	noise model
Padilla et al. [Padilla §5.1]	2.982 ± 0.047	1.994 ± 0.013	as published
ours, chain, known σ_i	2.999 ± 0.0416	1.9811 ± 0.0136	0.3 and 0.2 known
ours, exact, known σ_i	2.999 ± 0.0417	1.9810 ± 0.0134	(9.101)
ours, chain, one unknown σ	2.976 ± 0.0516	1.9805 ± 0.0171	σ column ignored
ours, exact, one unknown σ	2.977 ± 0.0514	1.9805 ± 0.0172	Student t (9.105)
typical width, unknown σ	0.0412–0.0545	0.0133–0.0193	20000 replications

Why they differ. *The centres.* Their data are a different random realisation of the same experiment, so their centres must differ from ours by amounts of order one standard deviation, and they do.

The widths cannot come from the known errors. With known errors the posterior covariance $\mathbf{F}^{-1}$ depends on the x_i and σ_i only, not on the y_i (9.102), and σ_a can exceed its floor $1/\sqrt{S}$ only through the factor $(1-\bar{x}_w^2/\overline{x^2}_w)^{-1/2}$. How large would that factor have to be? Write $r^2 = \bar{x}_w^2/\overline{x^2}_w$ for the squared ratio of the weighted mean of the x_i to their weighted root-mean-square. Then

$$\frac{0.047}{0.0416} \approx 1.13 \overset{(1)}{=} \frac{1}{1-r^2} = 1.13^2 \approx 1.28 \overset{(2)}{=} r^2 \approx 0.22, \qquad r \approx 0.47. \tag{9.106}$$

(1) The floor is $1/\sqrt{S}$ with $S = 16/0.3^2 + 16/0.2^2 = 578$, so $1/\sqrt{S} = 0.0416$; by (9.102) the ratio $\sigma_a\sqrt{S}$ is $(1-r^2)^{-1/2}$. Square it.
(2) Solve: $r^2 = 1 - 1/1.28$.

So the weighted mean of the x_i would have to be about half their root-mean-square. For 32 points drawn from $3 \times \mathcal{N}(0,1)$ the root-mean-square is about 3, while the mean scatters about zero with standard deviation about $3/\sqrt{32} \approx 0.53$; a typical r is therefore $0.53/3 \approx 0.18$, and 0.47 is more than two and a half times that. The replications confirm it: the known-error width σ_a has median 0.0420 with 68 per cent between 0.0416 and 0.0432, and 99 per cent of replications give less than Padilla's 0.047. Their posterior is not the posterior of (9.100).

The widths do come from one unknown σ. Over the same 20000 replications the unknown-σ model gives σ_a between 0.0412 and 0.0545 (68 per cent) and σ_b between 0.0133 and 0.0193; Padilla's widths sit at the 46th and 13th percentiles, perfectly ordinary.[4]

The lesson. It is not about this paper. Two analyses of the same kind of data can both be correct and quote different error bars because they put different hypotheses in the likelihood: "the errors are 0.3 and 0.2" is a stronger statement than "there is some common noise level", and the stronger statement buys tighter constraints. When reproducing a published number, find the model first.

9.17.6 Centring the predictor: the same posterior, a different chain

Kruschke points out that a regression chain can become very slow for a reason that has nothing to do with the physics: the arbitrary position of $x = 0$ [Kruschke §17.2.1.1]. Move the origin of x far from the data, and every line that goes through the data has to trade its slope against its intercept, because tilting a line about the data cloud moves its value at the distant $x = 0$ by a lot. Equation (9.102) puts a number on it. With x shifted by s, the weighted means become $\bar{x}_w = s$ and $\overline{x^2}_w = s^2 + v$, where $v = 2.99^2$ is the variance of our x values, so

$$\rho_{ab} = -\frac{s}{\sqrt{s^2+v}} = -\frac{10}{\sqrt{100+2.99^2}} \approx -0.96 \quad \text{for } s = 10,$$

and the intercept width grows from 0.0416 to 0.141. The simulation, with the same data plotted against $x - \bar{x} + 10$, gives $\rho = -0.955$.

The posterior is the same physical statement in both coordinates: the intercept at the new origin is the old one minus s times the slope, a linear change of variables [Kruschke eq. 17.2]. The chain is not the same. Run both with the same round proposal, of radius equal to the centred intercept width: the centred chain accepts 0.30 of its moves and has $\tau = 5.7$ steps; the shifted one accepts 0.12 and has $\tau = 342$ steps, a factor of about 60 in run time for the same precision (Figure 9.43). This is the narrow-diagonal problem of Section 9.11.2: a round proposal on a

[4] The paper also says the chains were run "with temperature $T = 2$". Read literally, a tempered chain samples $\mathcal{L}^{1/2}$ and would widen the known-error posterior by $\sqrt{2}$, to about 0.059 and 0.019; neither those numbers nor the known-error ones match the published widths, while the unknown-σ model does.

thin tilted ellipse must take steps of the ellipse's short width and then diffuses slowly along its length. There are two cures, and both are standard. Reparametrise so that the posterior is round (centre or standardise x, as Kruschke does, then transform back with eq. 17.2), or keep the parameters and give the proposal the shape of the posterior covariance, which is what CosmoMC does with the covariance of a previous run [37].

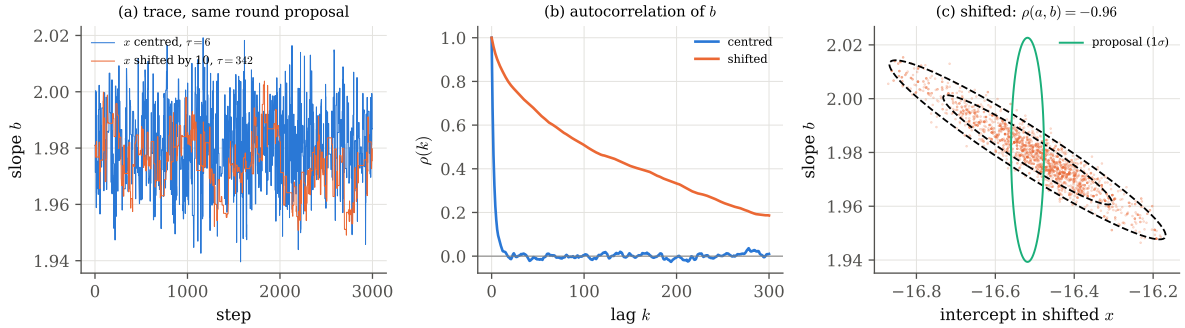


Figure 9.43: The same straight-line posterior sampled with the same round proposal in two coordinate systems. (a) Traces of the slope: with x centred the chain jumps freely; with the origin of x moved ten units from the data it creeps. (b) Their autocorrelation functions. (c) In shifted coordinates the posterior (ellipse, 68 and 95 per cent) is a thin diagonal ridge, and the proposal (green circle, one standard deviation) can only take steps of its short width.

9.18 A decaying source: curved posteriors and nuisance parameters

For the straight-line fit, a Gaussian likelihood and a model linear in its parameters gave a Gaussian posterior under the prior we used. In a decay model the unknown time scale enters inside an exponential, as in radioactive decay, cooling, capacitor charging and models for portions of a supernova light curve. The posterior then need not have a closed form. A Gaussian fit at the peak describes only the local shape, so we use the chain to explore the rest of the posterior and to account for uncertainty in the noise level.

The data. A source is read once an hour for fifteen hours. The count rate falls as $y(t) = Ae^{-t/\tau}$, and each reading has a Gaussian error of standard deviation σ. We simulate $N = 15$ readings at $t = 0, 1, \ldots, 14\,\mathrm{h}$ with $A = 100\,\mathrm{s}^{-1}$, $\tau = 4\,\mathrm{h}$ and $\sigma = 6\,\mathrm{s}^{-1}$ (Figure 9.44a). The prior is flat in $0 < A < 300\,\mathrm{s}^{-1}$ and $0.1 < \tau < 30\,\mathrm{h}$. With σ known, the derivation of (9.99) goes through unchanged with $a + bx_i$ replaced by $Ae^{-t_i/\tau}$:

$$\ln p(A, \tau \mid \boldsymbol{d}) = -\frac{\chi_0^2(A, \tau)}{2\sigma^2} + \text{const}, \qquad \chi_0^2(A, \tau) = \sum_{i=1}^{N} \left(y_i - Ae^{-t_i/\tau}\right)^2, \tag{9.107}$$

where χ_0^2 is the residual sum of squares, not yet divided by σ^2.

9.18.1 How far from Gaussian? The Laplace approximation and the chain

The question. Is a Gaussian centred at the best fit a good description of this posterior, and if not, does the chain notice? The Gaussian in question is the **Laplace approximation** [the Gaussian whose log matches $\ln p$ to second order at its peak]. It comes from a Taylor series of

$\ln p$ about the maximum $\hat{\boldsymbol{\theta}}$:

$$\ln p(\boldsymbol{\theta}) \overset{(1)}{=} \ln p(\hat{\boldsymbol{\theta}}) + (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^{\mathsf{T}} \nabla \ln p\big|_{\hat{\theta}} + \tfrac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^{\mathsf{T}} \left[\nabla\nabla^{\mathsf{T}} \ln p\right]_{\hat{\theta}} (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) + \dots$$
$$\overset{(2)}{=} \ln p(\hat{\boldsymbol{\theta}}) - \tfrac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^{\mathsf{T}} \mathbf{H} (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) + \dots, \qquad H_{ij} = -\frac{\partial^2 \ln p}{\partial\theta_i \partial\theta_j}\Big|_{\hat{\theta}}. \tag{9.108}$$

(1) Taylor's theorem in two variables.

(2) At a maximum the gradient vanishes. Defining $\mathbf{H}$ with a minus sign makes it positive definite there.

Dropping the dots leaves a Gaussian with mean $\hat{\boldsymbol{\theta}}$ and covariance $\mathbf{H}^{-1}$, the same object as $\mathbf{F}^{-1}$ in (9.101). For the line the dots were exactly zero. Here we find the peak with a simplex optimiser and $\mathbf{H}$ by finite differences: $A = 97.0 \pm 4.7$, $\tau = 4.31 \pm 0.34\,\mathrm{h}$, $\rho = -0.63$. Four chains of 20000 steps, with the Laplace covariance as the proposal shape, give $A = 96.8 \pm 4.7$, $\tau = 4.35 \pm 0.34\,\mathrm{h}$, $\rho = -0.63$ (acceptance 0.36). The widths agree with Laplace. The centre of τ does not: the posterior mean is above the peak. On a 400×400 grid, where the exact posterior can be normalised and its A direction summed out, the marginal of τ has mean 4.35 h, standard deviation 0.34 h and skewness 0.26; the chain's column of τ has skewness 0.32 (Figure 9.44c). The joint contours are bent, a banana in miniature (Figure 9.44b).

Example: why the tail is on the long-τ side. At the time $t = 4\,\mathrm{h}$ the model predicts $100e^{-1} = 36.8\,\mathrm{s}^{-1}$ for $\tau = 4\,\mathrm{h}$. Lower τ by one hour and the prediction falls to $100e^{-4/3} = 26.4$, a change of 10.4; raise it by one hour and it rises to $100e^{-4/5} = 44.9$, a change of only 8.1. The data resist a shorter decay time more strongly than a longer one, so the likelihood falls faster on the short side and leaves a longer tail on the long side. The reason is that τ enters as $1/\tau$ in the exponent, and $1/\tau$ compresses large values. A different parameter choice, the decay rate $\lambda = 1/\tau$, would give a more symmetric posterior, but then a flat prior in λ is a different prior from a flat prior in τ (the Jacobian $|\mathrm{d}\tau/\mathrm{d}\lambda| = 1/\lambda^2$ of Chapter 2), and the question "which is flat?" must be answered with the physics, not with convenience.

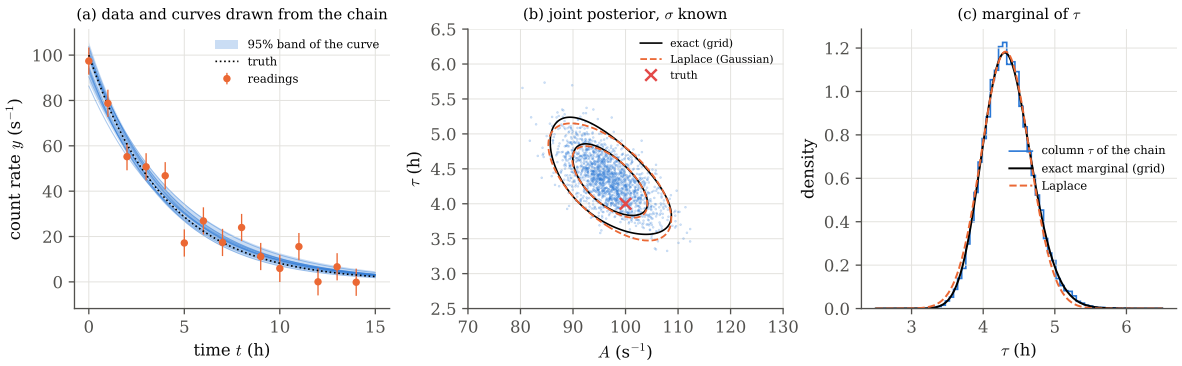


Figure 9.44: The decaying source with known noise. (a) The fifteen readings, the true curve (dotted), fifteen curves whose (A, τ) were drawn from the chain, and the band containing 95 per cent of the drawn curves at each time. (b) The joint posterior: chain samples, the exact contours computed on a grid (black), and the Laplace ellipses (dashed), which miss the bend. (c) The marginal of τ: the chain's column, the exact marginal and the Laplace Gaussian, which puts the centre too low and misses the tail at large τ.

9.18.2 The unknown noise level: three ways to treat a nuisance

So far σ was given. In a real measurement it is often known only roughly, and it matters: by (9.107) every posterior width is proportional to it. There are three things one can do with it.

Fix it at a guess. If the guess is wrong, every width is wrong by the same factor. With $\sigma = 3$ instead of 6, the marginal of τ has standard deviation 0.17 h instead of 0.34, and nothing in the output signals the problem (Figure 9.45a, dashed).

Integrate it out analytically. Hobson et al. do this for a model with the same structure [HobsonBMC eqs. 2.25–2.26]. With the prior $p(\sigma) \propto 1/\sigma$ (flat in $\ln\sigma$, the Jeffreys prior for a scale, Chapter 5),

$$p(A,\tau \mid \boldsymbol{d}) \overset{(1)}{=} \int_0^\infty (2\pi\sigma^2)^{-N/2} e^{-\chi_0^2/2\sigma^2} \frac{\mathrm{d}\sigma}{\sigma} \overset{(2)}{=} (2\pi)^{-N/2} \frac{1}{2}\Big(\frac{2}{\chi_0^2}\Big)^{N/2} \int_0^\infty u^{N/2-1} e^{-u}\,\mathrm{d}u$$
$$\overset{(3)}{=} \frac{\Gamma(N/2)}{2\,\pi^{N/2}} \left[\chi_0^2(A,\tau)\right]^{-N/2}. \tag{9.109}$$

(1) The likelihood (9.107) with its normalisation $(2\pi\sigma^2)^{-N/2}$ restored, because it depends on σ, times the prior $1/\sigma$ (its normalisation over a finite range of σ is a constant and is dropped).
(2) The substitution $u = \chi_0^2/2\sigma^2$ of (9.105), now with one more power of $1/\sigma$ from the prior: $\sigma^{-N-1}\mathrm{d}\sigma = \frac{1}{2}(2/\chi_0^2)^{N/2} u^{N/2-1}\mathrm{d}u$.
(3) The Gamma integral with $s = N/2$; $(2\pi)^{-N/2}2^{N/2} = \pi^{-N/2}$.

The shape $[\chi_0^2]^{-N/2}$ is exactly Hobson's eq. (2.26); their prefactor is the large-N form of ours, from Stirling's formula for $\Gamma(N/2)$. Compare with the line: a flat prior on σ gave the power $-(N-1)/2$ in (9.105), the scale prior $1/\sigma$ gives $-N/2$. Either way the unknown noise turns $e^{-\chi^2/2}$ into a power law in χ^2, which has heavier tails: the data must now also estimate their own scatter, and every parameter is a little less certain. On the grid the marginal of τ from (9.109) has mean 4.36 h and standard deviation 0.39 h, wider than the 0.34 h of the known-noise case.

Sample it and ignore its column. Add $\ln\sigma$ to the chain as a third parameter with a flat prior (the same $1/\sigma$ prior, written in the variable in which it is flat) and run:

Listing 9.11: Three parameters, or σ integrated out (`code/ch09/41_exponential.py`)

```
def lp_three(th):
    """(A, tau, ln sigma): flat in ln sigma is the Jeffreys prior 1/sigma."""
    A, tau, ls = th
    if not (-5 < ls < 5):
        return -np.inf
    base = lp_known((A, tau), 1.0)                # -chi0^2 / 2
    if not np.isfinite(base):
        return -np.inf
    return -N * ls + base * np.exp(-2 * ls)


def lp_marg(th):
    """sigma integrated out analytically: (N/2) ln chi0^2 with a minus sign."""
    A, tau = th
    if not (PRIOR["A"][0] < A < PRIOR["A"][1] and PRIOR["tau"][0] < tau < PRIOR["tau"][1]):
        return -np.inf
    return -0.5 * N * np.log(chi2_0(A, tau))
```

The function `lp_three` is $-N\ln\sigma - \chi_0^2/2\sigma^2$ in the variable $\ln\sigma$, and `lp_marg` is (9.109). By the key idea of Section 9.17.4, the τ column of the three-parameter chain is a sample of $p(\tau \mid \boldsymbol{d})$ with σ integrated out, so it must agree with the grid built on (9.109). It does: mean 4.36 h, standard deviation 0.39 h (four chains, acceptance 0.33, largest $\hat{R} = 1.0009$; Figure 9.45a). The analytic integral and the ignored column are two routes to the same marginal; the second needs no algebra and works for any prior.

The column we ignored is not useless either: it is the posterior of the noise level (Figure 9.45b), $\sigma = 6.4$ with 68 per cent between 5.4 and 8.0 s^{-1}. That is the honest statement of how well

fifteen readings determine their own scatter, and its width is the reason the marginal of τ widened. Hobson et al. use the same device for error bars that may be too small, scaling every σ_i by one unknown factor and marginalising it [HobsonBMC p. 52]; Padilla's notebook did exactly that for the straight line of Section 9.17.5.

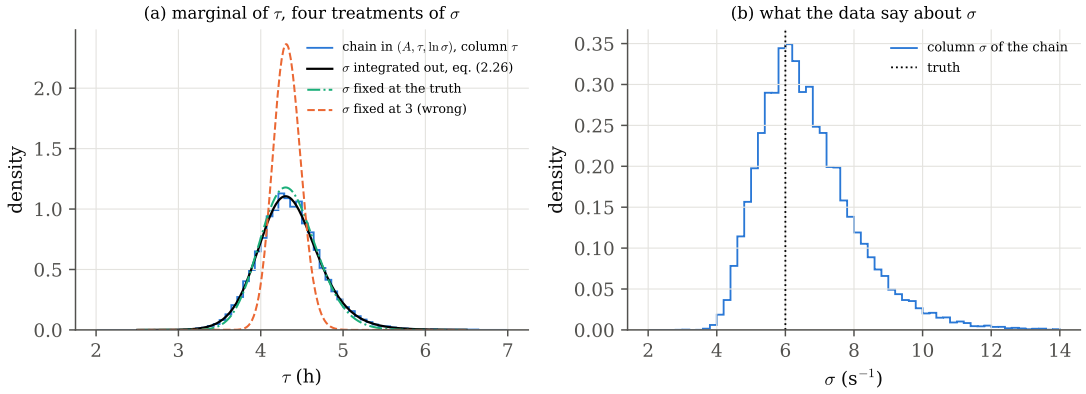


Figure 9.45: The noise level as a nuisance parameter. (a) The marginal of τ four ways: the τ column of the chain in $(A, \tau, \ln\sigma)$, which overlays the analytic marginal (9.109); σ fixed at its true value (narrower, since the uncertainty in σ is ignored); and σ fixed at a wrong value, half the truth (far too narrow). (b) The σ column of the same chain: what the data say about their own scatter.

Watch out: A fixed nuisance parameter is a hidden assumption

Fixing a nuisance parameter is marginalising it with a delta-function prior [HobsonBMC p. 42]. It is legitimate when the parameter really is known; when it is not, every width inherits the error of the guess, and the chain cannot tell. Before quoting a constraint, list what was held fixed. Section 9.22 will show that the error bar on $\ln(10^{10}A_s)$ from a CMB temperature spectrum changes by a factor of nine depending on whether one parameter is fixed.

9.18.3 Data sets that disagree: hyperparameters

Two laboratories measure the same quantity, each quotes an error bar, and the two results do not overlap. One of them has underestimated its errors, or both have, and nobody knows which. Averaging them with their stated weights gives a number with a small error bar that may be far from the truth; throwing one away by judgement is arbitrary. Padilla et al. pose this for the straight line [Padilla §§3.10, 5.2].

The question. When two data sets disagree and we do not know which is wrong, can the analysis itself discover that one does not belong and down-weight it, and can the data tell us whether doing so was justified?

The setup. Two cases, each with 16 points per data set and x drawn as $3 \times \mathcal{N}(0,1)$ as before:

	data set D_1	data set D_2
case 1 (agree)	$y = 3 + 2x$, $\sigma = 0.3$	$y = 3 + 2x$, $\sigma = 0.2$
case 2 (disagree)	$y = 3 + 2x$, $\sigma = 0.3$	$y = 3.5 + 1.5x$, $\sigma = 0.5$

Case 1 is the experiment of Section 9.17, in a new realisation. In case 2 the two sets come from different lines, but the analysis is not told so. Two models of the data are compared.

H_0: both data sets as stated. One line, and every point has the error bar its experiment quotes. This is the likelihood (9.99) of the straight line applied to all 32 points. In case 2 the single line lands between the two truths and fits neither; for our realisation $a = 3.276 \pm 0.066$, $b = 1.864 \pm 0.024$, with a best-fit $\chi^2 \approx 158$ for 32 points (Figure 9.46a, b). The posterior is confident and wrong, because the likelihood was told that both data sets have the stated errors and describe the same line.[5]

H_1: each data set may have its errors scaled by an unknown factor. Still one line, but data set k gets a weight α_k that multiplies its χ^2_k, which is the same as dividing its error bars by $\sqrt{\alpha_k}$. With $\alpha_k = 1$ the errors are as stated; a small α_k says "these error bars are underestimated". The weights are unknown and are marginalised, one per data set: this is the **hyperparameter method** [one extra weight per data set, unknown and integrated out]. We keep Padilla's letters: α_k for the weight, n_k for the number of points and χ^2_k for the χ^2 of data set k alone, as in (9.99); their covariance $\mathbf{V}$ of one data set is our $\mathrm{diag}(\sigma_i^2)$.

The prior on a weight. All we want to say about α is that it is positive and that the stated errors are right on average, $\mathbb{E}\,\alpha = 1$; nothing about how far a given experiment may stray. The least committal density with that one property is the maximum-entropy density of Section 2.11.3 with a flat measure on $\alpha > 0$. Maximise $-\int p \ln p \, \mathrm{d}\alpha$ subject to $\int p \, \mathrm{d}\alpha = 1$ and $\int \alpha \, p \, \mathrm{d}\alpha = 1$:

$$p(\alpha) \overset{(1)}{=} \frac{1}{Z} e^{-\lambda\alpha} \overset{(2)}{=} \lambda \, e^{-\lambda\alpha} \overset{(3)}{=} e^{-\alpha}, \qquad \alpha > 0. \tag{9.110}$$

(1) (2.69) with one constraint, $f(\alpha) = \alpha$, and a flat measure; λ is its Lagrange multiplier.
(2) Normalisation: $Z = \int_0^\infty e^{-\lambda\alpha} \mathrm{d}\alpha = 1/\lambda$.
(3) The mean of $\lambda e^{-\lambda\alpha}$ is $1/\lambda$, and it must be one, so $\lambda = 1$.

A mean of one centres the weights on "the errors are as quoted"; the exponential then lets an individual weight be small (errors underestimated) without any further assumption.

Their key equation: the likelihood with the weight integrated out. For one data set of n points,

$$\begin{aligned} p(D_k \,|\, \boldsymbol{\theta}, H_1) &\overset{(1)}{=} \int_0^\infty \frac{\alpha^{n/2}}{(2\pi)^{n/2}|\mathbf{V}|^{1/2}} \, e^{-\alpha\chi^2/2} \, e^{-\alpha} \, \mathrm{d}\alpha \overset{(2)}{=} \frac{1}{(2\pi)^{n/2}|\mathbf{V}|^{1/2}} \, \frac{\Gamma(n/2+1)}{(\chi^2/2+1)^{n/2+1}} \\ &\overset{(3)}{=} \frac{2\,\Gamma(n/2+1)}{\pi^{n/2}|\mathbf{V}|^{1/2}} \, (\chi^2+2)^{-(n/2+1)}. \end{aligned} \tag{9.111}$$

(1) Errors $\sigma_i/\sqrt{\alpha}$: each of the n Gaussian factors gains $\sqrt{\alpha}$ in its normalisation and α in its exponent. Times the prior (9.110).
(2) $\int_0^\infty \alpha^{s-1} e^{-c\alpha} \mathrm{d}\alpha = \Gamma(s)/c^s$ with $s = n/2+1$ and $c = \chi^2/2+1$.
(3) $(\chi^2/2+1)^{-(n/2+1)} = 2^{n/2+1}(\chi^2+2)^{-(n/2+1)}$, and $2^{n/2+1}/(2\pi)^{n/2} = 2/\pi^{n/2}$.

This is Padilla's eq. (50), and the H_1 likelihood of the two data sets is the product of one such factor per set. The Gaussian $e^{-\chi^2/2}$ has become the power law $(\chi^2+2)^{-(n/2+1)}$, the same thing that happened to the unknown σ in (9.109), now once per data set. A power law is forgiving: a line that misses one data set badly is penalised only logarithmically in that set's χ^2, so the posterior can follow the other.

[5] Padilla et al. quote $a = 3.528 \pm 0.056$, $b = 1.795 \pm 0.014$ for their realisation. By (9.102), errors of 0.3 and 0.5 allow no intercept width below $(16/0.3^2 + 16/0.5^2)^{-1/2} = 0.064$ for any choice of the x_i, so the published widths were not computed with these error bars; we do not try to reconstruct their setting.

What the posterior does. Under H_1 the posterior of (a, b) has two peaks (Figure 9.46c): one at the line of D_1, carrying 1.00 of the probability, and one near $(a, b) = (3.71, 1.61)$, towards the line of D_2, whose peak is lower by a factor $e^{7.2}$ and carries a fraction 0.0031. The better data set wins, because its smaller errors make it more informative. These weights come from integrating the posterior on a grid. Six chains started at random in the prior box do not find them: none of the 6 chains crossed between the peaks in 20 000 steps, each spending more than 98 per cent of its steps in a single peak. A chain started near the minor peak would stay there and report it as the whole posterior, and a plot of the points of several chains shows the peaks with whatever density the starting points happened to give them. This is the trap of Section 9.13 in practice. When the posterior is multimodal, histograms of chain points are not evidence about the relative weights of the modes; a tempered or nested sampler, or a direct integral as here, is needed.

Should we use H_1? The evidence ratio. Adding hyperparameters is a choice of model, and Chapter 5 answers such questions with the evidence $Z = \int \mathcal{L}(\boldsymbol{\theta})\, p(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}$ and the Bayes factor $K = Z_1/Z_0$. Both models share the flat prior on (a, b), here the box $-5 < a < 8$, $0 < b < 4$, and both Z are two-dimensional integrals, which we do by quadrature on a 1301×801 grid. For the disagreeing data sets (case 2) we find $\ln K = 40.6$, a factor of order 10^{17} in favour of H_1; for the agreeing ones (case 1), $K = 0.59$.

Where does a factor of 10^{17} come from? An integral over a likelihood with one sharp peak is the height of the peak times its width, so $\ln K$ is the difference of the two log-likelihoods at their peaks plus a difference of log-widths (the Occam terms of Chapter 5), which are a few units at most. The heights carry the effect. The normalisation $(2\pi)^{-n_k/2}|\mathbf{V}|^{-1/2}$ is common to the two models and cancels. With $n_k = 16$, so that $n_k/2 + 1 = 9$ and $\ln \Gamma(9) = \ln 40\,320 = 10.6$, case 2 gives:

- H_0 at its best line: $-\chi^2_{\min}/2 = -158/2 = -79$. An acceptable fit of 32 points would have χ^2 near 32 (the mean of a χ^2 is its number of degrees of freedom), so the single line is worse than an acceptable fit by $e^{-(158-32)/2} = e^{-63}$.
- H_1 at its main peak, the line of D_1: by line 2 of (9.111) each set contributes $\ln \Gamma(9) - 9\ln(\chi^2_k/2 + 1)$. D_1 fits its own line, $\chi^2_1 \approx 16$, giving $10.6 - 9\ln 9 = -9.2$. D_2 misses that line by $(3.5 + 1.5x) - (3 + 2x) = 0.5(1 - x)$, whose mean square over $x \sim 3 \times \mathcal{N}(0, 1)$ is $0.25(1 + 9) = 2.5$; adding the noise variance 0.25 and dividing by 0.5^2 gives about 11 per point, $\chi^2_2 \approx 176$, and a contribution $10.6 - 9\ln 89 = -29.8$. In all, -39.

The difference is $-39 - (-79) = 40$, against the quadrature's 40.6. The decisive item is the treatment of D_2: under H_0 a misfit of $\chi^2_2 \approx 176$ costs $176/2 = 88$ in $\ln \mathcal{L}$, under H_1 it costs $9\ln 89 \approx 40$ less a constant, because the power law charges the logarithm of χ^2 rather than χ^2 itself. No Occam penalty from two extra parameters can compensate for that. The same count for case 1, where both sets fit their common line with $\chi^2_k \approx 16$ each, gives H_1 two terms of -9.2 against H_0's $-32/2 = -16$, a difference of about -2: the weights buy nothing when the errors are right and cost a little. The power law's broader peak gives back part of that in the width term, and the quadrature gives $\ln K = -0.53$, mildly in favour of H_0.

Side by side. Padilla et al. quote Bayes factors for both cases.

	case 1 (data sets agree)	case 2 (data sets disagree)
Padilla et al., K [Padilla eqs. 71–72]	3	37
ours, $K = Z_1/Z_0$ by quadrature	0.59 ($\ln K = -0.53$)	$\ln K = 40.6$
verdict [33]	barely worth a mention	decisive for H_1

Why they differ. The verdicts agree; the case-2 numbers differ by $\ln(4\times10^{17}/37)\approx 37$ e-folds. The published values come from Padilla's eq. (51), which is not an evidence ratio. To see what it should have been as a ratio of likelihoods, divide (9.111) by the Gaussian likelihood of the same data set at the same $\boldsymbol{\theta}$:

$$\begin{aligned}\frac{p(D_k\,|\,\boldsymbol{\theta},H_1)}{p(D_k\,|\,\boldsymbol{\theta},H_0)} &\overset{(1)}{=} \frac{2\,\Gamma(n/2+1)\,\pi^{-n/2}|\mathbf{V}|^{-1/2}(\chi^2+2)^{-(n/2+1)}}{(2\pi)^{-n/2}|\mathbf{V}|^{-1/2}e^{-\chi^2/2}}\\ &\overset{(2)}{=} 2\,(2\pi)^{n/2}\pi^{-n/2}\,\Gamma(n/2+1)\,(\chi^2+2)^{-(n/2+1)}\,e^{+\chi^2/2}\\ &\overset{(3)}{=} 2^{n/2+1}\,\Gamma(n/2+1)\,(\chi^2+2)^{-(n/2+1)}\,e^{+\chi^2/2}. \end{aligned}\tag{9.112}$$

(1) Numerator (9.111); denominator the Gaussian likelihood (9.99) written with $\mathbf{V}=\mathrm{diag}(\sigma_i^2)$.
(2) $|\mathbf{V}|^{-1/2}$ cancels; dividing by $e^{-\chi^2/2}$ is multiplying by $e^{+\chi^2/2}$.
(3) $2\,(2\pi)^{n/2}/\pi^{n/2}=2\cdot 2^{n/2}=2^{n/2+1}$.

For two data sets the ratio is the product over k. Eq. (51) as printed reads $K=\prod_k 2^{n_k/2+1}\Gamma(n_k/2+1)e^{-\chi_k^2/2}/(\chi_k^2+2)$. Three things separate it from $K=Z_1/Z_0$, each sized here for case 2 at the H_0 best fit, where the two χ_k^2 add up to 158:

1. *The power of* (χ_k^2+2): 1 in print, $n_k/2+1=9$ in (9.112). The missing factor $(\chi_k^2+2)^{-8}$ per set is $8\sum_k\ln(\chi_k^2+2)$ e-folds, between 46 and 70 however the 158 is split between the two sets.
2. *The sign of* $\chi_k^2/2$: $e^{-\chi_k^2/2}$ in print, $e^{+\chi_k^2/2}$ in (9.112). Together the two sets differ by $e^{\sum_k\chi_k^2}=e^{158}$.
3. *A ratio at one point is not a ratio of integrals.* Even corrected, (9.112) compares the two likelihoods at one $\boldsymbol{\theta}$, while Z_1/Z_0 integrates each over its own prior, and the two models peak at different lines. At the H_0 best fit the corrected ratio is $\sum_k[9\ln 2+\ln\Gamma(9)-9\ln(\chi_k^2+2)]+79$, between 34 and 61 e-folds depending on the split, against $\ln K=40.6$; at another $\boldsymbol{\theta}$ it would be another number.

We cannot say which combination of these produced the published 37, since the paper does not give the point at which eq. (51) was evaluated.

The lesson. A Bayes factor is a ratio of integrals over the whole prior, and a likelihood ratio at one point, however carefully written, is a different number, here by tens of e-folds. A direct integral (or, in more dimensions, nested sampling or the importance-sampling identity of their eq. 48, Section 9.23.1) is the way to compute it.

9.19 The likelihood of a CMB power spectrum

The straight line and the decaying source had data that were measurements of the model plus noise. The CMB is different in kind. Cosmology does not predict the temperature in any direction of the sky; it predicts the *statistics* of the temperature field, its power spectrum C_ℓ, and the sky we see is one random realisation of it (Chapter 7). The data are therefore the measured spectrum $\widehat{C}_\ell$ of that one realisation, and the likelihood has to say how probable this $\widehat{C}_\ell$ is if the true spectrum were C_ℓ. On the full sky the answer is exact and short, and every practical CMB likelihood is an approximation to it, so we derive it first and then test the approximations against it.

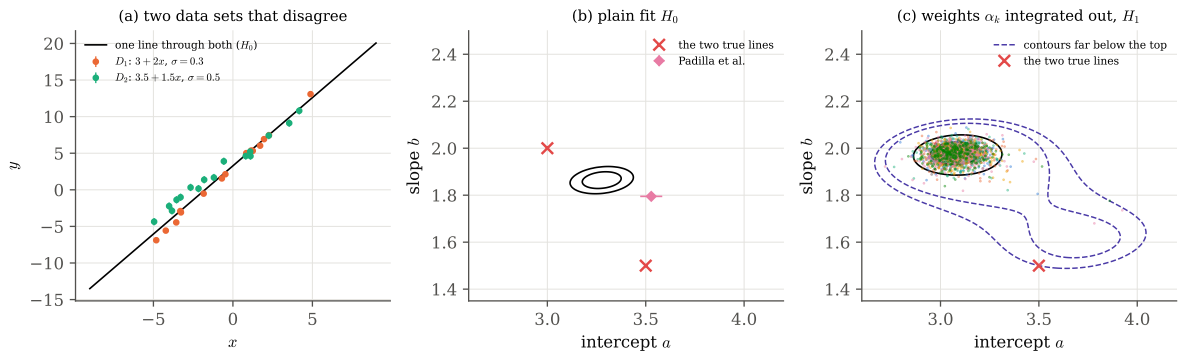


Figure 9.46: Two data sets drawn from different lines. (a) The data and the single line of H_0, which fits neither. (b) The H_0 posterior (68 and 95 per cent), far from both true lines (crosses), with the published result of Padilla et al. for their realisation. (c) With one marginalised weight per data set (H_1): the posterior concentrates on the line of the more precise data set; contours drawn far below the top (dashed) reveal the second, much lower peak towards the other line. Coloured points are the six chains, each confined to a single peak.

The question in words. It is the question of the straight line in Section 9.17, asked in a much larger space of hypotheses. The observed event is the whole data vector, here the measured spectrum $\widehat{C}_\ell$ at every multipole at once, standing in for the map it was computed from. The hypotheses are the cosmological models, one for every value of the ΛCDM parameter vector, as the candidate lines were one for every pair (a, b). To each hypothesis we put the same question: if nature had chosen this model, how probable would it have been to observe exactly this spectrum? That number, read as a function of the parameters, is the likelihood, and Bayes' theorem combines it with the prior into a posterior over the models (Section 5.7). What is new for the CMB is only the first step, the probability of a spectrum given a model.

9.19.1 The chain of dependence: parameters, spectrum, sky

Three levels connect the parameters to the data (Figure 9.47). The parameters $\boldsymbol{\theta}$ (amplitude and tilt of the primordial spectrum, densities, optical depth) fix the angular power spectrum through a Boltzmann code, $C_\ell = C_\ell(\boldsymbol{\theta})$, with no randomness: $C_\ell = \sum_m \int \mathrm{d}k\, T_m(\ell, k) P_m(k)$ in the notation of the cosmology notes, a weighted integral of the primordial power over transfer functions (Chapter T1 maps these objects). The spectrum fixes the *distribution* of the harmonic coefficients, $a_{\ell m} \sim \mathcal{N}(0, C_\ell)$, independent for different ℓ and m (statistical isotropy, Chapter 7). The coefficients give the map, and the map gives the data. Hobson et al. write the posterior of the parameters with the middle level integrated out [HobsonBMC eqs. 10.26–10.28]:

$$p(\boldsymbol{\theta} \mid \boldsymbol{d}) \overset{(1)}{=} \int \mathrm{d}C_\ell \; p(\boldsymbol{\theta}, C_\ell \mid \boldsymbol{d}) \overset{(2)}{=} \frac{1}{p(\boldsymbol{d})} \int \mathrm{d}C_\ell \; p(\boldsymbol{d} \mid C_\ell)\, p(C_\ell \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})$$
$$\overset{(3)}{=} \frac{p(\boldsymbol{\theta})}{p(\boldsymbol{d})} \int \mathrm{d}C_\ell \; p(\boldsymbol{d} \mid C_\ell)\, \delta\big(C_\ell - C_\ell(\boldsymbol{\theta})\big) \overset{(4)}{=} \frac{p(\boldsymbol{\theta})}{p(\boldsymbol{d})}\, p\big(\boldsymbol{d} \mid C_\ell = C_\ell(\boldsymbol{\theta})\big). \tag{9.113}$$

(1) The spectrum is an intermediate unknown; marginalise it.
(2) Bayes' theorem for the pair $(\boldsymbol{\theta}, C_\ell)$, with the product rule $p(\boldsymbol{d}, C_\ell, \boldsymbol{\theta}) = p(\boldsymbol{d} \mid C_\ell)\, p(C_\ell \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})$: once the spectrum is known, the data do not depend on how it was produced.
(3) The spectrum is a deterministic function of the parameters, so its conditional density is a delta function [HobsonBMC p. 240].
(4) The delta function picks the value $C_\ell(\boldsymbol{\theta})$.

So the likelihood of the parameters is the likelihood of the spectrum, evaluated at the spectrum

the parameters predict. All the statistics lives in $p(\boldsymbol{d} \mid C_\ell)$, and all the cosmology in the map $\boldsymbol{\theta} \mapsto C_\ell(\boldsymbol{\theta})$.

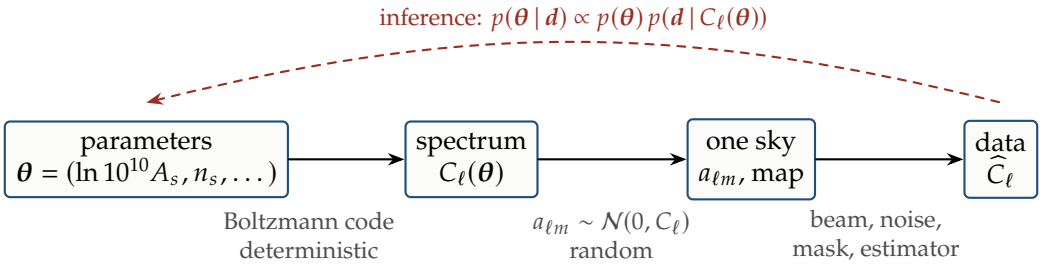


Figure 9.47: The CMB as a chain of dependence. Reading from left to right, the parameters determine the spectrum, which in turn sets the distribution of the harmonic coefficients from which an observed sky is drawn. The dashed arc traces inference in the reverse direction. Because $C_\ell(\boldsymbol{\theta})$ is fixed once the parameters are specified, their likelihood is the spectrum likelihood evaluated at that prediction.

9.19.2 The exact full-sky likelihood

Take one multipole ℓ of a full, noiseless sky. Its $2\ell+1$ real degrees of freedom are $a_{\ell 0}$, which is real with variance C_ℓ, and the real and imaginary parts x_m, y_m of $a_{\ell m}$ for $m = 1, \dots, \ell$, each with variance $C_\ell/2$ (the $m < 0$ coefficients are their complex conjugates and carry no new information; Chapter 7). The estimator is $\widehat{C}_\ell = \frac{1}{2\ell+1} \sum_{m=-\ell}^{\ell} |a_{\ell m}|^2 = \frac{1}{2\ell+1}\left[a_{\ell 0}^2 + 2\sum_{m\geq 1}(x_m^2 + y_m^2)\right]$. Write $\nu = 2\ell + 1$. The probability of the observed coefficients, given a candidate C_ℓ, is

$$p(\{a_{\ell m}\} \mid C_\ell) \overset{(1)}{=} \frac{e^{-a_{\ell 0}^2/2C_\ell}}{\sqrt{2\pi C_\ell}} \prod_{m=1}^{\ell} \frac{e^{-(x_m^2+y_m^2)/C_\ell}}{\pi C_\ell} \overset{(2)}{=} (2\pi)^{-1/2} \pi^{-\ell}\, C_\ell^{-\nu/2}\, \exp\!\left[-\frac{a_{\ell 0}^2 + 2\sum_{m\geq 1}(x_m^2 + y_m^2)}{2C_\ell}\right]$$

$$\overset{(3)}{=} \text{const} \times C_\ell^{-\nu/2}\, e^{-\nu \widehat{C}_\ell / 2C_\ell}, \qquad \text{so} \qquad -2\ln \mathcal{L}(C_\ell) = \nu\left[\frac{\widehat{C}_\ell}{C_\ell} + \ln C_\ell\right] + \text{const}. \tag{9.114}$$

(1) Independent Gaussians: $a_{\ell 0}$ with variance C_ℓ; x_m and y_m each with variance $C_\ell/2$, whose two densities multiply to $(\pi C_\ell)^{-1} e^{-(x^2+y^2)/C_\ell}$.

(2) Collect the powers of C_ℓ: $\frac{1}{2} + \ell = \nu/2$; write the exponents over a common denominator $2C_\ell$.

(3) The numerator of the exponent is $\nu \widehat{C}_\ell$ by the definition of $\widehat{C}_\ell$.

This is eq. (6) of Hamimeche and Lewis for one field [16]. Three facts follow at once. *The data enter only through* $\widehat{C}_\ell$. The density (9.114) is a constant times a function of $\widehat{C}_\ell$ and C_ℓ alone, which is the factorised form of Section 3.16.3: on the full sky the spectrum is a **sufficient statistic** [a summary that carries all the information the data have about C_ℓ], and nothing is lost by compressing a map of millions of pixels to one number per ℓ. *The likelihood is largest at* $C_\ell = \widehat{C}_\ell$: set the derivative $\nu[-\widehat{C}_\ell/C_\ell^2 + 1/C_\ell]$ to zero. *Different* ℓ *are independent*, so $-2\ln\mathcal{L}$ of a whole spectrum is the sum of (9.114) over ℓ.

The same density read the other way. At fixed C_ℓ, what is the distribution of the estimate $\widehat{C}_\ell$? The exponent of (9.114) contains $\nu\widehat{C}_\ell = a_{\ell 0}^2 + \sum_{m\geq 1}\left[(\sqrt{2}x_m)^2 + (\sqrt{2}y_m)^2\right]$, a sum of ν squares of independent Gaussians, each of variance C_ℓ ($\sqrt{2}x_m$ has variance $2 \cdot C_\ell/2$). Divided by C_ℓ it is a sum of ν squared standard Gaussians, so $y \equiv \nu\widehat{C}_\ell/C_\ell \sim \chi^2_\nu$ (Section 2.13.2). Change variables

from y to $\widehat{C}_\ell$:

$$p(\widehat{C}_\ell \mid C_\ell) \overset{(1)}{=} \frac{y^{\nu/2-1}e^{-y/2}}{2^{\nu/2}\Gamma(\nu/2)}\,\frac{\mathrm{d}y}{\mathrm{d}\widehat{C}_\ell} \overset{(2)}{=} \frac{1}{2^{\nu/2}\Gamma(\nu/2)}\Big(\frac{\nu\widehat{C}_\ell}{C_\ell}\Big)^{\nu/2-1} e^{-\nu\widehat{C}_\ell/2C_\ell}\,\frac{\nu}{C_\ell}$$
$$\overset{(3)}{=} \text{const} \times \widehat{C}_\ell^{\,(\nu-2)/2}\, C_\ell^{-\nu/2}\, e^{-\nu\widehat{C}_\ell/2C_\ell}. \tag{9.115}$$

(1) The χ^2_ν density of y, times the Jacobian of $y = \nu\widehat{C}_\ell/C_\ell$.
(2) Substitute y and $\mathrm{d}y/\mathrm{d}\widehat{C}_\ell = \nu/C_\ell$.
(3) Collect the powers: $C_\ell^{-(\nu/2-1)}C_\ell^{-1} = C_\ell^{-\nu/2}$; the constant holds ν and the Gamma function.

This is the scaled χ^2_ν law of Chapter 7 (HL eq. 8). As a function of C_ℓ it is (9.114) again: the extra factor $\widehat{C}_\ell^{\,(\nu-2)/2}$ is fixed once the data are fixed and changes nothing in the likelihood.

Several fields. With temperature and E-polarization observed together, each mode is a vector $\boldsymbol{a}_{\ell m}$ of two numbers with a 2×2 covariance $\mathbf{C}_\ell$, and the same three moves give the two-field likelihood. Take as the ν real modes $\boldsymbol{a}_{\ell 0}$ and the real and imaginary parts of the $m \geq 1$ coefficients, each multiplied by $\sqrt{2}$ as above, so that every mode has covariance $\mathbf{C}_\ell$ and density $(2\pi)^{-1}|\mathbf{C}_\ell|^{-1/2}\exp(-\frac{1}{2}\boldsymbol{a}^{\mathsf{T}}\mathbf{C}_\ell^{-1}\boldsymbol{a})$ (Eq. (2.78)). The modes are independent, so

$$-2\ln\mathcal{L} \overset{(1)}{=} \sum_{\text{modes}} \boldsymbol{a}^{\mathsf{T}}\mathbf{C}_\ell^{-1}\boldsymbol{a} + \nu\ln|\mathbf{C}_\ell| + \text{const} \overset{(2)}{=} \mathrm{Tr}\Big(\mathbf{C}_\ell^{-1}\sum_{\text{modes}}\boldsymbol{a}\boldsymbol{a}^{\mathsf{T}}\Big) + \nu\ln|\mathbf{C}_\ell| + \text{const}$$
$$\overset{(3)}{=} \nu\big[\mathrm{Tr}(\widehat{\mathbf{C}}_\ell\mathbf{C}_\ell^{-1}) + \ln|\mathbf{C}_\ell|\big] + \text{const}. \tag{9.116}$$

(1) Minus twice the logarithm of a product of ν Gaussian densities.
(2) A number equals its own trace, and $\mathrm{Tr}(\boldsymbol{a}^{\mathsf{T}}\mathbf{M}\boldsymbol{a}) = \mathrm{Tr}(\mathbf{M}\boldsymbol{a}\boldsymbol{a}^{\mathsf{T}})$ (a trace is unchanged when the factors are cycled); the trace of a sum is the sum of the traces.
(3) The estimate of the spectrum matrix is $\widehat{\mathbf{C}}_\ell = \nu^{-1}\sum_{\text{modes}}\boldsymbol{a}\boldsymbol{a}^{\mathsf{T}}$, and $\mathrm{Tr}(\mathbf{AB}) = \mathrm{Tr}(\mathbf{BA})$.

For one field the trace is $\widehat{C}_\ell/C_\ell$ and the determinant is C_ℓ, and (9.114) comes back (HL eqs. 6–7). The matrix $\nu\widehat{\mathbf{C}}_\ell$ is a sum of outer products of Gaussian vectors, the Wishart matrix met in Section 8.13.2; we use only the one-field form from here on.

Instrumental effects enter simply. White noise of power N_ℓ is another independent Gaussian field, so the observed coefficients have variance $C_\ell + N_\ell$, and (9.114) holds with $C_\ell \to C_\ell + N_\ell$ [HobsonBMC eq. 10.24]; with a beam and pixel window T_ℓ (Chapter 8), $C_\ell \to T_\ell^2 C_\ell + N_\ell$. On a fraction f_{sky} of the sky the number of independent modes is roughly $\nu = (2\ell+1)f_{\mathrm{sky}}$ (Knox's mode count, Chapter 8), and (9.114) with this ν is the standard idealised cut-sky likelihood. It is exact for a sky observed with isotropic noise; for a real mask it is an approximation, since the mask couples multipoles.

Working definition 9.12 (power-spectrum likelihood)

For a Gaussian sky observed with beam and pixel window T_ℓ and isotropic noise N_ℓ on a fraction f_{sky} of the sky, the likelihood of parameters $\boldsymbol{\theta}$ given the observed spectrum $\widehat{C}_\ell$ is

$$-2\ln\mathcal{L}(\boldsymbol{\theta}) = \sum_\ell \nu_\ell\Big[\frac{\widehat{C}_\ell}{T_\ell^2 C_\ell(\boldsymbol{\theta}) + N_\ell} + \ln\big(T_\ell^2 C_\ell(\boldsymbol{\theta}) + N_\ell\big)\Big] + \text{const}, \qquad \nu_\ell = (2\ell+1)f_{\mathrm{sky}},$$

with ν_ℓ [the number of independent real modes at multipole ℓ, exact on the full sky]. *Operational check:* at each ℓ separately it is smallest when the predicted $T_\ell^2 C_\ell + N_\ell$ equals $\widehat{C}_\ell$, and data simulated as $\widehat{C}_\ell = (T_\ell^2 C_\ell + N_\ell)\chi^2_\nu/\nu$ have exactly this likelihood.

Example: the likelihood of the quadrupole is very skewed. As a function of C_ℓ at fixed $\widehat{C}_\ell$, $\mathcal{L} \propto C_\ell^{-\nu/2} e^{-\beta/C_\ell}$ with $\beta = \nu\widehat{C}_\ell/2$, an inverse-gamma shape with index $\alpha = \nu/2 - 1$ and scale β (Chapter 2). Its peak is where the derivative of $\ln\mathcal{L} = -\frac{\nu}{2}\ln C_\ell - \beta/C_\ell$ vanishes, $-\nu/2C_\ell + \beta/C_\ell^2 = 0$, that is at $C_\ell = 2\beta/\nu = \widehat{C}_\ell$. Its mean, with a flat prior on C_ℓ, is a ratio of two integrals of the same kind, each done with the substitution $u = \beta/C_\ell$:

$$\int_0^\infty C_\ell^{-s} e^{-\beta/C_\ell}\,\mathrm{d}C_\ell \overset{(1)}{=} \beta^{1-s}\int_0^\infty u^{s-2}e^{-u}\,\mathrm{d}u \overset{(2)}{=} \beta^{1-s}\,\Gamma(s-1),$$

$$\mathbb{E}[C_\ell] \overset{(3)}{=} \frac{\int C_\ell \cdot C_\ell^{-\nu/2} e^{-\beta/C_\ell}\mathrm{d}C_\ell}{\int C_\ell^{-\nu/2} e^{-\beta/C_\ell}\mathrm{d}C_\ell} \overset{(4)}{=} \beta\,\frac{\Gamma(\nu/2-2)}{\Gamma(\nu/2-1)} \overset{(5)}{=} \frac{\beta}{\nu/2-2} = \frac{\nu\widehat{C}_\ell}{\nu-4} = \widehat{C}_\ell\,\frac{2\ell+1}{2\ell-3}. \tag{9.117}$$

(1) $C_\ell = \beta/u$, $\mathrm{d}C_\ell = -\beta u^{-2}\mathrm{d}u$, and $C_\ell^{-s} = \beta^{-s}u^s$; swapping the limits absorbs the minus sign.
(2) The Gamma integral, which converges for $s > 1$.
(3) The posterior with a flat prior is the normalised likelihood.
(4) Line 2 with $s = \nu/2 - 1$ in the numerator and $s = \nu/2$ in the denominator; the powers of β leave one factor β.
(5) $\Gamma(z+1) = z\,\Gamma(z)$ with $z = \nu/2 - 2$; then $\beta = \nu\widehat{C}_\ell/2$ and $\nu = 2\ell + 1$.

The numerator needs $\nu/2 - 1 > 1$, so the mean exists for $\nu > 4$, that is from the quadrupole up (HL below eq. 8). At $\ell = 5$ the mean is $1.571\,\widehat{C}_\ell$; at the quadrupole, $\ell = 2$ and $\nu = 5$, it is $5\widehat{C}_\ell$, five times the peak. The likelihood has a long tail towards large C_ℓ, because a large true spectrum can produce a small $\widehat{C}_\ell$ by chance much more easily than a small spectrum can produce a large one. The other direction is skewed too. Given C_ℓ, the most probable $\widehat{C}_\ell$ is the peak of (9.115): in the variable $y = \nu\widehat{C}_\ell/C_\ell$ the density is proportional to $y^{\nu/2-1}e^{-y/2}$, whose logarithm has derivative $(\nu/2-1)/y - 1/2$, zero at $y = \nu - 2$. So the most probable $\widehat{C}_\ell$ is $C_\ell(\nu-2)/\nu$: $0.818\,C_\ell$ at $\ell = 5$ and $0.6\,C_\ell$ at the quadrupole. Low multipoles usually come out below their true value, and their likelihood leans upwards to compensate. Figure 9.48 shows the shapes.

9.19.3 Approximations, and what each gets wrong

On a masked sky with anisotropic noise the exact likelihood is a Gaussian in millions of pixels with a covariance that depends on C_ℓ; evaluating it costs $O(N_{\mathrm{pix}}^3)$ operations and is impossible at high ℓ [HobsonBMC p. 238]. Every high-ℓ CMB analysis therefore replaces it by a function of the pseudo-C_ℓ estimates of Chapter 8 that is quadratic in some variable, so that a covariance matrix measured on simulations can be put in. The full sky is where these approximations can be tested, because there the exact answer is (9.114). Normalise it so that it is zero at $C_\ell = \widehat{C}_\ell$ (HL eq. 15) and write $u = \widehat{C}_\ell/C_\ell - 1$ for the fractional mismatch:

$$-2\ln\mathcal{L}_{\mathrm{exact}} \overset{(1)}{=} \nu\big[u - \ln(1+u)\big] \overset{(2)}{=} \nu\Big[\frac{u^2}{2} - \frac{u^3}{3} + \frac{u^4}{4} - \dots\Big]. \tag{9.118}$$

(1) $\widehat{C}_\ell/C_\ell + \ln C_\ell - (1 + \ln\widehat{C}_\ell) = (1+u) - \ln(1+u) - 1$, using $\ln C_\ell - \ln\widehat{C}_\ell = -\ln(1+u)$.
(2) $\ln(1+u) = u - u^2/2 + u^3/3 - \dots$.

The common approximations of HL eqs. (22)–(28) all start from one idea. On the full sky $\widehat{C}_\ell$ has mean C_ℓ and variance $2C_\ell^2/\nu$ (the variance of χ^2_ν is 2ν, and $\widehat{C}_\ell = C_\ell\chi^2_\nu/\nu$). A Gaussian with that mean and variance is quadratic in the data, and a quadratic form is what makes a cut-sky analysis possible: replace the one variance by a covariance matrix of the pseudo-C_ℓ measured on simulations and the formula still works. The approximations differ in where the variance comes from, or in which variable the Gaussian is written:

- $\mathcal{L}_S$, the symmetric Gaussian: variance $2\widehat{C}_\ell^2/\nu$, taken from the data. It costs nothing to evaluate, since the variance does not change with the model.
- $\mathcal{L}_Q$, the quadratic form: variance $2C_\ell^2/\nu$, taken from the model being tested, which is the correct variance at each C_ℓ.
- $\mathcal{L}_f$, the fixed fiducial: variance $2(C_\ell^{\rm f})^2/\nu$ from one fiducial spectrum, computed once; this is what a simulated covariance matrix in a real analysis amounts to.
- The log-normal: a Gaussian in $\ln\widehat{C}_\ell$ with variance $2/\nu$, because $\ln\widehat{C}_\ell$ is closer to Gaussian than $\widehat{C}_\ell$.
- WMAP: a weighted sum of $\mathcal{L}_Q$ and the log-normal [184], with weights fixed below.
- Gaussian$_D$: $\mathcal{L}_Q$ with the normalisation of its Gaussian kept, which adds the logarithm of the model variance, $\ln C_\ell$ once per multipole as HL print it (their eq. 25).
- The cube-root forms: a Gaussian in $\widehat{C}_\ell^{1/3}$, the power of a χ^2 variable whose distribution is most nearly symmetric (HL eq. 28, in two variants labelled $\alpha = \pm 1$).

To compare them with the exact form (9.118), write each in the same variable $u = \widehat{C}_\ell/C_\ell - 1$, so that $\widehat{C}_\ell = C_\ell(1+u)$, and expand to third order. For $\mathcal{L}_S$:

$$-2\ln\mathcal{L}_S \overset{(1)}{=} \frac{\nu}{2}\Big(\frac{\widehat{C}_\ell - C_\ell}{\widehat{C}_\ell}\Big)^2 \overset{(2)}{=} \frac{\nu}{2}\,\frac{u^2}{(1+u)^2} \overset{(3)}{=} \frac{\nu}{2}\,u^2(1-2u+3u^2-\dots) \overset{(4)}{=} \frac{\nu}{2}u^2 - \nu u^3 + \dots \qquad (9.119)$$

(1) A Gaussian in $\widehat{C}_\ell$ with mean C_ℓ and variance $2\widehat{C}_\ell^2/\nu$.
(2) $\widehat{C}_\ell - C_\ell = C_\ell u$ and $\widehat{C}_\ell = C_\ell(1+u)$; C_ℓ cancels.
(3) $(1+u)^{-2} = 1 - 2u + 3u^2 - \dots$.
(4) Keep terms up to u^3.

The others follow in the same way. $\mathcal{L}_Q$ is $\frac{\nu}{2}[(\widehat{C}_\ell - C_\ell)/C_\ell]^2 = \frac{\nu}{2}u^2$ exactly. The log-normal is $\frac{\nu}{2}\ln^2(1+u)$, and $\ln(1+u) = u - u^2/2 + \dots$ squares to $u^2 - u^3 + \dots$.

name	$-2\ln\mathcal{L}$	cubic term in u
exact	$\nu[u - \ln(1+u)]$	$-\nu u^3/3$
$\mathcal{L}_S$, symmetric Gaussian, variance from $\widehat{C}_\ell$	$\frac{\nu}{2}\big[(\widehat{C}_\ell - C_\ell)/\widehat{C}_\ell\big]^2 = \frac{\nu}{2}\frac{u^2}{(1+u)^2}$	$-\nu u^3$
$\mathcal{L}_Q$, variance from the model	$\frac{\nu}{2}\big[(\widehat{C}_\ell - C_\ell)/C_\ell\big]^2 = \frac{\nu}{2}u^2$	0
$\mathcal{L}_f$, variance from a fixed fiducial	$\frac{\nu}{2}\big[(\widehat{C}_\ell - C_\ell)/C_\ell^{\rm f}\big]^2$	(see text)
log-normal	$\frac{\nu}{2}\big[\ln(\widehat{C}_\ell/C_\ell)\big]^2 = \frac{\nu}{2}\ln^2(1+u)$	$-\nu u^3/2$
WMAP	$\frac{1}{3}\mathcal{L}_Q + \frac{2}{3}$ log-normal	$-\nu u^3/3$

Every one of them matches the exact $\nu u^2/2$, so all have the right width near the peak. They differ in the cubic term, which is the skewness of the previous example. What does a cubic term do? A negative u means a model C_ℓ above the measured $\widehat{C}_\ell$, a positive u one below it. A cubic coefficient more negative than the exact $-\nu/3$ adds to $-2\ln\mathcal{L}$ for $u<0$ and subtracts for $u>0$, so it penalises models above $\widehat{C}_\ell$ too much and models below too little, and the best fit is pulled low. A coefficient less negative than $-\nu/3$ pulls it high. So $\mathcal{L}_S$ ($-\nu$) biases the spectrum low; $\mathcal{L}_Q$ (0) penalises low C_ℓ too much and biases it high; the log-normal ($-\nu/2$) biases it low again, by less than $\mathcal{L}_S$. The WMAP combination was built to get the cubic term right [184]: weights w_Q and $w_{\rm LN} = 1 - w_Q$ must satisfy $w_Q\cdot 0 + w_{\rm LN}(-\nu/2) = -\nu/3$, so $w_{\rm LN} = 2/3$, which is why the weights are $1/3$ and $2/3$.

Example: how large is the bias of $\mathcal{L}_Q$? Fit an amplitude A, $C_\ell = AC_\ell^{\rm f}$, to a band of L multipoles on a noiseless full sky whose true spectrum is $C_\ell^{\rm f}$, the test of HL §IV A. Write $x_\ell = \widehat{C}_\ell / C_\ell^{\rm f}$, so that $\nu x_\ell \sim \chi^2_\nu$, with $\mathbb{E}\,x = 1$, $\mathrm{Var}\,x = 2/\nu$, $\mathbb{E}\,x^2 = 1 + 2/\nu$ and $\mathbb{E}\,x^3 = (\nu+2)(\nu+4)/\nu^2$ (moments of χ^2, Chapter 2). *The exact best fit.* The exact likelihood is $\sum \nu(x_\ell/A + \ln A)$, with derivative $\sum \nu(-x_\ell/A^2 + 1/A)$; it vanishes at $\hat{A} = \sum \nu x_\ell / \sum \nu$ (HL eq. 17), whose mean is exactly one. *The best fit of $\mathcal{L}_Q$.* $\mathcal{L}_Q$ is $\sum \frac{\nu}{2}(x_\ell/A - 1)^2$. Its derivative is

$$\frac{\mathrm{d}}{\mathrm{d}A} \sum \frac{\nu}{2}\Big(\frac{x_\ell}{A} - 1\Big)^2 = \sum \nu \Big(\frac{x_\ell}{A} - 1\Big)\Big(-\frac{x_\ell}{A^2}\Big),$$

and setting it to zero and multiplying by $-A^2$ gives $\sum \nu x_\ell^2/A = \sum \nu x_\ell$, so $\hat{A}_Q = \sum \nu x_\ell^2 \big/ \sum \nu x_\ell$, a ratio N/D of two random sums. *The mean of a ratio.* Write $D = \mathbb{E}\,D\,(1+\delta)$ with $\delta = (D - \mathbb{E}\,D)/\mathbb{E}\,D$ small, and expand $1/(1+\delta) = 1 - \delta + \delta^2 - \ldots$:

$$\mathbb{E}\,\frac{N}{D} = \frac{\mathbb{E}[N(1-\delta+\delta^2)]}{\mathbb{E}\,D} + \cdots \simeq \frac{\mathbb{E}\,N}{\mathbb{E}\,D} - \frac{\mathrm{Cov}(N,D)}{(\mathbb{E}\,D)^2} + \frac{\mathbb{E}\,N\ \mathrm{Var}\,D}{(\mathbb{E}\,D)^3},$$

using two facts. Since $\mathbb{E}\,\delta = 0$, the mean $\mathbb{E}[N\delta]$ equals $\mathrm{Cov}(N,D)/\mathbb{E}\,D$; and, to leading order, $\mathbb{E}[N\delta^2] \simeq \mathbb{E}\,N\ \mathrm{Var}\,D/(\mathbb{E}\,D)^2$. With $\Sigma \equiv \sum_\ell \nu_\ell$,

$$\mathbb{E}\,\hat{A}_Q \overset{(1)}{=} \frac{\Sigma + 2L}{\Sigma} - \frac{4\Sigma}{\Sigma^2} + \frac{(\Sigma+2L)\,2\Sigma}{\Sigma^3} \overset{(2)}{=} 1 + \frac{2(L-1)}{\Sigma} + O(\Sigma^{-2}). \tag{9.120}$$

(1) $\mathbb{E}\,N = \sum \nu(1 + 2/\nu) = \Sigma + 2L$; $\mathbb{E}\,D = \Sigma$; $\mathrm{Var}\,D = \sum \nu^2 \cdot 2/\nu = 2\Sigma$; $\mathrm{Cov}(N,D) = \sum \nu^2\,\mathrm{Cov}(x^2, x) = \sum \nu^2(\mathbb{E}\,x^3 - \mathbb{E}\,x^2) = \sum \nu^2(4/\nu + 8/\nu^2) \simeq 4\Sigma$.
(2) Collect terms: $2L/\Sigma - 4/\Sigma + 2/\Sigma$, dropping terms of order L/Σ^2.

For HL's bins of $L = 10$ starting at $\ell_{\rm min}$, $\Sigma = \sum_{\ell=\ell_{\rm min}}^{\ell_{\rm min}+9}(2\ell+1) = 20\ell_{\rm min} + 100$, so the bias is $18/(20\ell_{\rm min} + 100) = 0.9/(\ell_{\rm min} + 5)$: about $0.9/\ell$, just inside the tolerance $1/\ell$ of Section 9.19.4. At $\ell_{\rm min} = 1022$ the formula gives 8.8×10^{-4}; the simulation below gives 8.770×10^{-4}, and HL's Fig. 1 shows 9.100×10^{-4}.

The fixed fiducial needs no such calculation. With $C_\ell^{\rm f}$ equal to the truth, $\mathcal{L}_f = \sum \frac{\nu}{2}(x_\ell - A)^2$, with derivative $-\sum \nu(x_\ell - A)$, is maximised at $\sum \nu x_\ell / \sum \nu$, *identical* to the exact $\hat{A}$ in every realisation. The reason is visible in the two derivatives. The exact one is $\sum \nu(A - x_\ell)/A^2$, the fixed-fiducial one $\sum \nu(A - x_\ell)$: they differ by the positive factor $1/A^2$, so they vanish together. The same holds for any parameter. With a general model $C_\ell(\boldsymbol{\theta})$ the exact derivative is $\sum \nu(C_\ell - \widehat{C}_\ell)\,\partial C_\ell / C_\ell^2$ and the fixed-fiducial one is $\sum \nu(C_\ell - \widehat{C}_\ell)\,\partial C_\ell/(C_\ell^{\rm f})^2$, so if the fiducial equals the best-fit model, $C_\ell^{\rm f} = C_\ell(\hat{\boldsymbol{\theta}})$, the two agree term by term at $\hat{\boldsymbol{\theta}}$, and the exact best fit is also a best fit of $\mathcal{L}_f$. This is why a fixed fiducial covariance gives unbiased parameters even though its shape at each ℓ is wrong (HL eqs. 38–39).

Example: what the bias means for a parameter. A relative bias of $0.9/\ell$ in the spectrum is larger at low ℓ than at high ℓ. A parameter that moves power between low and high multipoles, like the tilt n_s, therefore reads it as a tilt: power pushed up at low ℓ looks like a redder spectrum. HL found that $\mathcal{L}_Q$ shifts n_s by about one standard deviation at Planck sensitivity while $\mathcal{L}_f$ does not (HL §III C). We test both statements with our own chains in Sections 9.21.5 and 9.22.6.

9.19.4 Reproducing Hamimeche and Lewis, Fig. 1

Their question. In 2008, with Planck about to fly, every high-ℓ CMB analysis used one of the Gaussian shortcuts of Section 9.19.3, because the exact likelihood cannot be evaluated on a

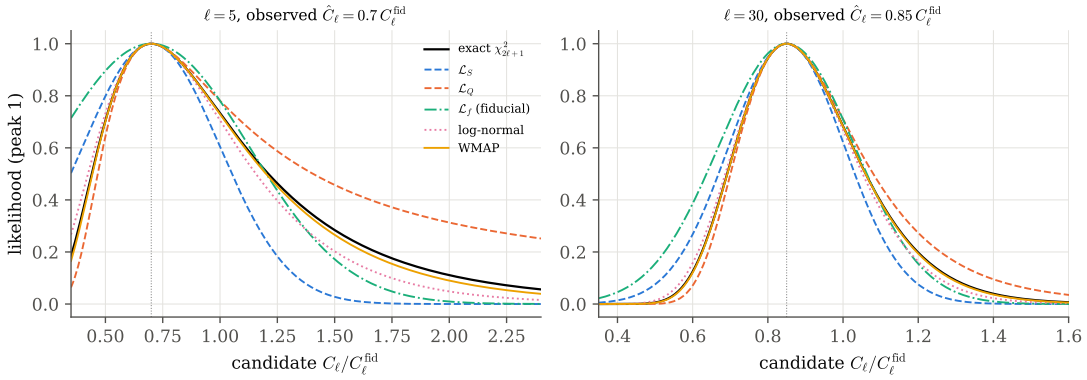


Figure 9.48: Likelihood of the true spectrum, in units of a fiducial one, after observing $\widehat{C}_\ell = 0.7\, C_\ell^{\rm fid}$ at $\ell = 5$ (left) and $0.85\, C_\ell^{\rm fid}$ at $\ell = 30$ (right). The exact likelihood (black) peaks at the observed value with a long tail to the right. The symmetric approximation $\mathcal{L}_S$ is too narrow on the high side; $\mathcal{L}_Q$ has far too heavy a tail there (it tends to a constant as C_ℓ grows); the fixed-fiducial $\mathcal{L}_f$ is symmetric; the WMAP combination is close to exact. At $\ell = 30$ all forms agree better: the distribution of $\widehat{C}_\ell$ becomes Gaussian as ℓ grows.

masked sky. Hamimeche and Lewis asked which of them is safe [16]: which ones return the same spectrum as the exact likelihood, to within an error small enough not to matter at Planck's precision? The cost of a wrong answer is a biased parameter: as the previous example showed, a bias that depends on ℓ reads as a tilt, and for $\mathcal{L}_Q$ it moves n_s by about one standard deviation.

Their setup, one item at a time. *The truth.* A fiducial spectrum $C_\ell^{\rm f}$, and many simulated full, noiseless skies drawn from it. On the full sky the exact likelihood is known, so each approximation can be compared with it sky by sky. *The test statistic.* The data are cut into bins of ten multipoles, $\ell_{\rm min} \le \ell \le \ell_{\rm min} + 9$, and in each bin one amplitude is fitted, $C_\ell = A C_\ell^{\rm f}$, by maximising the exact likelihood and each approximation in turn. An amplitude is the simplest parameter that a bias in the spectrum can move. *Bias and scatter.* For approximation i, the bias is the mean of $A_i - A_{\rm exact}$ over the simulated skies, and the scatter is the root-mean-square of the same difference. The bias says whether the approximation is systematically off; the scatter says how far it is off on a typical sky. *The tolerance.* An error matters only if it is comparable with the cosmic-variance error of the amplitude itself. For an amplitude fitted to all multipoles up to ℓ, the exact $\hat{A} = \sum \nu x_\ell / \sum \nu$ has variance $\sum \nu^2 (2/\nu)/(\sum \nu)^2 = 2/\sum \nu$, and $\sum_{\ell'=2}^{\ell} (2\ell'+1) = (\ell+1)^2 - 4 \approx \ell^2$, so its error is about $\sqrt{2}/\ell$. HL take $1/\ell_{\rm min}$ as the line below which a bias in one bin is harmless (their eqs. 19–20).

Their key equations. The exact best fit and the best fits of the approximations are those derived in Section 9.19.3: $\hat{A} = \sum \nu x_\ell / \sum \nu$ for the exact likelihood and for $\mathcal{L}_f$, and $\hat{A}_Q = \sum \nu x_\ell^2 / \sum \nu x_\ell$ for $\mathcal{L}_Q$, whose bias is (9.120) (HL eqs. 17–20). The symbols map directly: x_ℓ is the ratio $\widehat{C}_\ell / C_\ell^{\rm f}$ of the simulated to the fiducial spectrum, $\nu = 2\ell + 1$ the number of modes at ℓ on the full sky, and $L = 10$ the width of a bin.

What we compute. Only x_ℓ matters, and $\nu x_\ell \sim \chi^2_\nu$ exactly ((9.115)), so we draw the x_ℓ directly from χ^2 distributions, 10000 realisations per bin as in the paper, for 70 values of $\ell_{\rm min}$ between 2 and 1500. The best fits of $\mathcal{L}_S$, $\mathcal{L}_Q$ and the log-normal have closed forms; for Gaussian$_D$, WMAP and the cube-root form we use a vectorised golden-section search, as HL did. The likelihood functions are:

Listing 9.12: The exact likelihood and its approximations at one multipole (code/ch09/42_cl_likelihoods.py)

```
def m2lnL(kind, x, A, l):
    """-2 ln L (up to a constant) at multipole l, observed x = C_hat / C_fid, candidate C = A C_fid."""
    nu = 2 * l + 1.0
    if kind == "exact":
        return nu * (x / A + np.log(A))
    if kind == "S":
        return nu / 2 * ((x - A) / x) ** 2
    if kind == "Q":
        return nu / 2 * ((x - A) / A) ** 2
    if kind == "D":
        return nu / 2 * ((x - A) / A) ** 2 + np.log(A)
    if kind == "f":                                  # fixed fiducial in the denominator (24)
        return nu / 2 * (x - A) ** 2
    if kind == "LN":
        return nu / 2 * np.log(x / A) ** 2
    if kind == "WMAP":
        return (1 / 3) * m2lnL("Q", x, A, l) + (2 / 3) * m2lnL("LN", x, A, l)
    if kind == "cube":
        al = -1.0
        c = ((2 * l + al) / nu) ** (1 / 3)
        return nu * 4.5 * c * ((x / A) ** (1 / 3) - c) ** 2 + (1 - al) * np.log(A)
    if kind == "HL":                                 # g(x)^2 form, eq. (44); equals exact - const
        z = x / A
        g2 = 2 * (z - np.log(z) - 1)
        return nu / 2 * g2
    raise ValueError(kind)
```

For the published curves we read the vector paths of HL's Fig. 1 directly from the PDF and convert them to data coordinates with the tick marks.

Side by side. Figure 9.49 overlays the two, and the table compares them at $\ell_{\min} = 1022$ (and the scatter also at $\ell_{\min} = 102$):

	bias $\lvert\langle A_i - A_{\text{exact}}\rangle\rvert$		scatter $\langle (A_i - A_{\text{exact}})^2\rangle^{1/2}$	
	HL Fig. 1	ours	HL Fig. 1	ours ($\ell_{\min} = 102$)
$\mathcal{L}_S$	0.0018	0.00175	0.002	0.00194 (0.0187)
$\mathcal{L}_Q$	9.100×10^{-4}	8.770×10^{-4}	0.001	9.700×10^{-4} (0.00936)
Gaussian$_D$	4.000×10^{-4}	3.890×10^{-4}	5.800×10^{-4}	5.690×10^{-4} (0.00546)
WMAP	noise	5.500×10^{-8}	5.000×10^{-6}	4.480×10^{-6} (1.350×10^{-4})
cube root	noise	6.430×10^{-8}	1.400×10^{-6}	7.510×10^{-7} (2.330×10^{-5})

In units of $1/\ell$ the biases are 1.79 ($\mathcal{L}_S$, outside the tolerance), 0.90 ($\mathcal{L}_Q$, as predicted by (9.120)) and 0.40 (Gaussian$_D$).

Why they differ.

1. *Reading precision.* The symmetric Gaussian, the quadratic form and Gaussian$_D$ agree with the paper to a few per cent, which is the precision of reading a curve off a plot.[6]

[6] For Gaussian$_D$ we used HL's eq. (25) as printed, $\mathcal{L}_Q$ plus $\ln C_\ell$ once per multipole; this is the form that reproduces their curve.

2. *Monte Carlo noise.* The WMAP and cube-root biases are of order 10^{-7}. A mean of 10^4 differences whose scatter is a few times 10^{-6} has a Monte Carlo error of $4\times10^{-6}/\sqrt{10^4} = 4\times10^{-8}$, so these biases are that noise: both curves, the paper's and ours, jitter because they measure nothing but the random draws. The scatter of the WMAP form matches.

3. *An unspecified variant.* Our cube-root scatter, 7.510×10^{-7}, is about half the published 1.400×10^{-6}. HL plot the cube-root form without saying which of their two variants ($\alpha = \pm1$ in eq. 28) is shown, and we used $\alpha = -1$. Either way it is three orders of magnitude inside the tolerance, and the conclusion is the same.

A final check. HL's own new approximation (their eqs. 41–50) is built on $g(x) = \mathrm{sign}(x-1)\sqrt{2(x-\ln x-1)}$, with $x = \widehat{C}_\ell/C_\ell$ the ratio to the model. By (9.118) with $1+u = x$, the exact $-2\ln\mathcal{L}$ minus its minimum is $\nu[(x-1)-\ln x] = \nu(x-\ln x-1)$, and $g(x)^2 = 2(x-\ln x-1)$, so

$$-2\ln\mathcal{L}_{\text{exact}} = \tfrac{\nu}{2}\,g(x)^2.$$

On the full sky the exact likelihood is therefore a Gaussian in the variable g, with variance $2/\nu$: the same quadratic shape as the shortcuts, but in a variable chosen so that it is exact. On a cut sky the variance $2/\nu$ is replaced by a covariance of the g values measured on simulations, as for the other forms. Our best fits with it equal the exact ones to 1.500×10^{-8}, the precision of the search.

The lesson. All the shortcuts are right at the peak; what separates safe from unsafe is the cubic term, and the "biases" of the safe ones in the published figure are Monte Carlo noise, which only rerunning the test reveals.

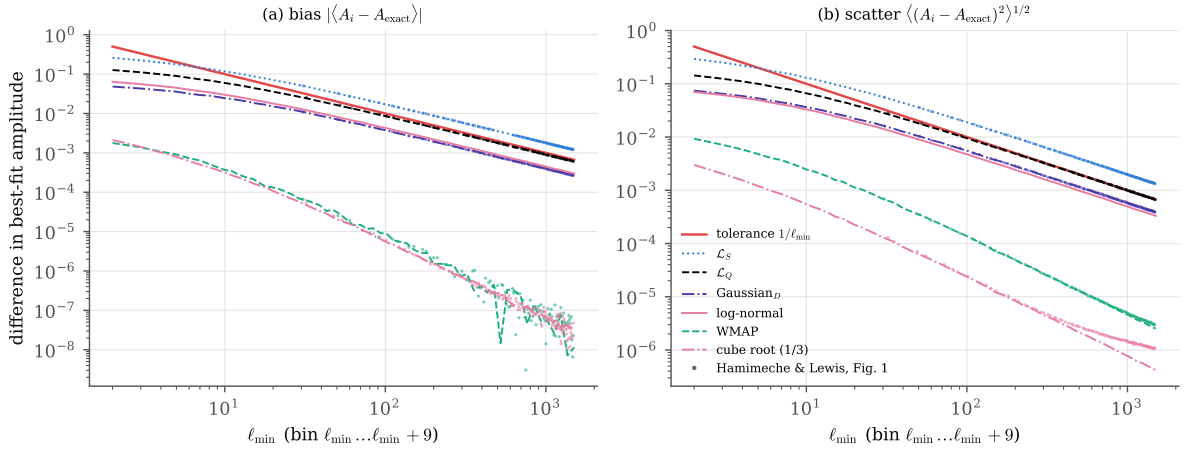


Figure 9.49: Reproduction of Fig. 1 of Hamimeche and Lewis. For bins of ten multipoles starting at $\ell_{\min}$, the bias (a) and the scatter (b) of the best-fit amplitude under each approximation, relative to the exact likelihood, from 10^4 simulated full skies per bin (lines), with the curves read from the published figure (dots) and the tolerance $1/\ell_{\min}$ (red). The symmetric $\mathcal{L}_S$ fails; $\mathcal{L}_Q$ sits on the tolerance; WMAP and the cube-root forms are far inside, and their biases are Monte Carlo noise.

What Planck does. The Planck 2018 temperature likelihood is a hybrid split at $\ell = 30$ [75]. Below it, where the skewness of Figure 9.48 is large, Planck uses the exact likelihood in a form that survives the mask and the noise [185]. Call $\boldsymbol{s}$ the full-sky signal, the sky we would see with no mask and no noise, and $\boldsymbol{d}$ the masked, noisy data. If we knew $\boldsymbol{s}$ the problem would be the one already solved: its spectrum, which we call σ_ℓ (the spectrum of a full sky, as $\widehat{C}_\ell$ was),

would enter (9.114). We do not know $\boldsymbol{s}$, so we average over what the data allow it to be. With a flat prior on C_ℓ,

$$p(C_\ell \mid \boldsymbol{d}) \overset{(1)}{=} \int p(C_\ell \mid \boldsymbol{s}, \boldsymbol{d})\, p(\boldsymbol{s} \mid \boldsymbol{d})\, \mathrm{d}\boldsymbol{s} \overset{(2)}{=} \int p(C_\ell \mid \boldsymbol{s})\, p(\boldsymbol{s} \mid \boldsymbol{d})\, \mathrm{d}\boldsymbol{s}$$
$$\overset{(3)}{=} \frac{1}{N_\mathrm{G}} \sum_{i=1}^{N_\mathrm{G}} p(C_\ell \mid \boldsymbol{s}^i) \overset{(4)}{=} \frac{1}{N_\mathrm{G}} \sum_{i=1}^{N_\mathrm{G}} \prod_{\ell=2}^{29} \frac{\mathrm{const}}{\sigma_\ell^i} \left(\frac{\sigma_\ell^i}{C_\ell}\right)^{\nu/2} e^{-\nu\sigma_\ell^i/2C_\ell}. \tag{9.121}$$

(1) The full sky is an unknown intermediate; marginalise it, with the product rule.
(2) Once the full sky is known, the masked and noisy data add nothing about its spectrum.
(3) Replace the integral over $p(\boldsymbol{s} \mid \boldsymbol{d})$ by an average over N_G samples $\boldsymbol{s}^i$ drawn from it; the average approaches the integral as N_G grows (Section 9.6).
(4) $p(C_\ell \mid \boldsymbol{s})$ depends on $\boldsymbol{s}$ only through its spectrum σ_ℓ, so it is (9.114) with $\widehat{C}_\ell \to \sigma_\ell^i$ and $\nu = 2\ell + 1$, normalised over C_ℓ: the integral $\int C_\ell^{-\nu/2} e^{-\nu\sigma/2C_\ell} \mathrm{d}C_\ell$ is proportional to $\sigma^{1-\nu/2}$ by line 1 of (9.117), so the normalised density carries the factor $\sigma^{\nu/2-1}$, which is written as $(1/\sigma)\,\sigma^{\nu/2}$. Different ℓ are independent given $\boldsymbol{s}$.

The samples $\boldsymbol{s}^i$ come from a Gibbs sampler (Section 9.16.1) that alternates between drawing a full sky given the data and the current spectrum, and drawing a spectrum given that sky; each $\boldsymbol{s}^i$ is one full sky compatible with the masked, noisy map, and σ_ℓ^i is its spectrum. The last line is Planck 2018 V, eq. 2, the Blackwell–Rao estimator: the exact full-sky likelihood averaged over what the mask and the noise hide. The factor $(\sigma_\ell^i)^{\nu/2-1}$, which a single sky could drop because it does not depend on C_ℓ, now matters, because it differs from sample to sample. Above $\ell = 30$ Planck uses a Gaussian in the cross-spectra with a covariance computed for a fixed fiducial model (their eq. 20): the multi-field, cut-sky version of $\mathcal{L}_f$. Their footnote 12 records that the symmetric shape biases n_s by about 0.1σ, and that they ignore it. Section 9.22.6 measures the bias of this choice in our idealised setting.

9.20 A model spectrum fast enough for a chain

Equation (9.113) says that every step of the chain needs the spectrum $C_\ell(\boldsymbol{\theta})$ at the proposed parameters. The honest way to get it is a Boltzmann code. CAMB [8] computes a lensed temperature spectrum to $\ell = 2500$ in about 0.9 s on the laptop. A run of four chains of 40 000 steps, the size used in Section 9.22, would then take some forty hours, and the whole point of a chain is that it can be rerun when something changes. Research codes pay this price on a cluster (Section 9.23). We replace the Boltzmann code by an **emulator** [a cheap function that reproduces the code's output over the region of parameter space the chain will visit], built once from a few dozen CAMB calls and checked against CAMB before it is trusted.

The question. How do we build a function that returns $C_\ell(\boldsymbol{\theta})$ in microseconds, and how do we know it is accurate enough? "Accurate enough" has a precise meaning: the emulator's error must change $-2\ln\mathcal{L}$ by much less than one anywhere the posterior has weight, because a change of one in $-2\ln\mathcal{L}$ is what moves a parameter by one standard deviation.

9.20.1 Expand the logarithm

We use a Taylor series around the fiducial Planck 2018 model $\boldsymbol{\theta}_0$, in six parameters $\boldsymbol{\theta} = (\ln 10^{10} A_s,\, n_s,\, H_0,\, \omega_b,\, \omega_c,\, \tau)$. The first two describe the primordial fluctuations: A_s is their amplitude at the pivot wavenumber k_* and n_s their tilt, how that amplitude changes with

scale (Section 8.7.3). H_0 is the present expansion rate. $\omega_b = \Omega_b h^2$ and $\omega_c = \Omega_c h^2$ are the physical densities of baryons and of cold dark matter, their density parameters times h^2 with $h = H_0/(100\,\mathrm{km\,s^{-1}\,Mpc^{-1}})$, which fix their actual mass per unit volume. τ is the optical depth to reionisation: a CMB photon scatters off a free electron after the Universe was reionised with probability $1 - e^{-\tau} \approx \tau$. The fiducial values, those of CAMB's input everywhere in the book, are

$\ln(10^{10}A_s)$	n_s	H_0 ($\mathrm{km\,s^{-1}\,Mpc^{-1}}$)	ω_b	ω_c	τ
3.044	0.9649	67.36	0.02237	0.1200	0.0544

with $A_s = 2.100 \times 10^{-9}$ at $k_* = 0.05\,\mathrm{Mpc^{-1}}$. The question is: a series for what? Look at how the parameters enter. The primordial power is $P(k) = A_s(k/k_*)^{n_s-1}$ with the pivot $k_* = 0.05\,\mathrm{Mpc^{-1}}$, and the spectrum is a weighted integral of it, $C_\ell = \int \mathrm{d}k\, T(\ell,k)\, P(k)$ with a transfer function $T \geq 0$ (the scalar term of the cosmology notes' $\sum_m \int T_m P_m$). Reionisation multiplies the spectrum at $\ell \gg 10$ by $e^{-2\tau}$ (cosmology notes). So, writing $x = \ln(k/k_*)$,

$$\ln C_\ell \overset{(1)}{=} \ln A_s - 2\tau + \ln \int \mathrm{d}k\, T(\ell,k)\, e^{(n_s-1)x},$$
$$\frac{\partial \ln C_\ell}{\partial n_s} \overset{(2)}{=} \frac{\int \mathrm{d}k\, T\, e^{(n_s-1)x}\, x}{\int \mathrm{d}k\, T\, e^{(n_s-1)x}} \equiv \langle x \rangle_\ell, \qquad \frac{\partial^2 \ln C_\ell}{\partial n_s^2} \overset{(3)}{=} \langle x^2 \rangle_\ell - \langle x \rangle_\ell^2 \geq 0. \tag{9.122}$$

(1) A_s and $e^{-2\tau}$ are overall factors (for $\ell \gg 10$, ignoring lensing, which depends weakly on A_s too), and $(k/k_*)^{n_s-1} = e^{(n_s-1)x}$.

(2) Differentiate the logarithm of the integral: the derivative of the exponent brings down x. The ratio is the average of x over the normalised weight $T\, e^{(n_s-1)x}$, the range of wavenumbers that feed multipole ℓ.

(3) Differentiate the ratio once more (quotient rule): the result is the variance of x under the same weight.

So $\ln C_\ell$ is exactly linear in $\ln A_s$ and τ, and nearly linear in n_s: its curvature in n_s is the variance of $\ln k$ over the wavenumbers that contribute to one multipole, a small number. The spectrum itself is *not* linear in these parameters ($C_\ell \propto A_s$ is linear, but $e^{-2\tau}$ and $e^{(n_s-1)x}$ are not). The expansion we use is therefore

$$\ln C_\ell(\boldsymbol{\theta}) \simeq \ln C_\ell(\boldsymbol{\theta}_0) + \sum_i D_{i\ell}\, \delta\theta_i + \tfrac{1}{2} \sum_{ij} H_{ij\ell}\, \delta\theta_i \delta\theta_j, \qquad D_{i\ell} = \frac{\partial \ln C_\ell}{\partial \theta_i}, \quad H_{ij\ell} = \frac{\partial^2 \ln C_\ell}{\partial \theta_i \partial \theta_j}, \tag{9.123}$$

at first or second order, with all derivatives at $\boldsymbol{\theta}_0$.

Example: the derivatives read as physics. CAMB gives, by central differences, $\partial \ln C_\ell/\partial\tau = -2.00$ averaged over $500 \leq \ell \leq 2000$: the factor $e^{-2\tau}$ exactly. It gives $\partial \ln C_\ell/\partial n_s = 0.55$ at $\ell = 1000$ and -3.61 at $\ell = 10$; by (9.122) these are average values of $\ln(k/k_*)$, so multipole 1000 is fed by wavenumbers about $e^{0.55} \approx 1.7$ times the pivot and multipole 10 by wavenumbers some forty times smaller. Somewhere between, at a multipole of a few hundred, the average is zero: that multipole sees the pivot, and changing n_s leaves it unchanged, which is the meaning of the pivot (Chapter 8). The second derivative at $\ell = 1000$ is 0.08, a spread of about 0.3 in $\ln k$. The second derivative in $\ln A_s$ is at most 0.041 for $30 \leq \ell \leq 2000$: it would be zero without lensing, whose strength also grows with A_s. Figure 9.50 plots all six first derivatives, each multiplied by that parameter's expected error so that the curves show how much one standard deviation of each parameter changes the spectrum.

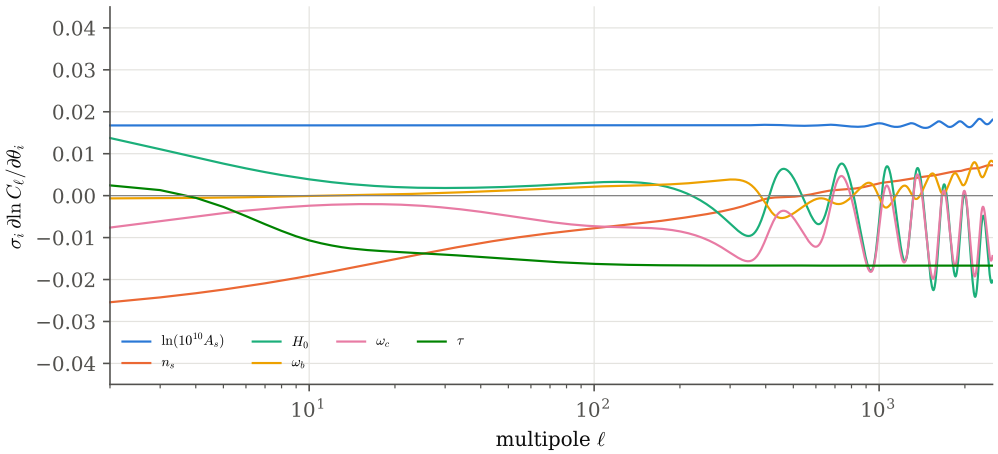


Figure 9.50: What one standard deviation of each parameter does to $\ln C_\ell$. The amplitude (blue) and the optical depth (dark green) give flat, opposite responses above $\ell \approx 30$: a temperature spectrum alone cannot separate them. The tilt n_s (orange) rotates the spectrum about a multipole of several hundred. H_0, ω_b and ω_c move and reshape the acoustic peaks, hence the oscillations.

9.20.2 Building it

The derivatives come from central differences. For one parameter with step h, $f(\theta \pm h) = f \pm hf' + \frac{1}{2}h^2 f'' \pm \frac{1}{6}h^3 f''' + \ldots$, so

$$\frac{f(\theta+h)-f(\theta-h)}{2h} \overset{(1)}{=} f' + \tfrac{1}{6}h^2 f''', \qquad \frac{f(\theta+h)-2f(\theta)+f(\theta-h)}{h^2} \overset{(2)}{=} f'' + \tfrac{1}{12}h^2 f'''', \quad (9.124)$$

(1) Subtract the two series: even powers cancel.
(2) Add them: odd powers cancel.

both with errors of order h^2. Mixed second derivatives use the four corners $(\pm h_i, \pm h_j)$. For $d = 6$ parameters that is $1 + 2d + 4\binom{d}{2} = 73$ CAMB calls, about a minute; the steps (0.02 in $\ln 10^{10}A_s$, 0.01 in n_s, 0.5 in H_0, 0.0003 in ω_b, 0.003 in ω_c, 0.01 in τ) are those of the finite-difference study of Chapter 8. Evaluating the emulator is two small matrix products:

Listing 9.13: The emulator: ln C to first or second order (`code/ch09/lib_cmbemu.py`)

```
    def cl(self, theta):
        dt = np.asarray(theta, float) - self.fid
        if self.linear_in_C:                          # Taylor series of C itself, first order
            return self.c0 * (1.0 + dt @ self.D)
        x = dt @ self.D
        if self.order == 2:
            x = x + 0.5 * np.einsum("i,ijl,j->l", dt, self.H, dt)
        out = np.exp(self.lnc0 + x)
        out[:2] = 0.0
        return out
```

It takes 54 μs, 16496 times faster than CAMB, which turns forty hours of chain into a few seconds.

9.20.3 How accurate is accurate enough?

To check the emulator we need the scale on which it must be right, and that scale is set by the data. Our target experiment is Planck-like: one channel with a 7.22′ Gaussian beam and

white noise of 33 μK arcmin, the 143 GHz values of Planck 2018 [186], on $f_{\rm sky} = 0.57$ of the sky, $2 \le \ell \le 2500$. Its noise power, deconvolved by the beam, is $N_\ell = \Delta_T^2\, e^{\ell(\ell+1)\sigma_b^2}$, with Δ_T the depth in μK rad and $\sigma_b = \mathrm{FWHM}/\sqrt{8\ln 2}$ (Chapter 8). How far does a model error δC_ℓ move $-2\ln\mathcal{L}$? Expand the exact likelihood (9.114), with $C_\ell \to C_\ell + N_\ell$ and $\nu = (2\ell+1)f_{\rm sky}$, about the true spectrum and average over data; at one multipole, with $C = C_\ell + N_\ell$,

$$\begin{aligned}\mathbb{E}\Big[-2\ln\mathcal{L}(C+\delta C)\Big] - \mathbb{E}\Big[-2\ln\mathcal{L}(C)\Big] &\overset{(1)}{=} \nu\Big[\frac{C}{C+\delta C} + \ln(C+\delta C) - 1 - \ln C\Big] \\ &\overset{(2)}{=} \nu\Big[-\epsilon + \epsilon^2 + \epsilon - \tfrac{1}{2}\epsilon^2 + O(\epsilon^3)\Big] \overset{(3)}{=} \frac{\nu}{2}\Big(\frac{\delta C}{C_\ell + N_\ell}\Big)^2 + O(\epsilon^3),\end{aligned} \tag{9.125}$$

(1) $\mathbb{E}\,\widehat{C}_\ell = C$ (the estimator is unbiased), so the data enter only through their mean.
(2) With $\epsilon = \delta C/C$: $1/(1+\epsilon) = 1 - \epsilon + \epsilon^2 - \dots$ and $\ln(1+\epsilon) = \epsilon - \epsilon^2/2 + \dots$.
(3) The linear terms cancel, because the true spectrum maximises the expected likelihood.

Summed over ℓ this is $\Delta\chi^2 = \sum_\ell \frac{\nu}{2}[\delta C_\ell/(C_\ell + N_\ell)]^2$, the change of $-2\ln\mathcal{L}$ that the emulator error alone produces. The same expansion measures a change of parameters. A small step $\delta\boldsymbol{\theta}$ changes the spectrum by $\delta C_\ell = \sum_i \partial_i C_\ell\, \delta\theta_i$, with $\partial_i \equiv \partial/\partial\theta_i$, and substituting it,

$$\Delta\chi^2 \overset{(1)}{=} \sum_\ell \frac{\nu}{2}\Big(\sum_i \frac{\partial_i C_\ell\, \delta\theta_i}{C_\ell + N_\ell}\Big)^2 \overset{(2)}{=} \sum_{ij} \delta\theta_i\, \delta\theta_j \sum_\ell \frac{\nu}{2}\,\frac{\partial_i C_\ell\, \partial_j C_\ell}{(C_\ell + N_\ell)^2} \overset{(3)}{=} \delta\boldsymbol{\theta}^{\mathsf{T}}\mathbf{F}\,\delta\boldsymbol{\theta}, \tag{9.126}$$

(1) (9.125) with $\delta C = \sum_i \partial_i C_\ell\, \delta\theta_i$.
(2) The square of a sum is a double sum, $(\sum_i a_i)^2 = \sum_{ij} a_i a_j$; the $\delta\theta$ do not depend on ℓ and come out of the sum over ℓ.
(3) Name the inner sum F_{ij}.

with the **Fisher matrix**

$$F_{ij} = \sum_\ell \frac{(2\ell+1)f_{\rm sky}}{2}\,\frac{\partial_i C_\ell\, \partial_j C_\ell}{(C_\ell + N_\ell)^2}\,\Big[+\frac{1}{\sigma_\tau^2}\ \text{for } i = j = \tau\Big], \tag{9.127}$$

which Chapter 10 develops in full. The bracket is the Gaussian prior on τ, of width $\sigma_\tau = 0.0086$, that stands in for Planck's low-ℓ polarization ("lowE") measurement [25]. A Gaussian prior adds $(\tau - \tau_{\rm obs})^2/\sigma_\tau^2$ to $-2\ln\mathcal{L}$; a step $\delta\tau$ raises that term by $\delta\tau^2/\sigma_\tau^2$, which is $\delta\boldsymbol{\theta}^{\mathsf{T}}\mathbf{F}\,\delta\boldsymbol{\theta}$ with $1/\sigma_\tau^2$ added to $F_{\tau\tau}$ alone. Without the prior a temperature spectrum cannot separate A_s from τ (Figure 9.50). The posterior near its peak is then $\propto e^{-\delta\boldsymbol{\theta}^{\mathsf{T}}\mathbf{F}\delta\boldsymbol{\theta}/2}$, a Gaussian with covariance $\mathbf{F}^{-1}$, so the error of parameter i with all the others free is $\sigma_i = \sqrt{(\mathbf{F}^{-1})_{ii}}$, not $1/\sqrt{F_{ii}}$; the second is the error with all the others held fixed, and the two differ whenever parameters are correlated (Section 10.4.1, (10.24)). Holding fixed a parameter that is in fact unknown is what produces the factor of nine in $\ln(10^{10}A_s)$ of Section 9.22. The expected errors σ_i set the natural size of a step: 0.01674 in $\ln 10^{10}A_s$, 0.005293 in n_s, 0.9365 in H_0, 0.0002014 in ω_b, 0.002157 in ω_c and 0.008346 in τ. With only $(\ln 10^{10}A_s, n_s)$ free and the rest fixed, F is the 2×2 corner and the errors shrink to 0.0018 and 0.0026, with correlation -0.73; we return to these numbers in Section 9.22.

Now the test. We compare three emulators, the series in C_ℓ itself to first order, the series in $\ln C_\ell$ to first order, and the series in $\ln C_\ell$ to second order, against direct CAMB calls, at two kinds of points: along each parameter axis at ±1, ±3 and $\pm5\sigma_i$, and at 40 random points drawn from the Fisher Gaussian with its widths doubled, which reach out to 7.3σ from the fiducial.

$\Delta\chi^2$ from the emulator error	axes, $\pm1\sigma$ (max)	axes, $\pm5\sigma$ (max)	random (median)	random (max)
first order in C_ℓ	0.054	35.5	0.0535	2
first order in $\ln C_\ell$	0.0485	33	0.0484	1.98
second order in $\ln C_\ell$	5.290×10^{-5}	0.515	6.550×10^{-4}	0.0964

With six free parameters the first-order series are not good enough: at 5σ along one axis they are wrong by tens in χ^2, and at random points within the posterior they reach $\Delta\chi^2 \approx 2$. Their errors oscillate with the acoustic peaks (Figure 9.51a): H_0, ω_b and ω_c move the peaks sideways, and a sideways shift of an oscillating function is not linear in the size of the shift. The second-order series in $\ln C_\ell$ stays below $\Delta\chi^2 = 0.1$ everywhere we tested (Figure 9.51b). With only $(\ln 10^{10} A_s, n_s)$ free, for which (9.122) promised near-linearity, the first-order series in $\ln C_\ell$ is already enough: its worst error at the four corners $(\pm5\sigma, \pm5\sigma)$ is $\Delta\chi^2 = 2.890 \times 10^{-5}$. These are the two emulators we use: first order for two or three parameters, second order for six.

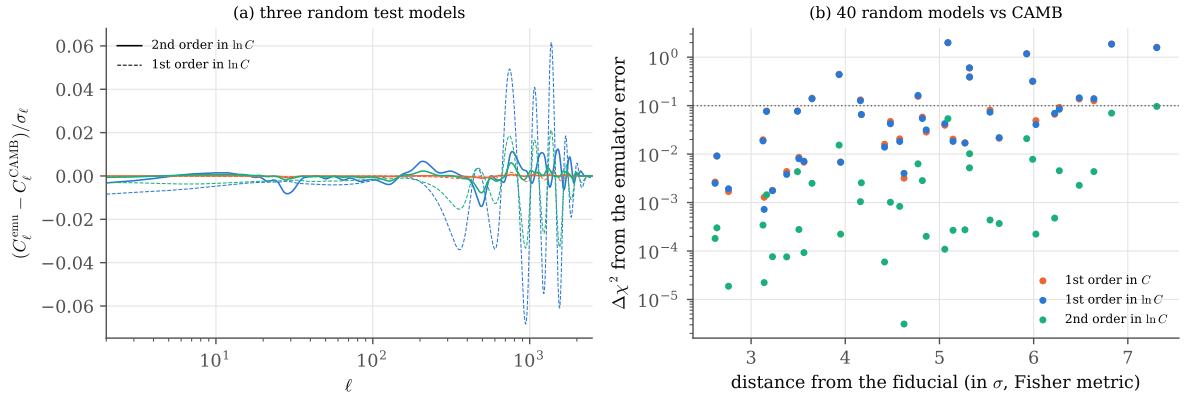


Figure 9.51: Testing the emulator against CAMB. (a) For three random models, the emulator error at each ℓ in units of the error bar of one multipole: second order in $\ln C_\ell$ (solid) against first order (dotted); the first-order errors oscillate with the acoustic peaks. (b) The $\Delta\chi^2$ produced by the emulator error at 40 random models, against their distance from the fiducial; the dotted line is $\Delta\chi^2 = 0.1$.

Example: what $\Delta\chi^2 = 0.1$ does to a posterior. An error $\Delta\chi^2$ in $-2\ln\mathcal{L}$ multiplies the posterior density at that point by $e^{-\Delta\chi^2/2}$. If the error were a constant, it would cancel in the normalisation; what matters is how it varies across the posterior. Errors that vary by 0.1 across the posterior reweight its regions by at most $e^{0.05} \approx 1.05$: a five per cent reweighting of the tails, which moves means by a few hundredths of a standard deviation. Section 9.23.1 measures this directly by reweighting the six-parameter chain with CAMB itself.

Laptop versus research. Our emulator costs 73 Boltzmann calls and is valid only within a few standard deviations of the fiducial model; it could not explore a prior much wider than the posterior or a cosmology far from Planck's. Research analyses either call CAMB or CLASS at every step, with the fast and slow parameters of Section 9.16 updated at different rates [37], or train a neural-network emulator on 10^5 Boltzmann runs spread over the whole prior, as CosmoPower does [187]. Either way the principle is the one used here: validate the fast model against the slow one on the posterior itself.

9.21 Cosmological parameters from one simulated sky

Chapter 8 built a small CMB experiment on the laptop and studied what it does to the power spectrum: a 30′ beam, white noise of 200 μK arcmin, maps at HEALPix resolution $N_{\rm side} = 256$, and a Galactic band $|b| < 20^\circ$ masked with edges tapered over 5°. It produced the sampling distribution of the measured spectrum, including a covariance of the masked bandpowers from 2000 simulated skies. Everything needed for inference is now in place: the likelihood (Section 9.19), a fast model (Section 9.20), and the sampler. We take one simulated sky, observe it three ways, and ask what each observation says about two parameters of the primordial spectrum, the amplitude $\ln(10^{10}A_s)$ and the tilt n_s, with the other four ΛCDM parameters held at their true values.

The question. The reasoning is that of Section 9.19: every candidate pair of amplitude and tilt is a hypothesis about the sky, and the measured spectrum is asked how probable it would have been under each. How does the posterior of $(\ln 10^{10}A_s, n_s)$ change from (a) a perfect, noiseless view of the whole sky, to (b) the same sky seen through the beam and the noise, to (c) the noisy sky behind the Galactic mask? And are the chains' answers the ones the Fisher matrix predicts?

9.21.1 One sky, three observations, three likelihoods

One set of $a_{\ell m}$ is drawn from the fiducial spectrum up to $\ell = 767$, and the analysis uses $2 \le \ell \le 601$ (Figure 9.52). The three data sets and their likelihoods are:

(a) The sky itself. $\widehat{C}_\ell$ is the spectrum of the $a_{\ell m}$ themselves, and the likelihood is the exact full-sky form (9.114) with $\nu = 2\ell + 1$, summed over ℓ.

(b) The full sky with beam and noise. $\widehat{C}_\ell^{\rm obs}$ is the spectrum of the observed map, and the likelihood is the same exact form with the model spectrum replaced by what the instrument should see, $T_\ell^2 C_\ell(\boldsymbol{\theta}) + N_\ell$, where T_ℓ is the beam and pixel window and N_ℓ the white-noise power (Section 9.19.2). Nothing is approximated: on the full sky with isotropic noise this is still exact.

(c) The masked noisy sky. The data are the 30 MASTER bandpowers $\widehat{D}_b$ of Chapter 8, in bins of 20 multipoles, from the masked map ($f_{\rm sky} = 0.658$ of the sky has nonzero weight). The model predicts them through the bandpower windows, $D_b(\boldsymbol{\theta}) = \sum_\ell \mathcal{F}_{b\ell} C_\ell(\boldsymbol{\theta})$, which encode what the binning, the mask deconvolution and the beam do to a theory spectrum [168]. Their distribution is no longer known exactly, so we use a Gaussian with the covariance $\widehat{\mathbf{C}}$ measured on the 2000 simulations of Chapter 8. That covariance is an unbiased estimate, but its inverse, which is what the likelihood uses, is not: inverting magnifies the directions in which the simulations happened to scatter too little, and for n simulations of p bandpowers the inverse comes out too large on average by the factor $(n-1)/(n-p-2)$ (Section 8.13.2, (8.78)). A χ^2 built with it would be too large, and the errors too small. We therefore multiply the inverse by the reciprocal, the Hartlap factor $(n-p-2)/(n-1) = 0.9845$ for $n = 2000$ and $p = 30$ [6]. With two thousand simulations for thirty bins the correction is small: every χ^2 shrinks by 1.5 per cent and every error grows by $1/\sqrt{0.9845} \approx 1.008$, under one per cent. The likelihood is

$$-2\ln\mathcal{L} = \big(\widehat{\boldsymbol{D}} - \boldsymbol{D}(\boldsymbol{\theta})\big)^{\mathsf{T}} \, \frac{n-p-2}{n-1} \, \widehat{\mathbf{C}}^{-1} \, \big(\widehat{\boldsymbol{D}} - \boldsymbol{D}(\boldsymbol{\theta})\big). \tag{9.128}$$

The covariance is fixed, computed once at the fiducial model: this is the $\mathcal{L}_f$ of Section 9.19.3 on a cut sky, the form Planck uses at high ℓ. Binning by 20 makes each bandpower an average of about $20 \times (2\ell+1) f_{\rm sky}$ modes, close to Gaussian except in the first bin or two, where the cost of the approximation is concentrated.

In code the three log-posteriors are three short functions (the prior is a wide box):

Listing 9.14: The three likelihoods of the simulated sky (`code/ch09/44_cmb_mcmc.py`)

```
def lp_full(th):
    if not inprior(th):
        return -np.inf
    c = emu.cl(th)
    return -0.5 * np.sum(nu[use] * (chat_sky[use] / c[use] + np.log(c[use])))


def lp_noisy(th):
    if not inprior(th):
        return -np.inf
    c = T2 * emu.cl(th) + Nw                          # what the instrument's spectrum
    should be
    return -0.5 * np.sum(nu[use] * (chat_obs[use] / c[use] + np.log(c[use])))


def lp_mask(th):
    if not inprior(th):
        return -np.inf
    r = Db - Fwin @ emu.cl(th)
    return -0.5 * r @ Cinv @ r
```

The model spectrum `emu.cl` is the first-order series in $\ln C_\ell$ in the two free parameters, which Section 9.20.3 showed to be accurate to $\Delta\chi^2 < 10^{-4}$ out to five standard deviations.

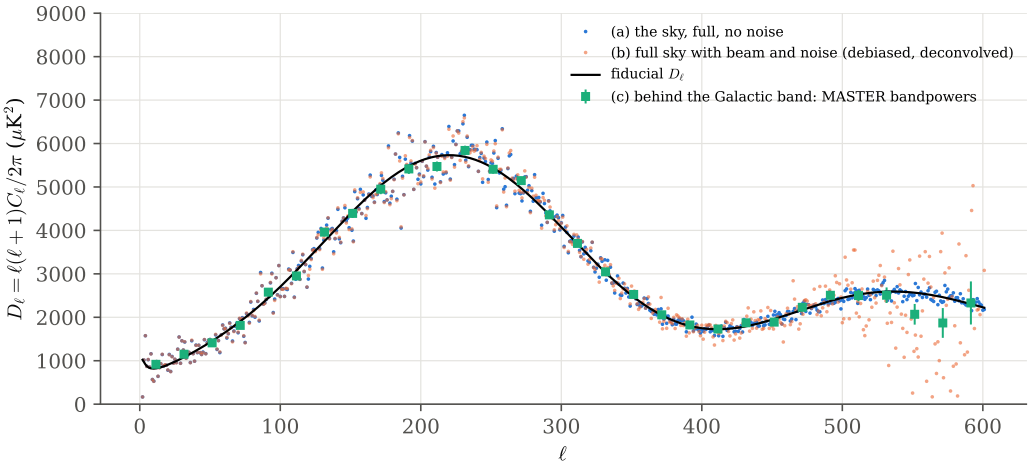


Figure 9.52: The data of the three analyses, as $D_\ell = \ell(\ell+1)C_\ell/2\pi$. (a) The spectrum of the simulated sky itself, scattering about the fiducial model (black) by cosmic variance. (b) The spectrum of the observed full-sky map, with the noise power subtracted and the beam divided out for display; above $\ell \approx 400$ the noise dominates and the scatter grows. (c) MASTER bandpowers from the masked map, with error bars from the 2000 simulations.

9.21.2 The chains and their diagnostics

For each case we run 4 chains of 20000 steps, started at random up to six expected standard deviations from the truth, with the proposal shaped by the Fisher covariance of that case and scaled by $2.38/\sqrt{2}$ (Section 9.11.2), and drop 2000 steps of burn-in. Each case takes about a second. Before reading off any number we look at the diagnostic panel that Kruschke recommends [Kruschke §8.2.5], here for the masked case (Figure 9.53): the traces of the four chains overlap after a few hundred steps; the autocorrelation falls to zero within about thirty steps; $\hat{R}$ for n_s drops below 1.01 well before the end; and the four chains' marginals agree. In

numbers: acceptance 0.36, largest $\hat{R} = 1.0006$, $\tau = 8.0$ steps, about 9336 effective samples of n_s. The other cases look the same (the table below).

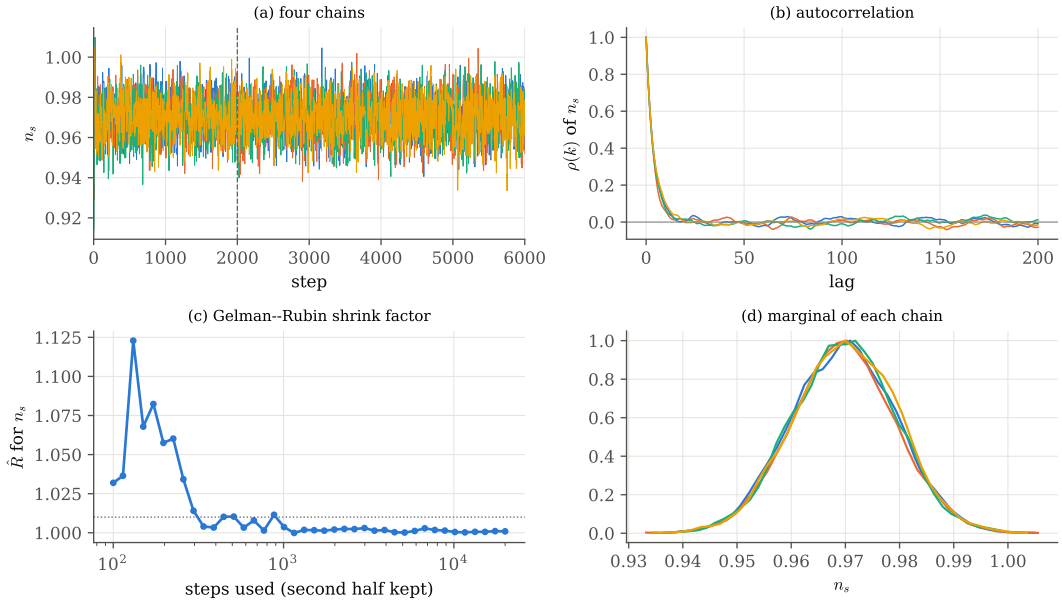


Figure 9.53: Diagnostics of the masked-sky chains, for n_s. (a) Traces of the four chains from dispersed starting points; the dashed line marks the end of burn-in. (b) Autocorrelation functions after burn-in. (c) The Gelman–Rubin $\hat{R}$ computed from the second half of the first n steps, against n; the dotted line is 1.01. (d) The marginal of n_s from each chain separately.

case	$\ln(10^{10}A_s)$	n_s	ρ	Fisher σ's, ρ	acc., $\hat{R}$	τ, ESS
(a) sky	3.0446 ± 0.0029	0.9649 ± 0.0050	0.59	0.0029, 0.0049, 0.59	0.36, 1.0013	8.4, 9471
(b) noisy	3.0400 ± 0.0063	0.9604 ± 0.0072	0.83	0.0063, 0.0072, 0.83	0.36, 1.0005	6.4, 10259
(c) masked	3.0519 ± 0.0083	0.9698 ± 0.0095	0.83	0.0082, 0.0094, 0.83	0.36, 1.0006	8.0, 9336
truth	3.0445	0.9649				

The library cross-checks agree: `emcee` on the masked likelihood gives $\ln(10^{10}A_s) = 3.0518 \pm 0.0083$ and $n_s = 0.9698 \pm 0.0093$, and `getdist` [188], the analysis package of CosmoMC and Cobaya, puts the 68 per cent interval of n_s at $[0.9603, 0.9792]$, the same as our percentiles $[0.9603, 0.9792]$.

9.21.3 Reading the triangle plot

The standard display of a posterior with several parameters is the **triangle plot** [one-dimensional marginals on the diagonal, two-dimensional marginals of each pair below it]. By the key idea of Section 9.17.4 every panel is a histogram of one or two columns of the chain. Ours is written by hand in matplotlib (`triangle` in `lib_mcmc9c.py`) so that nothing is hidden, and two choices in it need a word.

Where to draw the 68 per cent contour. A two-dimensional histogram H of the samples, smoothed slightly, estimates the density on a grid. The 68 per cent region is the set of highest density holding 68 per cent of the samples: sort the cells from the highest down, add up their contents until the total reaches 68 per cent, and the height of the last cell added is the contour level.

Listing 9.15: The contour level that encloses a given probability (`code/ch09/lib_mcmc9c.py`)

```
def hpd_levels(H, probs=(0.68, 0.95)):
    """Heights h_p of a (normalised-histogram) density H such that the region H >= h_p holds
    probability p.  Sort the cells from the highest down and accumulate their mass."""
    flat = np.sort(H.ravel())[::-1]
    cum = np.cumsum(flat) / flat.sum()
    return [flat[np.searchsorted(cum, p)] for p in probs]
```

Where to draw the Fisher ellipse. For a two-dimensional Gaussian with covariance $\mathbf{C}$, the quantity $\Delta\chi^2 = (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})^{\mathsf{T}}\mathbf{C}^{-1}(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})$ is a sum of two squared independent standard Gaussians (rotate and rescale to the principal axes), so it follows χ^2_2, whose cumulative distribution is $1 - e^{-c/2}$ (Chapter 2). The ellipse $\Delta\chi^2 = c$ encloses probability p when

$$1 - e^{-c/2} = p \quad \Longleftrightarrow \quad c = -2\ln(1-p): \qquad p = 0.683 \to c = 2.30, \qquad p = 0.954 \to c = 6.18.$$

In two dimensions the "1σ" ellipse is not $\Delta\chi^2 = 1$, which encloses only $1 - e^{-1/2} = 39$ per cent; the one-dimensional 68.3 per cent of $\pm 1\sigma$ needs $\Delta\chi^2 = 2.30$ when two parameters are shown together.

Figure 9.54 shows the three posteriors with the Fisher ellipses of each case centred on the truth.

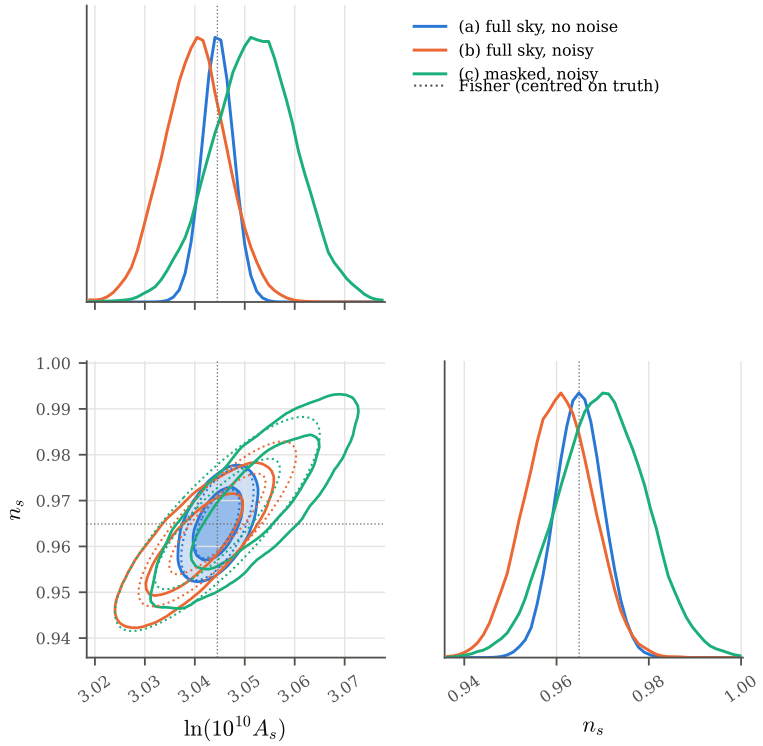


Figure 9.54: Hand-made triangle plot of $(\ln 10^{10}A_s, n_s)$ from one simulated sky observed three ways: the sky itself (filled blue), the full sky with beam and noise (red), and the noisy sky behind the Galactic mask (green). Diagonal: one-dimensional marginals. Lower panel: 68 and 95 per cent contours of the chains, with the Fisher ellipses of each case centred on the truth (dotted) and the true values (dashed lines).

9.21.4 Understanding the three posteriors

The centres. The three analyses see the same sky, so their centres are not independent draws; they move only through what the noise adds and what the mask hides. Relative to their own

widths, the noisy case sits −0.7 and −0.6 standard deviations from the truth in $\ln(10^{10}A_s)$ and n_s, the masked case +0.9 and +0.5: ordinary fluctuations.

The widths. From the sky itself to the noisy sky the error of $\ln(10^{10}A_s)$ grows from 0.0029 to 0.0063, a factor of about two, and that of n_s from 0.0050 to 0.0072, a factor of about one and a half. Where does the information go? The Fisher matrix (9.127) counts it multipole by multipole. With the beam divided out, so that N_ℓ is the deconvolved noise power, each multipole enters with the weight

$$w_\ell = \frac{\nu}{2}\left(\frac{C_\ell}{C_\ell + N_\ell}\right)^2,$$

which is the full-sky $\nu/2$ where the signal dominates and falls as $(C_\ell/N_\ell)^2$ where the noise does. For our experiment, 30′ and 200 μK arcmin, the deconvolved noise $N_\ell = \Delta_T^2 e^{\ell(\ell+1)\sigma_b^2}$ (Section 9.20.3) is below a twentieth of the signal at $\ell = 300$, equals it near $\ell \approx 460$, and is twelve times larger at $\ell = 600$. Because the number of modes $2\ell+1$ grows with ℓ, the multipoles above the crossover carried much of the total: $\sum_\ell w_\ell$, which is the amplitude entry F_{AA}, keeps only 47 per cent of its noiseless value. The tilt's entry F_{nn} keeps 97 per cent, since it weights each multipole also by the square of its lever arm in $\ln k$ (next paragraph), which is largest at the low multipoles the noise does not touch. The tilt's error still grows, through the correlation: with the high multipoles gone, amplitude and tilt can trade more freely. These weights reproduce the chains' widths to about ten per cent; the remainder comes from the pixel window and the other details of the simulation that this count leaves out.

The mask then removes a third of the sky. On a fraction $f_{\rm sky}$ of the sky each multipole has about $(2\ell+1)f_{\rm sky}$ independent modes instead of $2\ell+1$ (Knox's mode count, Section 8.4.2), so every w_ℓ, and with them the whole Fisher matrix, is multiplied by $f_{\rm sky}$, and every error by $1/\sqrt{f_{\rm sky}} = 1/\sqrt{0.658} \approx 1.23$. The chains give $0.0083/0.0063 \approx 1.32$ for both parameters. The extra seven per cent is the taper: pixels near the mask edge are kept with weights below one, so they carry less than their share of modes.

In every case the chain and the Fisher matrix agree to the second decimal. That follows from two facts already shown. By (9.122) the parameters enter $\ln C_\ell$ almost linearly, and a likelihood that is quadratic in a quantity linear in the parameters is a Gaussian in the parameters, as for the straight line; and with hundreds of multipoles each term of the sum is close to its quadratic expansion (9.125). So the posterior is nearly Gaussian, and its covariance is the inverse Fisher matrix.

The correlation: a straight line in disguise. Why is the correlation +0.59, and why does it grow to +0.83 with noise? Give a name to the derivative of (9.122): for each multipole let

$$s_\ell \equiv \frac{\partial \ln C_\ell}{\partial n_s} = \langle \ln(k/k_*) \rangle_\ell,$$

the average distance in $\ln k$ from the pivot of the wavenumbers that feed multipole ℓ (in Section 9.20.1 the integration variable $\ln(k/k_*)$ was called x; s_ℓ is its average, one number per multipole). Near the fiducial model, to first order,

$$\ln C_\ell(\boldsymbol{\theta}) \simeq \ln C_\ell(\boldsymbol{\theta}_0) + \delta \ln A_s + s_\ell\, \delta n_s,$$

a straight line in the variable s_ℓ, with intercept $\delta \ln A_s$ and slope δn_s. Its Fisher matrix follows from (9.127) in two lines:

$$\partial_i C_\ell\, \partial_j C_\ell \overset{(1)}{=} C_\ell^2 \frac{\partial \ln C_\ell}{\partial \theta_i}\frac{\partial \ln C_\ell}{\partial \theta_j} \overset{(2)}{=} C_\ell^2 \begin{pmatrix} 1 & s_\ell \\ s_\ell & s_\ell^2 \end{pmatrix}_{ij}, \qquad F \overset{(3)}{=} \sum_\ell w_\ell \begin{pmatrix} 1 & s_\ell \\ s_\ell & s_\ell^2 \end{pmatrix}. \tag{9.129}$$

(1) $\partial C_\ell = C_\ell\, \partial \ln C_\ell$.
(2) $\partial \ln C_\ell / \partial \ln A_s = 1$ and $\partial \ln C_\ell / \partial n_s = s_\ell$, from the line above.
(3) Put this into (9.127): $\frac{\nu}{2} C_\ell^2/(C_\ell + N_\ell)^2$ is the weight w_ℓ of the previous paragraph.

This is exactly the matrix of the straight-line fit (9.102), with s_ℓ in place of x_i and w_ℓ in place of $1/\sigma_i^2$. So the result of Section 9.17.2 applies unchanged:

$$\rho(\ln A_s, n_s) = -\frac{\bar{s}_w}{\sqrt{\overline{s^2}_w}},$$

with $\bar{s}_w$ and $\overline{s^2}_w$ the w-weighted means of s_ℓ and s_ℓ^2. The pivot $k_* = 0.05\,\mathrm{Mpc}^{-1}$ is the origin $s = 0$ of this line, and it falls at a multipole of a few hundred (Section 9.20.1). Our data stop at $\ell = 601$ and most of their weight is at lower multipoles, where $s_\ell < 0$: the weighted mean is negative and the correlation positive. Noise removes the high multipoles, nearest the pivot, so $\bar{s}_w$ moves further from zero and the correlation grows. A Planck-like experiment measures out to $\ell = 2500$, where $s_\ell > 0$, and the sign flips (Section 9.22). Choosing the pivot is choosing the origin of s, Kruschke's centring of Section 9.17.6 in cosmological dress: a pivot at the weighted centre of the data would make amplitude and tilt uncorrelated.

9.21.5 Do the Gaussian approximations bias n_s?

On the noiseless full sky we know the exact likelihood, so we can test the approximations of Section 9.19.3 on parameters. Running the chain with $\mathcal{L}_Q$ (variance from the model) and with $\mathcal{L}_f$ (variance from the fixed fiducial) instead of the exact form, on the same sky, gives the posteriors of Figure 9.55a: $\mathcal{L}_f$ coincides with the exact one, while $\mathcal{L}_Q$ moves n_s by -1.15 standard deviations. One sky could be a fluke, so we repeat the comparison of best fits on 500 skies drawn from the exact χ^2 distribution. Relative to the exact best fit, $\mathcal{L}_Q$ shifts n_s by -1.25 ± 0.17 standard deviations (mean and scatter over skies) and $\ln(10^{10} A_s)$ by $+0.38$; $\mathcal{L}_f$ shifts n_s by -0.000 ± 0.009 (Figure 9.55b). This is HL's statement, "$\mathcal{L}_Q$ biases parameters like the spectral index by around one sigma, $\mathcal{L}_f$ is unbiased" [16], reproduced on our experiment. The mechanism is the $0.9/\ell$ upward bias of the spectrum derived in (9.120): largest at low ℓ, so it is read as extra power on large scales, a lower n_s and a slightly higher amplitude.

The same 500 skies give a check of the exact analysis itself. Its best-fit n_s scatters about the truth with mean $+0.03$ and standard deviation 1.05 in units of the Fisher error: unbiased, and with error bars that mean what they say, to the precision of 500 skies (about ± 0.03 on the standard deviation).

9.22 Reproducing Planck 2018: the amplitude and tilt of the primordial spectrum

How did the fluctuations that grew into galaxies start out? The simplest models of the early Universe predict a primordial spectrum that is almost a power law, $P(k) = A_s (k/k_*)^{n_s - 1}$, described by two numbers: its amplitude A_s at the pivot wavenumber $k_* = 0.05\,\mathrm{Mpc}^{-1}$ and its tilt n_s (Section 8.7.3). A tilt of exactly one would mean the same strength of fluctuations on every scale; a tilt slightly below one, more on large scales, is what the simplest models of inflation predict. A CMB temperature map is the cleanest view of that spectrum we have: the sky is one realisation of $C_\ell(\boldsymbol{\theta})$, and the spectrum is a weighted integral of $P(k)$ (Section 9.20.1).

The question. How well does Planck's temperature map, with its low-multipole polarization, measure the amplitude and the tilt, and how much of the answer depends on which other

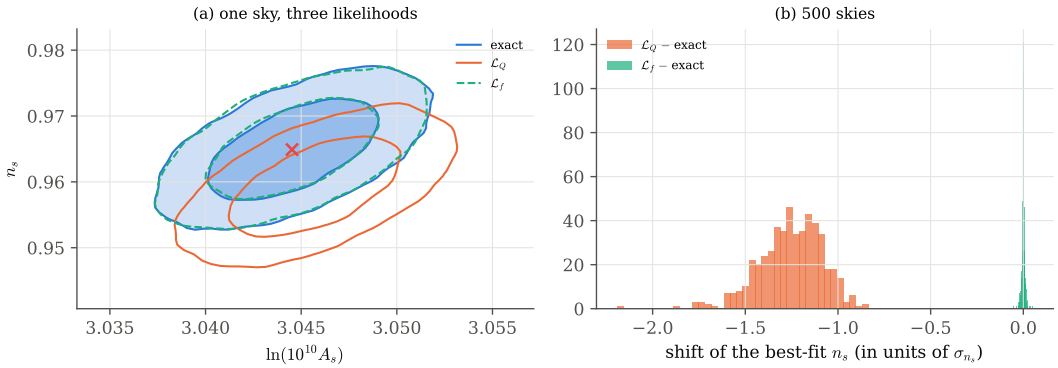


Figure 9.55: Gaussian approximations on the noiseless full sky. (a) One sky, three likelihoods: exact (filled), $\mathcal{L}_Q$ and $\mathcal{L}_f$; $\mathcal{L}_f$ lies on the exact contours, $\mathcal{L}_Q$ is displaced to lower n_s. (b) Over 500 skies, the shift of the best-fit n_s from its exact value, in units of the expected error: $\mathcal{L}_Q$ is off by more than one standard deviation in every sky, $\mathcal{L}_f$ by less than a hundredth.

parameters are left free? We cannot reproduce Planck's centres: they belong to the one real sky. We can reproduce the widths and the correlation, which depend on the experiment and on the model, not on the sky, and we can trace every difference to a cause.

9.22.1 Planck's setup and likelihood

The maps. Planck's high-ℓ temperature likelihood uses three frequency channels, 100, 143 and 217 GHz, each observed in two halves of the mission. Spectra are formed only between different halves, so the noise of one half is uncorrelated with the other and adds no power on average (Section 9.23.2).

The masks. Each channel has its own mask, removing the Galactic plane and point sources, and leaving a different fraction of the sky, because dust emission grows with frequency [75].

The foregrounds and the instrument. What remains on the sky is not only the CMB. Planck describes dust, point sources, the thermal and kinetic Sunyaev–Zeldovich effects, the cosmic infrared background and the relative calibrations of the channels with about fifteen nuisance parameters, sampled alongside the six cosmological ones.

The low multipoles. Below $\ell = 30$, the Blackwell–Rao likelihood of Section 9.19.4 for temperature, and the lowE polarization likelihood, which measures the optical depth $\tau = 0.0506 \pm 0.0086$ on its own [25].

channel	100 GHz	143 GHz	217 GHz
sky fraction of the temperature mask	0.66	0.57	0.47

range	likelihood
$2 \le \ell < 30$	Blackwell–Rao; lowE for τ
$\ell \ge 30$	Gaussian, fiducial covariance

Their key equation. Above $\ell = 30$ Planck's likelihood is a Gaussian in the vector $\widehat{C}$ of all the binned cross-spectra, with a covariance $\mathbf{\Sigma}$ computed once for a fiducial model (Planck 2018 V, eq. 20): $-2\ln\mathcal{L} = (\widehat{C} - C(\theta))^{\mathsf{T}}\mathbf{\Sigma}^{-1}(\widehat{C} - C(\theta)) + \text{const}$, where $C(\theta)$ holds the cosmological spectrum plus the foreground model, scaled by the calibrations. The symbol map to our notation is short. Their $\widehat{C}$ is our $\widehat{C}_\ell$ for one channel and no binning; their $C(\theta)$ is our $C_\ell(\theta)$

plus the noise power N_ℓ (we use one map, whose spectrum contains its noise, where Planck uses cross-spectra without it) and with no foregrounds; their $\mathbf{\Sigma}$ is diagonal for us, with the $f_{\rm sky}$ mode count of Section 9.19.2. Then

$$-2\ln\mathcal{L} \overset{(1)}{=} \sum_\ell \frac{\left(\widehat{C}_\ell - C_\ell(\boldsymbol{\theta}) - N_\ell\right)^2}{\Sigma_{\ell\ell}} \overset{(2)}{=} \sum_\ell \frac{(2\ell+1)f_{\rm sky}}{2}\left[\frac{\widehat{C}_\ell - C_\ell(\boldsymbol{\theta}) - N_\ell}{C_\ell^{\rm f} + N_\ell}\right]^2. \tag{9.130}$$

(1) One channel, one spectrum per multipole, and no coupling between multipoles: $\mathbf{\Sigma}$ is diagonal.

(2) The variance of $\widehat{C}_\ell$ is $2(C_\ell + N_\ell)^2/\nu$ with $\nu = (2\ell+1)f_{\rm sky}$ (Section 9.19.3), evaluated once at the fiducial spectrum $C_\ell^{\rm f}$.

This is the fixed-fiducial $\mathcal{L}_f$ of Section 9.19.3, with the noise added. On our idealised data the exact form is available, so our runs use it, and Section 9.22.6 measures what replacing it by (9.130) costs. What we drop, then, is the second and third channels, the mode coupling of a real mask, and all the foreground and calibration parameters.

The numbers to reproduce. With this likelihood, temperature plus lowE gives $\ln(10^{10}A_s) = 3.040 \pm 0.016$ and $n_s = 0.9626 \pm 0.0057$ (Planck 2018 VI, Table 2, column TT+lowE [25]). The covariance matrix that the Planck team distributes with its chains adds the correlation between the two, +0.03, with widths 0.0164 and 0.0057.

9.22.2 The simulated experiment and its likelihood

The spectrum. The true spectrum is computed by CAMB at the fiducial parameters (not by the emulator, so the emulator's error is part of the test). The noise is that of the 143 GHz channel, 7.22′ and 33 μK arcmin [186], and the sky fraction is $f_{\rm sky} = 0.57$, the fraction left by Planck's 143 GHz temperature mask [75]. The measured spectrum is drawn as $\widehat{C}_\ell = (C_\ell + N_\ell)\,\chi^2_\nu/\nu$ with $\nu = (2\ell+1)f_{\rm sky}$ at every $2 \le \ell \le 2500$: the distribution of Section 9.19.2 with the mode count of a partial sky.

The likelihood. For these data the $f_{\rm sky}$-scaled exact form is the true likelihood,

$$-2\ln\mathcal{L}(\boldsymbol{\theta}) = \sum_{\ell=2}^{2500}(2\ell+1)f_{\rm sky}\left[\frac{\widehat{C}_\ell}{C_\ell(\boldsymbol{\theta})+N_\ell} + \ln\big(C_\ell(\boldsymbol{\theta})+N_\ell\big)\right] + \frac{(\tau-\tau_{\rm obs})^2}{\sigma_\tau^2}. \tag{9.131}$$

The last term stands in for Planck's low-ℓ polarization likelihood (lowE), which on its own measures $\tau = 0.0506 \pm 0.0086$ [25]. We simulate that measurement too: $\tau_{\rm obs}$ is drawn from a Gaussian of width $\sigma_\tau = 0.0086$ about the true $\tau = 0.0544$, and the draw gave $\tau_{\rm obs} = 0.0714$, two standard deviations high. In code:

Listing 9.16: The Planck-like log posterior (`code/ch09/45_planck_like.py`)

```
def make_lp(names, order):
    em = ce.Emulator(order=order, names=names, lmax=LMAX)
    it = names.index("tau") if "tau" in names else None
    lo = em.fid - np.array([0.5, 0.1, 10, 0.003, 0.02, 0.06])[[ce.NAMES.index(n) for n in
    names]]
    hi = em.fid + np.array([0.5, 0.1, 10, 0.003, 0.02, 0.06])[[ce.NAMES.index(n) for n in
    names]]
    lo = np.where(np.array(names) == "tau", 0.01, lo)

    def lp(th):
        th = np.asarray(th)
        if np.any(th < lo) or np.any(th > hi):
```

```
            return -np.inf
        c = em.cl(th)[use] + Nl[use]
        out = -0.5 * np.sum(nu[use] * (chat[use] / c + np.log(c)))
        if it is not None:
            out -= 0.5 * ((th[it] - tau_obs) / SIG_TAU) ** 2
        return out
    return lp, em
```

Three runs, each asking a different question:

- P2: *if we knew everything else*, how well would the spectrum fix the amplitude and the tilt? Only $(\ln 10^{10}A_s, n_s)$ are free; H_0, ω_b, ω_c and the optical depth τ are fixed at the truth.
- P3: what changes when the reionisation optical depth τ is left free, constrained only by its lowE prior?
- P6: the full problem, with all six ΛCDM parameters free.

Every run has 4 chains of 40000 steps, a proposal shaped by the Fisher covariance, and 4000 steps of burn-in. The diagnostics of Section 9.12 are below; to keep τ for the optical depth, the integrated autocorrelation time (written τ in Section 9.17.3) is called t_{ac} here.

run	free	emulator	time (s)	acceptance	largest $\hat{R}$	t_{ac} (steps)
P2	$\ln 10^{10}A_s, n_s$	first order	5	0.35	1.0001	8
P3	$+\tau$	first order	5	0.32	1.0004	10
P6	all six	second order	12	0.28	1.0006	up to 23

All three chains can be trusted. Their widths, against Planck's:

	$\sigma(\ln 10^{10}A_s)$	$\sigma(n_s)$	$\rho(\ln A_s, n_s)$
Planck 2018 TT+lowE, Table 2	0.016	0.0057	
Planck chains' covariance	0.0164	0.0057	+0.03
P2: 2 free, others fixed	0.0018	0.0026	−0.73
Fisher	0.0018	0.0026	−0.73
P3: $+\tau$, lowE prior	0.0164	0.0026	−0.06
Fisher	0.0165	0.0026	−0.06
P6: all 6 free	0.017	0.0053	−0.14
Fisher	0.0167	0.0053	−0.15
ratio P6 / Planck	1.04	0.93	

With all six parameters free we recover Planck's widths to within ten per cent. Reading down the table, freeing τ (P2 to P3) multiplies the amplitude's error by nine and leaves the tilt's alone; freeing the other three (P3 to P6) doubles the tilt's error and leaves the amplitude's alone. The next two subsections explain each step, and the third the correlation (Figure 9.56).

9.22.3 Why the amplitude needs τ

Above $\ell \approx 30$ the temperature spectrum depends on A_s and τ only through $A_s e^{-2\tau}$ (Section 9.20.1), that is, through the combination $c = \ln(10^{10}A_s) - 2\tau$. A temperature spectrum measures c precisely and says almost nothing about $\ln A_s$ and τ separately (the opposite flat responses in Figure 9.50). In P3 the chain's column of $\ln(10^{10}A_s) - 2\tau$ has standard deviation

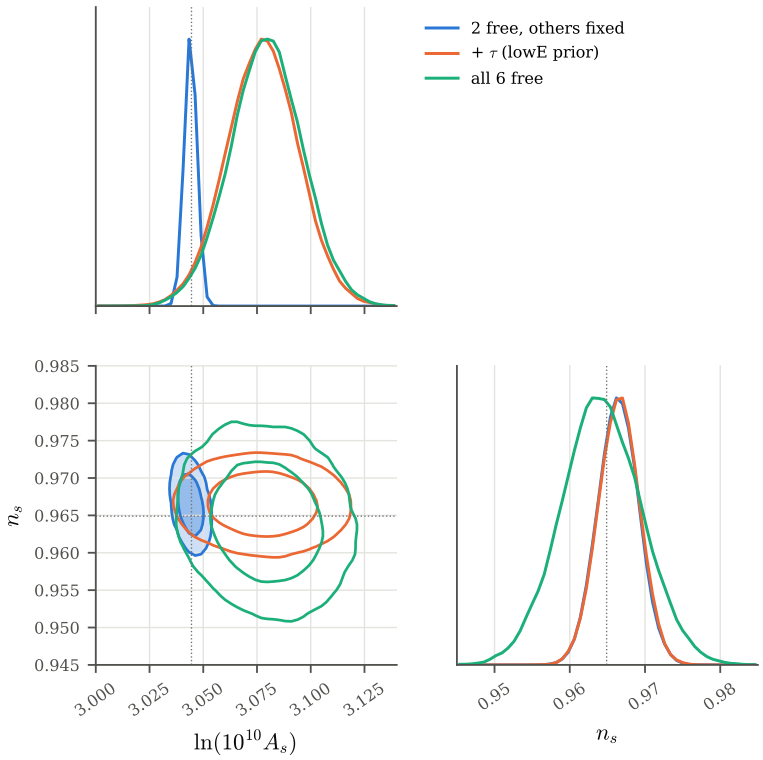


Figure 9.56: $(\ln 10^{10}A_s, n_s)$ from the Planck-like simulated spectrum, with two parameters free and the rest fixed (filled blue), with τ free under the lowE prior (red), and with all six ΛCDM parameters free (green). Dashed lines: the true values. Freeing τ stretches the amplitude ninefold; freeing H_0, ω_b and ω_c doubles the tilt's error and removes most of the correlation.

0.0018, the same as $\sigma(\ln A_s)$ in P2, where τ was fixed and the whole of c was attributed to A_s. The amplitude itself is then pinned only by lowE. Write $\ln(10^{10}A_s) = c + 2\tau$. Are c and τ independent? Change variables from $(\ln 10^{10}A_s, n_s, \tau)$ to (c, n_s, τ), a shift with unit Jacobian, so flat priors stay flat. Above $\ell = 30$ the TT term of (9.131) depends on the amplitude and the optical depth only through c; the lowE term depends only on τ. The likelihood is then a function of (c, n_s) times a function of τ, and with flat priors so is the posterior: c and τ are independent, up to the weak information TT has on τ below $\ell = 30$. The chain agrees: the correlation of its c and τ columns in P3 is -0.05. So

$$\sigma^2(\ln 10^{10}A_s) \overset{(1)}{=} \sigma_c^2 + 4\sigma_\tau^2 \overset{(2)}{=} 0.0018^2 + 4\times 0.0082^2 \quad\Rightarrow\quad \sigma(\ln 10^{10}A_s) \approx 0.0165,$$
$$\rho(\ln A_s, \tau) \overset{(3)}{=} \frac{2\sigma_\tau^2}{\sigma(\ln A_s)\,\sigma_\tau} \approx 0.99. \tag{9.132}$$

(1) The variance of a sum of independent quantities, with the factor 2 squared.
(2) $\sigma_c = 0.0018$ from the chain; $\sigma_\tau = 0.0082$, the posterior width of τ, which is the lowE width slightly narrowed by the weak information TT has on τ at $\ell < 30$.
(3) $\mathrm{Cov}(c + 2\tau, \tau) = 2\sigma_\tau^2$.

The chain gives 0.0164 and +0.99. The published 0.016 is the same calculation with Planck's $\sigma_\tau = 0.0080$: our width is larger by about the ratio of the two τ errors, $0.0082/0.0080 \approx 1.03$, which is the origin of the 1.04 of the table. The amplitude of the primordial spectrum, as measured by the CMB, is a measurement of the optical depth to reionisation.

Example: why our amplitude is high. Our six-parameter chain centres on $\ln(10^{10}A_s) = 3.0794$, while the truth is 3.0445: a shift of two standard deviations. By the same argument it must follow the simulated lowE measurement, which came out high by $\tau_{\rm obs} - \tau_{\rm true} = 0.0714 - 0.0544 = 0.017$; the posterior of τ is 0.070519, and $2 \times (0.0705 - 0.0544) = 0.032$ is the shift of $\ln A_s$. A two-sigma fluctuation of one external measurement moved the amplitude by two sigma, and TT could not object, because TT fixes only c.

9.22.4 Why the tilt needs the other parameters

With ω_b, ω_c and H_0 fixed (P3), $\sigma(n_s) = 0.0026$; with them free (P6), 0.0053, a factor 2.0. How much a parameter's error grows when others are freed can be read from the Fisher matrix (9.127). Split the parameters into n_s and the rest, r. The inverse of a matrix split this way has the Schur-complement form already used in (8.77):

$$\sigma^2(n_s) \overset{(1)}{=} (\mathbf{F}^{-1})_{nn} \overset{(2)}{=} \frac{1}{F_{nn} - \boldsymbol{F}_{nr}\mathbf{F}_{rr}^{-1}\boldsymbol{F}_{rn}} \overset{(3)}{=} \frac{1}{F_{nn}\,(1-R^2)}, \qquad R^2 \equiv \frac{\boldsymbol{F}_{nr}\mathbf{F}_{rr}^{-1}\boldsymbol{F}_{rn}}{F_{nn}}. \tag{9.133}$$

(1) The marginal error, with all the other free parameters integrated out (Section 9.20.3).
(2) Block inversion: the (n, n) element of the inverse is one over the Schur complement of $\mathbf{F}_{rr}$.
(3) Factor out F_{nn}; $1/\sqrt{F_{nn}}$ is the error with everything else fixed.

R^2 is a number between zero and one: the fraction of the effect of n_s on the spectrum, measured in the Fisher metric, that some combination of the other parameters can reproduce (the multiple correlation of the derivative curves of Figure 9.50). For our experiment $1/\sqrt{F_{nn}} = 0.0018$. With $\ln A_s$ and τ free, $R^2 = 0.53$ and $\sigma(n_s) = 0.0018/\sqrt{0.47} = 0.0026$; with all five free, $R^2 = 0.89$ and $\sigma(n_s) = 0.0018/\sqrt{0.11} = 0.0053$, the chains' values. A doubling of the error is a factor of four in $1 - R^2$: of the part of n_s's effect that amplitude and optical depth cannot mimic, H_0, ω_b and ω_c together mimic three quarters.

Which of them? Freed one at a time (with $\ln A_s$ and τ), each moves $\sigma(n_s)$ by less than four per cent: 0.0027 for H_0 or ω_b, 0.0026 for ω_c. Freed together, H_0 and ω_c give 0.0049, and adding ω_b completes the 0.0053. The physical reading follows from these numbers. Changing H_0 or ω_c alone moves the acoustic peaks sideways in ℓ, which the spectrum measures far too well to allow; changing both together, in the proportion that keeps the angular size of the sound horizon fixed, leaves the peaks in place and changes the matter density, which alters the driving of the oscillations and so the heights of the first peaks relative to the later ones. A change of relative heights across ℓ is what a tilt does, so n_s becomes partly degenerate with that combination; the baryon density, which raises the odd peaks against the even ones, takes a smaller share. This is the pitfall of Section 9.18.2 on a cosmological scale: a nuisance parameter fixed at its true value hides an uncertainty, and the chain cannot tell. The remaining seven per cent between our 0.0053 and Planck's 0.0057 is what we left out (Section 9.22.1): the fifteen foreground and calibration parameters that Planck's official TT chains sample alongside the six (they are listed in the covariance file we used), and three frequency channels with different noise, beams and masks [75]. By (9.133), each nuisance parameter that can mimic part of the effect of n_s at high ℓ widens it a little. Our one clean channel with no foregrounds is slightly too optimistic.

The other widths tell the same story (Figure 9.57b): H_0 to ± 0.93 against Planck's 0.92, ω_b to 0.0002 against 0.00022, ω_c to 0.0022 against 0.0021, τ to 0.0083 against 0.0080, all within about ten per cent.[7]

[7] Planck samples 100 $\theta_{\rm MC}$, the angular size of the sound horizon, where we sample H_0. The marginal of $(\ln A_s, n_s)$ does not care how the integrated-out parameters are labelled: a change of variables $\phi \to \psi(\phi)$ inside $\int p(\boldsymbol{\theta}, \phi)\,\mathrm{d}\phi$

9.22.5 The correlation, and why the pivot is at $0.05\,\mathrm{Mpc}^{-1}$

With the other parameters fixed the correlation is strongly negative, -0.73, the opposite sign to the toy experiment of Section 9.21.4, and the straight-line formula $\rho = -\bar{s}_w/\sqrt{\overline{s^2}_w}$ of that section explains the sign. Planck's information extends to $\ell \approx 2000$, beyond the multipole of a few hundred that sees the pivot, so the weighted mean of $s_\ell = \langle \ln(k/k_*)\rangle_\ell$, the lever arm of multipole ℓ in $\ln k$, is positive and $\rho < 0$; the toy stopped at $\ell = 601$, where $s_\ell < 0$, and had $\rho > 0$. With τ free the correlation drops to -0.06, with all six free to -0.14 (Fisher -0.15), and Planck finds $+0.03$. Why does it nearly vanish?

Moving the pivot. Section 8.7.3 answered this with two lines. The same spectrum read at another wavenumber k', a distance $L = \ln(k'/k_*)$ away in $\ln k$, has amplitude $\ln A(k') = \ln A_s + (n_s - 1)L$, by $P(k) = A_s(k/k_*)^{n_s-1}$; its covariance with the tilt is

$$\mathrm{Cov}\big(\ln A(k'), n_s\big) = \mathrm{Cov}(\ln A_s, n_s) + L\ \mathrm{Var}(n_s), \tag{9.134}$$

zero at $L_{\mathrm{dec}} = -\mathrm{Cov}(\ln A_s, n_s)/\mathrm{Var}(n_s)$ (8.46). Dividing by the two widths, the correlation at the pivot itself is

$$\rho(\ln A_s, n_s) = \frac{\mathrm{Cov}(\ln A_s, n_s)}{\sigma(\ln A_s)\,\sigma(n_s)} = -L_{\mathrm{dec}}\,\frac{\sigma(n_s)}{\sigma(\ln A_s)}. \tag{9.135}$$

So the correlation at k_* is small if the decorrelation wavenumber is close to k_*, *or* if the amplitude's error is large compared with L_{dec} times the tilt's.

The numbers from our chains. In P2, $\sigma(\ln A_s) = 0.0018$, $\sigma(n_s) = 0.0026$ and $\rho = -0.73$ give $L_{\mathrm{dec}} = 0.50$, so $k_{\mathrm{dec}} = 0.05\,e^{0.50} = 0.082\,\mathrm{Mpc}^{-1}$, the value Section 8.7.3 found for a Planck-like experiment by Fisher matrix. In P6, $L_{\mathrm{dec}} = 0.43$ and $k_{\mathrm{dec}} = 0.077\,\mathrm{Mpc}^{-1}$: freeing the other parameters hardly moves the decorrelation scale. What changes is the ratio in (9.135): $0.0026/0.0018 = 1.5$ in P2 but $0.0053/0.017 = 0.32$ in P6, so $\rho = -0.50 \times 1.5 = -0.73$ becomes $-0.43 \times 0.32 = -0.14$. The drop is the factor of nine in the amplitude's error of Section 9.22.3: the optical depth adds the same uncertainty $2\sigma_\tau$ to $\ln A$ at every pivot, and that common part is uncorrelated with the tilt. This is also why P3, with only τ freed, already has $\rho = -0.06$. Planck's own $+0.03$ corresponds to $L_{\mathrm{dec}} = -0.03 \times 0.0164/0.0057 \approx -0.09$, a k_{dec} near $0.046\,\mathrm{Mpc}^{-1}$; their extra nuisance parameters, which absorb high-ℓ power, plausibly move the weighted centre of the tilt's information to lower ℓ, but with ρ this small the difference is within what the two analyses' different ingredients can produce.

A pivot far away. The same formula prices the alternative. At WMAP's pivot, $k' = 0.002\,\mathrm{Mpc}^{-1}$, $L = \ln(0.002/0.05) = -3.2$, and our P6 chain gives $\rho\big(\ln A(k'), n_s\big) = -0.76$ with $\sigma\big(\ln A(k')\big) = 0.025$, half as large again as at k_*; in P2 the correlation there is -0.99. By the centring argument of Section 9.17.6, the amplitude read far from the data's weighted centre swings with every error in the slope.

That is the reason for the convention. The pivot $k_* = 0.05\,\mathrm{Mpc}^{-1}$ defines what A_s means, the amplitude of the primordial spectrum at that wavenumber. Any other choice describes the same spectrum, but a pivot near the centre of the data's information, within a factor of two of k_{dec} in either direction, makes the amplitude and the tilt nearly independent, which makes them easier to quote, to combine and to sample.

9.22.6 Planck's Gaussian approximation, tested

Planck's high-ℓ likelihood is a Gaussian with a fiducial covariance (Section 9.19.4), with the remark that its symmetric shape biases n_s by about 0.1σ [75]. In our setting we can measure

only multiplies the prior by the Jacobian $\left|\mathrm{d}\phi/\mathrm{d}\psi\right|$, so flat priors in different bases differ slightly, and with Planck's data the likelihood dominates.

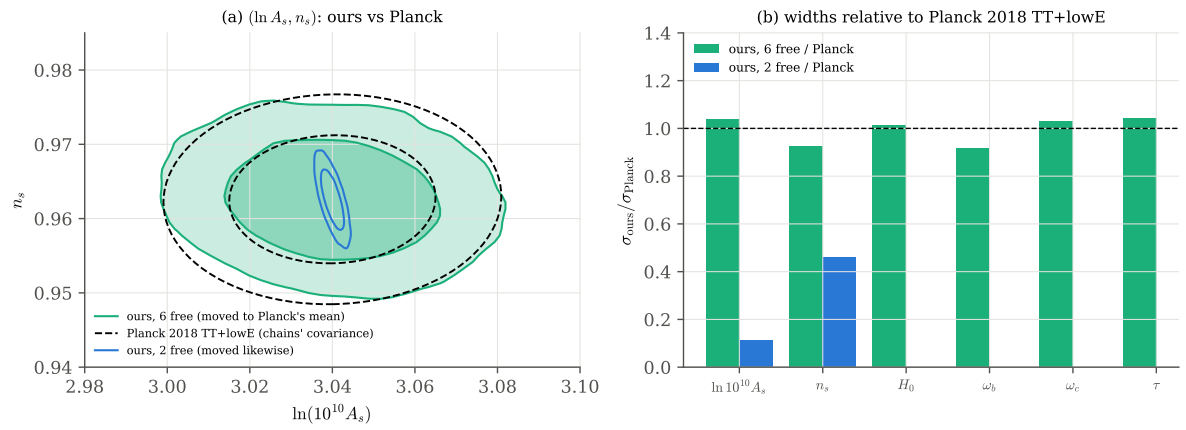


Figure 9.57: Ours against Planck 2018 TT+lowE. (a) The 68 and 95 per cent contours of our six-parameter run (filled green) and two-parameter run (blue), both shifted to Planck's centre since the centres belong to different skies, with the Gaussian of the official chains' covariance (black dashed). (b) Posterior widths of all six parameters divided by Planck's: with everything free (green) they are within about ten per cent; fixing the other four (blue) makes $\ln A_s$ nine times and n_s twice too precise.

the effect of that choice, because we know the exact likelihood. For 300 simulated Planck-like spectra we find the best-fit $(\ln A_s, n_s)$ under the exact likelihood (9.131) (others fixed), under a hybrid that keeps it at $\ell < 30$ and uses a Gaussian with the fixed fiducial variance $2(C_\ell^{\rm f} + N_\ell)^2/\nu$ above, and under the same hybrid with $\mathcal{L}_Q$ above $\ell = 30$. In units of $\sigma(n_s)$, the fixed-fiducial Gaussian shifts n_s by $+0.000 \pm 0.000$ on average (scatter 0.004 between spectra); $\mathcal{L}_Q$ shifts it by -1.307 ± 0.007.

The second number is HL's warning again, now at Planck's sensitivity: the variance must not be taken from the model being tested. The first says that in an idealised experiment the fixed-fiducial Gaussian costs nothing in the best fit, as HL's eqs. (38)–(39) predict. Planck's 0.1σ therefore comes from what our test leaves out: real cut-sky pseudo-spectra whose distribution is not a scaled χ^2 at $30 \le \ell < 100$, a fiducial model that is not exactly the truth, and posterior means from chains rather than best fits (the exact posterior is slightly skewed at low ℓ, the Gaussian is not). We have not reproduced that 0.1σ; we have shown that it is not a property of the Gaussian form itself, and that a tenth of a standard deviation is the scale of effect these choices produce.

quantity	Planck 2018	ours, and why they differ
$\sigma(\ln 10^{10}A_s)$	0.016	0.017: set by σ_τ, ours 0.0083, theirs 0.0080
$\sigma(n_s)$	0.0057	0.0053: no foreground or calibration nuisance parameters
$\rho(\ln A_s, n_s)$	+0.03	−0.14: same reason; small either way (pivot)
$\sigma(n_s)$, others fixed	not quoted	0.0026: a fixed parameter hides uncertainty
bias of the high-ℓ Gaussian on n_s	$\approx 0.1\sigma$ (V, fn. 12)	$+0.000\sigma$ in the ideal case; -1.307σ for $\mathcal{L}_Q$

9.23 How it is done in research

A published constraint is the end of a long pipeline, and the chain is only one stage of it. Once a chain exists, one trick that research pipelines use constantly takes a few lines: reweight the chain instead of rerunning it whenever the data or the theory change a little. Around the chain

sit the data products, the covariance, the likelihood code and the checks, and each of them has a laptop-sized counterpart in the analyses above.

9.23.1 Importance sampling: reusing a chain

A chain is a sample from one posterior, $p_{\text{old}}(\boldsymbol{\theta})$. Often we want averages over a slightly different one, $p_{\text{new}}(\boldsymbol{\theta})$: the same data with a more accurate theory code, or the same data plus a new measurement. The identity of Chapter 6 does it without a new run [Padilla eqs. 45–47]: for any function f,

$$\mathbb{E}_{\text{new}}[f] \overset{(1)}{=} \int f(\boldsymbol{\theta})\,\frac{p_{\text{new}}(\boldsymbol{\theta})}{p_{\text{old}}(\boldsymbol{\theta})}\,p_{\text{old}}(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \overset{(2)}{=} \frac{\sum_t w_t f(\boldsymbol{\theta}_t)}{\sum_t w_t}, \qquad w_t = \frac{\mathcal{L}_{\text{new}}(\boldsymbol{\theta}_t)}{\mathcal{L}_{\text{old}}(\boldsymbol{\theta}_t)}. \tag{9.136}$$

(1) Multiply and divide by p_{old}; this needs $p_{\text{old}} > 0$ wherever $p_{\text{new}} > 0$.
(2) The integral is an average over p_{old}, which the chain samples. The posteriors are known only up to their evidences, so we normalise the weights by their sum (the self-normalised estimator); with the same prior in both, the prior cancels in $p_{\text{new}}/p_{\text{old}}$ and only the likelihood ratio remains.

The price is precision. If N independent draws carry weights w_t, the weighted mean of f has variance $\sigma_f^2 \sum_t w_t^2/(\sum_t w_t)^2$ (the variance of $\sum_t a_t f_t$ with $a_t = w_t/\sum w$ is $\sigma_f^2 \sum a_t^2$), the same as an unweighted mean of

$$N_{\text{eff}} = \frac{\left(\sum_t w_t\right)^2}{\sum_t w_t^2} \tag{9.137}$$

draws, the **effective sample size** of Chapter 6. Equal weights give $N_{\text{eff}} = N$; one dominant weight gives $N_{\text{eff}} \approx 1$. Lewis and Bridle made this a standard part of the CosmoMC workflow [37]: run the chain with a fast approximation, then reweight with the expensive calculation, or add a data set by multiplying in its likelihood.

Example: checking the emulator on the posterior itself. The six-parameter chain of Section 9.22 used the second-order emulator. We pick 400 of its 144000 samples at random, recompute the likelihood with CAMB at each (3.56 s per call, 1422 s in all), and form $w_t = \mathcal{L}_{\text{CAMB}}/\mathcal{L}_{\text{emu}}$. The values of $-2\ln w_t$ scatter with standard deviation 0.044 and never deviate from their mean by more than 0.32 (Figure 9.58a): the effective sample size is 400 of 400, and the reweighted means move by at most 0.00 standard deviations. This is the check promised in Section 9.20.3, done where it matters, on the posterior. Reweighting the whole chain would take 142 hours of CAMB on the laptop, which is why we subsample; a cluster does it in parallel.

Where we have met H_0 before. The supernovae of Section 3.17.4 fixed only $M - 5\log_{10} H_0$; the distance ladder of Section T2.12 supplied M, and the real ladder gives 73.04 ± 1.04, against 67.36 ± 0.54 from Planck's full CMB data, a difference of 4.8σ between two independent Gaussians (Section T2.12.3). The ladder value is the same here, and so is the question, whether the CMB and the ladder can be describing one H_0. Two things are different. The CMB side is our own chain, run P6 of Section 9.22, which sampled H_0 directly from one simulated temperature spectrum and the τ prior. And the ladder enters not as a second number to subtract but as a likelihood that multiplies the posterior, the Bayesian combination of Chapter 5. The numbers differ for one reason: our chain has the width of temperature data alone, 0.93, close to Planck's 0.92 for TT+lowE, where the full data with polarization and lensing give 0.54, and its centre, 67.06, is one draw of a simulated sky about the fiducial 67.36. The same gap measured with a wider error bar is fewer standard deviations, 4.3 instead of 4.8.

Example: adding a measurement of H_0. Our CMB chain says $H_0 = 67.06 \pm 0.93\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$. Suppose an independent experiment reports a Gaussian measurement $H \pm s_H$. The new posterior is the old one times $e^{-(H_0-H)^2/2s_H^2}$, so the weights are exactly that factor. Two cases (Figure 9.58b, c):

- a compatible measurement, 67.0 ± 1.0: $N_{\mathrm{eff}} = 127474$, 88.5 per cent of the chain, and the reweighted $H_0 = 67.02 \pm 0.68$ agrees with a fresh chain run on the combined posterior, 67.03 ± 0.68;
- the SH0ES distance-ladder value 73.04 ± 1.04 [24], 4.3 standard deviations from the CMB value: $N_{\mathrm{eff}} = 634$, 0.4 per cent of the chain. The reweighted $H_0 = 69.74 \pm 0.64$ against the fresh chain's 69.79 ± 0.71. How far can we trust the reweighted numbers? Reweight 8 pieces of the chain separately and see how much they scatter: the centre is uncertain by about 0.083 and the width by about 0.027, against 0.007 and 0.0033 in the compatible case. The centre agrees with the fresh chain, but the width comes out narrower by about twice its own uncertainty. The combined posterior lives in the far tail of the old one, where the chain has few samples, and the samples that are missing are the ones that would widen it.

How small should we expect N_{eff} to be? Take the simplest case: the old posterior of H_0 is a Gaussian $\mathcal{N}(\mu_0, s_0^2)$, and the new measurement multiplies it by $\mathcal{L}(h) = e^{-(h-H)^2/2s_H^2}$, so the weights are $w = \mathcal{L}$. For a long chain the sums in (9.137) become averages over the old posterior, $N_{\mathrm{eff}}/N \to (\mathbb{E}_{\mathrm{old}}\,\mathcal{L})^2/\mathbb{E}_{\mathrm{old}}[\mathcal{L}^2]$, and we need the average of a power $\mathcal{L}^k$ for $k = 1$ and $k = 2$. With $\Delta = \mu_0 - H$ the offset between the old centre and the new measurement,

$$\mathbb{E}_{\mathrm{old}}\big[\mathcal{L}^k\big] \overset{(1)}{=} \int \frac{e^{-(h-\mu_0)^2/2s_0^2}}{\sqrt{2\pi}\,s_0}\, e^{-k(h-H)^2/2s_H^2}\,\mathrm{d}h \overset{(2)}{=} \sqrt{2\pi}\,\frac{s_H}{\sqrt{k}} \int \mathcal{N}\big(h;\mu_0,s_0^2\big)\,\mathcal{N}\big(h;H,s_H^2/k\big)\,\mathrm{d}h$$
$$\overset{(3)}{=} \sqrt{2\pi}\,\frac{s_H}{\sqrt{k}}\,\frac{e^{-\Delta^2/2(s_0^2+s_H^2/k)}}{\sqrt{2\pi(s_0^2+s_H^2/k)}} \overset{(4)}{=} \sqrt{\frac{s_H^2}{s_H^2+ks_0^2}}\; e^{-k\Delta^2/2(s_H^2+ks_0^2)}. \tag{9.138}$$

(1) The average of $\mathcal{L}^k$ over the old Gaussian.
(2) $\mathcal{L}^k$ is a Gaussian in h of variance s_H^2/k, missing its normalisation $1/(\sqrt{2\pi}\,s_H/\sqrt{k})$; put it in and compensate outside.
(3) The integral of a product of two normal densities in h is a normal density in the difference of their centres, with the two variances added (complete the square in h, or read it as the density of the sum of two independent Gaussians).
(4) Cancel $\sqrt{2\pi}$, and multiply numerator and denominator of the variance by k.

Now form the ratio with $k = 1$ in the numerator (squared) and $k = 2$ in the denominator:

$$\frac{N_{\mathrm{eff}}}{N} \overset{(1)}{=} \frac{\dfrac{s_H^2}{s_H^2+s_0^2}\,e^{-\Delta^2/(s_H^2+s_0^2)}}{\sqrt{\dfrac{s_H^2}{s_H^2+2s_0^2}}\,e^{-\Delta^2/(s_H^2+2s_0^2)}} \overset{(2)}{=} \frac{s_H\sqrt{s_H^2+2s_0^2}}{s_H^2+s_0^2}\,\exp\!\left[-\frac{\Delta^2}{s_H^2+s_0^2}+\frac{\Delta^2}{s_H^2+2s_0^2}\right]. \tag{9.139}$$

(1) Square (9.138) at $k = 1$: the square root goes and the exponent doubles to $-\Delta^2/(s_H^2+s_0^2)$. At $k = 2$ the exponent is $-2\Delta^2/2(s_H^2+2s_0^2) = -\Delta^2/(s_H^2+2s_0^2)$.
(2) Prefactor: $\frac{s_H^2}{s_H^2+s_0^2}\cdot\frac{\sqrt{s_H^2+2s_0^2}}{s_H}$; exponent: subtract the denominator's.

The prefactor is the cost of the narrower width alone, the exponential the cost of the offset. Put in the numbers, with $s_0 = 0.93$ from our chain. *The compatible case:* $s_H = 1$ and $\Delta \approx 0$, so the

exponential is one and the prefactor is $1 \times \sqrt{1 + 2 \times 0.865}/(1 + 0.865) = 1.652/1.865 = 0.89$, as found. *The SH0ES case:* $s_H = 1.04$, $s_H^2 = 1.082$, $\Delta = 6.0$. The two variances are $s_H^2 + s_0^2 = 1.946$ and $s_H^2 + 2s_0^2 = 2.811$, the exponent is $-36/1.946 + 36/2.811 = -18.50 + 12.81 = -5.7$, and the prefactor $1.04 \times \sqrt{2.811}/1.946 = 0.90$; together $0.90 \times e^{-5.7} = 0.003$, or 0.3 per cent, against the 0.4 ± 0.1 per cent measured (the error again from the scatter between pieces of the chain). The fraction falls like a Gaussian in the offset Δ: reweighting is safe for new data that are "broadly consistent" with the old posterior, as Padilla et al. put it, and fails quickly for data in tension, where a new run is needed.

The same weights give an evidence ratio. Write the old posterior with its normalisation, the evidence $Z_{\text{old}} = \int \mathcal{L}_{\text{old}}\, \pi \, \mathrm{d}\boldsymbol{\theta}$ of Chapter 5, as $p_{\text{old}} = \mathcal{L}_{\text{old}}\, \pi / Z_{\text{old}}$, with π the prior common to both analyses. The plain average of the weights over the old chain is then

$$\mathbb{E}_{\text{old}}[w] \overset{(1)}{=} \int \frac{\mathcal{L}_{\text{new}}}{\mathcal{L}_{\text{old}}} \frac{\mathcal{L}_{\text{old}}\, \pi}{Z_{\text{old}}} \, \mathrm{d}\boldsymbol{\theta} \overset{(2)}{=} \frac{\int \mathcal{L}_{\text{new}}\, \pi \, \mathrm{d}\boldsymbol{\theta}}{Z_{\text{old}}} \overset{(3)}{=} \frac{Z_{\text{new}}}{Z_{\text{old}}}. \tag{9.140}$$

(1) The definition of the weight, averaged over the normalised old posterior.
(2) $\mathcal{L}_{\text{old}}$ cancels, wherever it is not zero: as for (9.136), p_{old} must be positive wherever p_{new} is.
(3) The numerator is the evidence of the new analysis.

So $\frac{1}{N} \sum_t w_t$ estimates the ratio of the evidences of the two analyses [Padilla eq. 48], the Bayes factor of Section 9.18.3, without a new integral. Its precision is governed by the same spread of weights: when a few samples carry all the weight, N_{eff} is small and the average of the w_t is as unreliable as the reweighted means, so a small N_{eff} also means an untrustworthy evidence ratio.

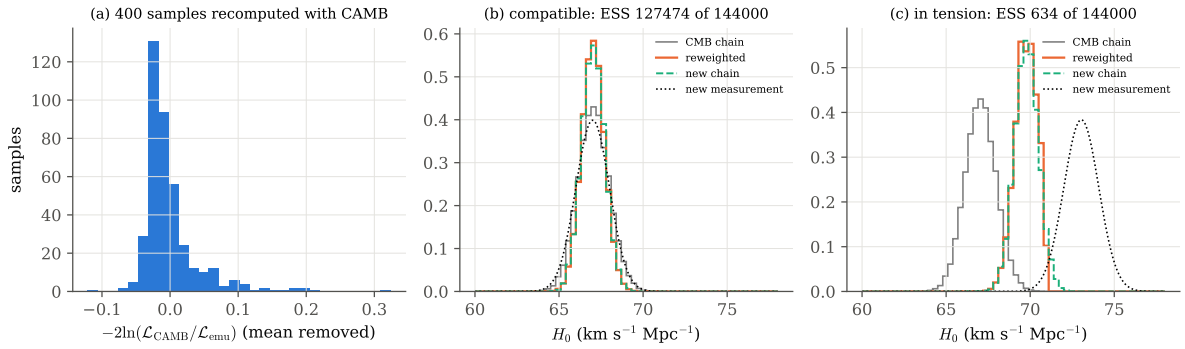


Figure 9.58: Importance sampling a six-parameter CMB chain. (a) $-2\ln(\mathcal{L}_{\text{CAMB}}/\mathcal{L}_{\text{emu}})$ at 400 random samples: a spread of a few hundredths, so the emulator chain needs no correction. (b) Adding a compatible H_0 measurement (dotted): the reweighted chain (red) and a new chain (dashed) agree. (c) Adding a measurement in tension: only 0.4 per cent of the samples carry weight, and the reweighted histogram is noisy and too narrow.

9.23.2 The research pipeline, stage by stage

A CMB temperature-and-polarization analysis of the Planck kind runs in seven stages [25, 75]. Each one produces something definite and is checked in a definite way; the laptop counterpart of every stage is drawn in Figure 9.59 and its cost is tabulated after it.

1. *Data products.* Maps at several frequencies (for Planck's high-ℓ likelihood, 100, 143 and 217 GHz), split into halves of the mission so that cross-spectra have no noise bias; a mask for each frequency (Galactic and point sources); beam window functions; and simulations: 1000 sky realisations and 300 end-to-end simulations of the full instrument and processing (FFP10).

2. *Spectra and covariance.* Cross-spectra of the half-mission maps for every frequency pair, corrected for the mask by a MASTER-type deconvolution, binned (Chapter 8); their covariance computed analytically for a fiducial model and corrected with simulations.

3. *Likelihood.* A hybrid: below $\ell = 30$, the Blackwell–Rao temperature likelihood from Gibbs samples of the sky and a simulation-based polarization likelihood (lowE); above, the Gaussian of Section 9.19.4 with the fiducial covariance, and a model of foregrounds and calibrations with about fifteen nuisance parameters for temperature alone.

4. *Theory.* A Boltzmann code (CAMB or CLASS) called by the sampler for each proposed cosmology, with the nuisance parameters updated more often than the cosmological ones because they are cheap [37].

5. *Sampling.* Metropolis–Hastings with a proposal covariance learned from earlier runs, several chains, and a stopping rule such as $\hat{R} - 1 < 0.01$; a nested sampler when the evidence is needed. The codes are CosmoMC [37], its successor Cobaya [189], and MontePython [190].

6. *Products.* The chains themselves, with tables of marginal means and limits and triangle plots made by `getdist` [188], and the parameter covariance, which is the file we compared with in Section 9.22.

7. *Checks.* The pipeline is run on simulations with known inputs and must recover them on average (the 300 FFP10 simulations); independent likelihood codes (Plik and CamSpec) must agree; results from different multipole ranges and frequency combinations must be consistent; null spectra must be compatible with noise; and importance sampling adds other data (CMB lensing, BAO) or checks the accuracy settings of the Boltzmann code.

How this is used in real research: verification and research

The straight line, the decaying source and the noiseless full sky were *verifications*: the posterior was known exactly (a Gaussian, a grid, a Student t, an exact likelihood) and the chain had to reproduce it, which it did to the precision its effective sample size allows. The masked sky and the six-parameter Planck-like run are the *research* mode: no closed-form posterior exists, the covariance comes from simulations, and the chain with its diagnostics is the instrument that produces the constraint. In real analyses the loop is exactly this one, with a Boltzmann code in place of the emulator: CosmoMC and Cobaya [37, 189] and MontePython [190] propose a cosmology, call CAMB or CLASS, evaluate the Planck likelihood code, and accept or reject with the Metropolis rule; the chains are analysed with `getdist`. Fisher forecasts (Chapter 10) are checked against such chains, and the same machinery, fed with simulated covariances of new summaries, will constrain parameters with topological statistics in Chapter 12.

What each laptop reduction costs. Every reduction trades something for speed, as the table shows.

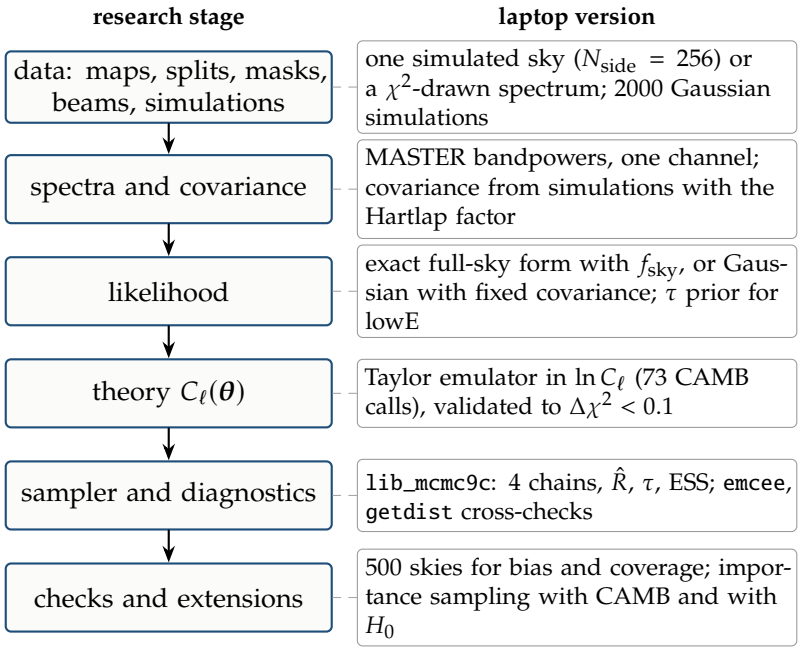


Figure 9.59: The stages of a CMB parameter analysis (left) and the reduced version run on a laptop (right). The statistics, the sampler and the diagnostics are the same; the data, the theory code and the size of the simulations are reduced.

our choice	research version	cost of the reduction
Taylor emulator, 73 CAMB calls	Boltzmann code at every step, or a neural emulator trained over the whole prior	valid only within a few σ of the fiducial model
one simulated χ^2 spectrum, $f_{\rm sky}$ mode count	real cut-sky cross-spectra at three frequencies with a calibrated covariance	no mode coupling or foregrounds: $\sigma(n_s)$ about 7 per cent too small
temperature only, Gaussian prior on τ	TT, TE, EE, lensing; a non-Gaussian lowE likelihood	weaker constraints; a symmetric τ posterior
no nuisance parameters	about fifteen foreground and calibration parameters marginalised	widths and correlations slightly optimistic
toy sky: $N_{\rm side} = 256$, $\ell \le 601$, 2000 Gaussian simulations	$N_{\rm side} = 2048$, $\ell \le 2500$, 300 end-to-end plus 1000 sky simulations	fewer modes; covariance noise of order $\sqrt{2/n}$ on each element
chains of $4 \times (2\text{–}4) \times 10^4$ steps in seconds	chains of 10^5 or more steps, days on a cluster with a Boltzmann code	none at the level of the diagnostics shown
importance check on 400 samples	reweighting the full chain in parallel	a check of the bulk, not of the far tails

Recap

- For a linear model with Gaussian errors the posterior is exactly Gaussian, with mean the weighted least-squares fit and covariance $(\mathbf{A}^{\mathsf{T}}\mathbf{W}\mathbf{A})^{-1}$, the Fisher matrix; the chain reproduces it, as does `emcee`.
- A column of a chain is a sample of that parameter's marginal: marginalising a nuisance parameter means ignoring its column. The marginal is wider than the slice at the best fit by $1/\sqrt{1-\rho^2}$.

- Published error bars depend on the noise model: Padilla et al.'s straight-line widths are those of one unknown common σ (a Student-t posterior), not of the stated errors.
- Slope and intercept correlate as $-\bar{x}_w/\sqrt{\overline{x^2}_w}$; centring x (or choosing a pivot) removes the correlation and speeds the chain by orders of magnitude.
- A nonlinear model gives a skewed posterior that the Laplace (peak Gaussian) approximation misses and the chain does not. An unknown noise level, integrated out, turns $e^{-\chi^2/2}$ into $(\chi^2)^{-N/2}$; sampling it and ignoring its column gives the same marginal. Fixing it at a wrong value gives wrong widths without warning.
- Hyperparameters with prior $e^{-\alpha}$ give $(\chi^2_k + 2)^{-(n_k/2+1)}$ per data set and let a posterior follow the consistent data; the Bayes factor is a ratio of integrals, not of likelihoods at a point. Multimodal posteriors defeat chains that never cross between modes.
- On the full sky the CMB likelihood is exact: $-2\ln\mathcal{L} = \sum_\ell (2\ell+1)[\widehat{C}_\ell/C_\ell + \ln C_\ell]$, with $C_\ell \to T_\ell^2 C_\ell + N_\ell$ for beam and noise and $2\ell+1 \to (2\ell+1)f_{\rm sky}$ for a partial sky; it is skewed at low ℓ.
- Gaussian approximations differ from it in the cubic term in $\widehat{C}_\ell/C_\ell - 1$. $\mathcal{L}_Q$ biases a spectrum up by $0.9/\ell$ in bins of ten and n_s by more than one standard deviation; a fixed fiducial covariance ($\mathcal{L}_f$, Planck's choice) is unbiased in the ideal case. Hamimeche and Lewis's Fig. 1 is reproduced to a few per cent.
- An emulator in $\ln C_\ell$ is natural ($\ln A_s$ and τ enter linearly, n_s almost) and must be validated with $\Delta\chi^2 = \sum \frac{\nu}{2}[\delta C/(C+N)]^2 \ll 1$: first order for $(\ln A_s, n_s)$, second order for six parameters.
- The amplitude and tilt are a straight-line fit in $x_\ell = \langle \ln k/k_* \rangle_\ell$; the sign of their correlation is the side of the pivot the data lie on.
- A Planck-like TT+lowE simulation reproduces Planck 2018's widths to ten per cent with six parameters free; $\sigma(\ln A_s) \approx 2\sigma_\tau$ because TT measures only $A_s e^{-2\tau}$; fixing the other parameters makes the amplitude nine times and the tilt twice too precise.
- Importance sampling reweights a chain by $\mathcal{L}_{\rm new}/\mathcal{L}_{\rm old}$; the effective sample size $(\sum w)^2/\sum w^2$ shows when it works (compatible data, accurate approximations) and when it fails (tension).

Reading the literature

- **Markov chains and MCMC, textbook level.** Wasserman, *All of Statistics*, Chapters 23 (Markov chains) and 24 (Metropolis–Hastings, Gibbs) [Wasserman pp. 381–420]; Kruschke, *Doing Bayesian Data Analysis*, Chapters 7 (MCMC with the island-hopping politician, diagnostics) and 17 (regression, centring, posterior predictive lines) [Kruschke Chs. 7, 8, 17]; Lambert, *A Student's Guide to Bayesian Statistics*, Chapters 12–14 on the typical set, $\hat{R}$ and effective sample size [Lambert Chs. 12–14].
- **The original algorithms.** Metropolis et al. on the canonical ensemble [36]; Hastings on asymmetric proposals [170]; Peskun on the optimality of the Metropolis acceptance [172]; Roberts, Gelman and Gilks on the $2.38/\sqrt{d}$ scaling [173]; Gelman and Rubin, and Brooks and Gelman, on multiple chains [38, 176]; Sokal on autocorrelation times [175]; Geman and Geman on Gibbs sampling [180]; Duane et al. and Neal on Hamiltonian Monte Carlo [181, 182]; Goodman and Weare, and Foreman-Mackey et al. (`emcee`), on affine-invariant ensembles [174, 183].

- **MCMC in cosmology.** Lewis and Bridle on CosmoMC, proposal covariances, fast and slow parameters and importance sampling [37], and their chapter in Hobson et al., *Bayesian Methods in Cosmology* [HobsonBMC Ch. 3]; Dunkley et al. on convergence and efficiency of cosmological chains [171]; Padilla et al. for a worked tour from coins to straight lines to cosmology [40]; Torrado and Lewis on Cobaya [189]; Audren et al. on MontePython [190]; Lewis on `getdist` [188]; Spurio Mancini et al. on neural emulators [187].

- **CMB likelihoods.** Hobson et al., Chapter 10, on the CMB as a hierarchical model, from maps to spectra to parameters [HobsonBMC pp. 237–240]; Hamimeche and Lewis on the exact full-sky likelihood and its approximations [16]; Bond, Jaffe and Knox on offset log-normal forms [191]; Verde et al. on the WMAP likelihood [184]; Chu et al. on the Blackwell–Rao estimator [185]; Hivon et al. on MASTER [168]; Hartlap et al. on inverting simulated covariances [6].

- **Planck.** Planck 2018 V on the likelihoods [75] and VI on the cosmological parameters [25]; Planck 2018 I for the instrument [186]; CAMB [8].

Chapter summary

A posterior that can be evaluated point by point but not integrated is known only in principle. A Markov chain turns it into a sample. The chain is a random walker that remembers only where it is; if the target density is stationary for its hopping rule, the walker forgets where it started and its time averages converge to averages over the target. Metropolis–Hastings supplies such a rule for any target: propose a step, accept it with the ratio of densities, and repeat the current point on refusal. Without the accept–reject step the walker would diffuse forever and sample nothing; the step is the bias that makes it linger where the probability is high, and because only ratios enter, the evidence never has to be computed. A finite chain is trusted only after its diagnostics: burn-in, several chains that agree ($\hat{R}$), and the autocorrelation time that converts N steps into N/τ independent ones. Then every marginal is a histogram of a column, and every constraint is read off. On a straight line, a decaying source and a full-sky spectrum the chain reproduces exact answers; on a masked sky and a six-parameter Planck-like spectrum it is the instrument that produces them, reproducing Planck's published widths once the same parameters are left free. In the pipeline

$$\text{physical model} \to \text{field} \to \text{summary } S \to P(S \mid \theta) \to \boxed{\text{infer physics}},$$

Markov chain Monte Carlo is the last arrow: it turns the sampling distribution of Chapters 7 and 8 into posterior constraints. Chapter 10 gives the fast Gaussian forecast that the chains check, Chapter 12 applies the same chains to topological summaries, and Chapter 13 runs them on Planck data.

Markov chains

concept	operational meaning	used later in
Markov chain	next state drawn from $p(x, \cdot)$ depending only on the current state; rows of P are pmfs	Chapters 12 and 13
stationary distribution	$\pi P = \pi$; probability flowing in balances flowing out	Chapters 10 and 13
detailed balance	$\pi_i p_{ij} = \pi_j p_{ji}$; sufficient for stationarity; no net currents	Chapter 13
irreducible, aperiodic, positive recurrent	every state reachable, no cycles, finite return time: the chain converges	Chapter 13
convergence (forgetting)	$\|p_t - \pi\|_{\mathrm{TV}} \sim \|\lambda_2\|^t$; any start is forgotten	Chapter 13
ergodic theorem	time average along one chain $\to$ average over π	Chapters 10, 12 and 13
autocorrelation time	$\tau = 1 + 2\sum \rho_k$; variance of a chain mean $\sigma^2\tau/N$	Chapters 12 and 13
free random walk	spreads as $\sqrt{t}$, no stationary law: the walker MH must bias	Chapter 11

The biased random walker

concept	operational meaning	used later in
Metropolis–Hastings	propose $y \sim q(y \mid x)$, accept with $\min(1, f(y)q(x \mid y)/f(x)q(y \mid x))$, else repeat x	Chapters 11 to 13
only ratios	evidence and partition function cancel; work with $\ln f$	Chapter 13
Hastings factor	Jacobian of an asymmetric proposal; omitting it samples the wrong target	Chapter 13
step size	round step $2.38/\sqrt{D}$ target widths, acceptance ≈ 0.23; shape the proposal by the covariance	Chapters 10 and 13
burn-in	cut where $\ln p$ settles near $\ln p_{\max} - D/2$, not at the peak	Chapter 13
ESS, MCSE	N/τ effective draws; Monte Carlo error $\mathrm{SD}\sqrt{\tau/N}$; thinning never helps	Chapters 12 and 13
$\hat{R}$	several overdispersed chains; $\sqrt{\hat{V}/W} \to 1$; blind to modes no chain found	Chapter 13
start vs prior	the start is forgotten, the prior is part of the target	Chapters 10 and 13
canonical ensemble	Metropolis samples $e^{-H/kT}$; MCMC time is not physical time	Chapter 12
Gibbs, HMC, ensembles	always-accepted conditional moves; Hamiltonian flights; affine-invariant walkers (`emcee`)	Chapters 11 and 13

MCMC at work

concept	operational meaning	used later in
marginal from a chain	histogram one column, ignore the others	Chapters 10, 12 and 13
linear-Gaussian posterior	exact Gaussian with covariance $(\mathbf{A}^\mathsf{T}\mathbf{W}\mathbf{A})^{-1}$ = inverse Fisher matrix	Chapter 10
centring, pivot	move the origin to the data's weighted centre to decorrelate parameters	Chapters 10 and 13
Laplace approximation	Gaussian from the curvature at the peak; misses skewness of nonlinear models	Chapter 10
nuisance noise level	integrate out $((\chi^2)^{-N/2})$ or sample and ignore; never fix at a guess	Chapters 11 and 13
hyperparameters	one marginalised weight per data set, $(\chi^2_k + 2)^{-(n_k/2+1)}$; Bayes factor by integration	Chapter 11
exact full-sky C_ℓ likelihood	$\sum(2\ell+1)[\widehat{C}_\ell/C_\ell + \ln C_\ell]$; $C_\ell \to T^2 C_\ell + N$; $\nu \to (2\ell+1) f_{\rm sky}$	Chapters 10 and 13
Gaussian C_ℓ likelihoods	fixed fiducial covariance unbiased; model-dependent variance ($\mathcal{L}_Q$) biases n_s by $\sim 1\sigma$	Chapter 13
emulator	Taylor series of $\ln C_\ell$; validate by $\Delta\chi^2 = \sum \frac{\nu}{2}(\delta C/(C+N))^2 \ll 1$ on the posterior	Chapters 10 and 13
triangle plot	1-D and 2-D marginals; 68/95 per cent levels by sorted cells; $\Delta\chi^2 = 2.30, 6.18$ in 2-D	Chapters 10, 12 and 13
$A_s e^{-2\tau}$ degeneracy	TT measures $\ln A_s - 2\tau$; $\sigma(\ln A_s) \approx 2\sigma_\tau$	Chapters 10 and 13
importance sampling	reweight by $\mathcal{L}_{\rm new}/\mathcal{L}_{\rm old}$; $N_{\rm eff} = (\sum w)^2/\sum w^2$; fails for data in tension	Chapters 11 and 13

Chapter 10

The Fisher Matrix: Forecasting What Data Can Teach

Where this chapter is going

Everything so far has started from data. In Chapter 3 a data set gave a likelihood, the likelihood had a peak, and the width of the peak was the error bar. In Chapter 9 a simulated temperature spectrum gave a posterior over six cosmological parameters, and a Markov chain mapped it out. When designing a satellite we have to reason in the other direction. A mission may cost a billion euros and take fifteen years to build, yet its beam size, noise level and sky coverage must be chosen before it collects a single photon. How precisely would each proposed design measure the parameters we care about?

For a specified signal model, noise model and fiducial set of parameters, we can predict the distribution of the data each design would collect. Averaging the curvature of $-\ln \mathcal{L}$ over those possible data sets gives the **Fisher matrix** at the fiducial parameters. When the likelihood is well approximated locally by a Gaussian and the matrix is invertible, its inverse gives the forecast parameter covariance. This lets us compare designs on a common footing.

Making the idea precise takes a few steps. The Fisher matrix has two faces, the covariance of the score and the average curvature, and they are equal. No unbiased method can beat its inverse (the Cramér–Rao bound), and maximum likelihood reaches it when the data are informative. For Gaussian data there is a closed formula, which splits the information into a part carried by the mean and a part carried by the covariance. Reading the matrix needs care: the error with the other parameters fixed is not the error with them free, and the difference is a Schur complement. Priors and independent experiments are added as matrices, a change of parameters is a Jacobian sandwich, and a nuisance parameter is removed by inverting. Then we forecast the six ΛCDM parameters for a Planck-like experiment, check the forecast against what Planck actually achieved, reproduce Knox's 1995 forecast and the classic degeneracies of the CMB, and test the whole method against the Markov chain of Chapter 9. Part 10b applies the same machinery to gravitational waves. The last part turns it on a tiny twist of the CMB polarization: a forecast of how small a rotation angle a Planck-like map can detect, and a thousand simulated masked skies that check the forecast (Sections 10.15 and 10.17), the rehearsal for the real-data measurement of Chapter 13 (Figure 10.1).

In the pipeline of the book (physical model → field or data set → what matters → summary → sampling distribution → inference), the Fisher matrix lives on the last two arrows, but it uses them *before* the data exist. It takes the sampling distribution $p(\boldsymbol{d} \mid \boldsymbol{\theta})$ of Chapters 7 and 8 and asks how wide the inference of Chapter 9 will be, on average, for an experiment that has not been built.

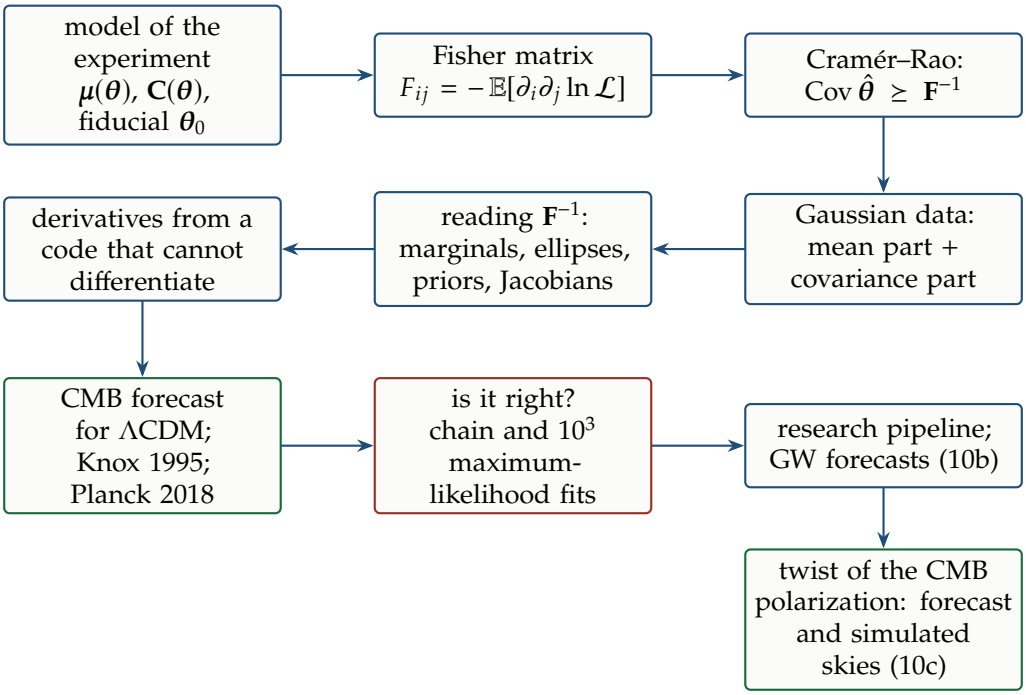


Figure 10.1: The route of the chapter. A model of the experiment and a guess for the truth give the Fisher matrix; the Cramér–Rao bound says what its inverse means; Gaussian data give it in closed form; a short grammar tells us how to read and combine it; finite differences supply the derivatives. The green boxes are the worked applications on the CMB, the last of them a forecast and simulation of a rotation of the polarization; the red box is the test of whether the forecast can be trusted.

10.1 How sharp will the likelihood be? The Fisher matrix

Two teams propose an experiment to measure the angular distribution of a scattering process. The theory says that the cosine $x = \cos\theta$ of the scattering angle has the density

$$f(x;\alpha,\beta) = \frac{1 + \alpha x + \beta x^2}{\mathcal{N}(\alpha,\beta)}, \qquad -0.95 \leq x \leq 0.95, \tag{10.1}$$

where α and β are the two numbers the theory wants measured and $\mathcal{N}$ makes the density integrate to one over the range the detector sees [Cowan §6.8, (6.26)–(6.27)]. One team promises 2000 events, the other 500 events with a wider acceptance. Which proposal will measure β better, and by how much? Nobody has taken data. Yet the answer can be worked out exactly, today, from (10.1) alone. That is what a forecast is: a statement about data nobody has yet taken, computed from the model of the experiment.

The question in words. In Section 3.12 we learned to read an error bar off one data set: the log-likelihood $\ln\mathcal{L}(\theta)$ has a peak at the maximum-likelihood estimate $\hat{\theta}$, and the sharper the peak, the smaller the error. Here there is no data set, so there is no peak to look at. The question becomes: *if the parameters have some definite true value and we run this experiment many times, how sharp will the peak of the log-likelihood be, on average?*

The hidden assumption. The question already contains an assumption: "the parameters have some definite true value". A forecast is always made *at* a chosen point θ_0, the **fiducial model** [the parameter values we assume nature has chosen, used to simulate or average over the data]. The width of the likelihood generally depends on where we sit. For the angular

distribution, a measurement of β near $\beta = 0$ and near $\beta = 5$ are different problems. So every forecast here reads: "*if* the truth is $\boldsymbol{\theta}_0$, the experiment will measure $\boldsymbol{\theta}$ to within ...". Changing $\boldsymbol{\theta}_0$ changes the answer; for a linear model it does not, and that is the only case where the fiducial does not matter [HobsonBMC p. 107].

10.1.1 From the curvature of one likelihood to the average curvature

Start with one data set $\boldsymbol{d}$ and m parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$. Near its maximum $\hat{\boldsymbol{\theta}}$, the log-likelihood $\ell(\boldsymbol{\theta}) = \ln \mathcal{L}(\boldsymbol{\theta})$ is a smooth hill. Expand it in a Taylor series about the top [Sivia (3.17)–(3.19)], [HobsonBMC (5.13)–(5.14)]:

$$\ell(\boldsymbol{\theta}) \overset{(1)}{=} \ell(\hat{\boldsymbol{\theta}}) + \sum_i \frac{\partial \ell}{\partial \theta_i}\Big|_{\hat{\boldsymbol{\theta}}} (\theta_i - \hat{\theta}_i) + \frac{1}{2} \sum_{ij} \frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j}\Big|_{\hat{\boldsymbol{\theta}}} (\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j) + \ldots$$

$$\overset{(2)}{=} \ell(\hat{\boldsymbol{\theta}}) - \frac{1}{2} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^{\mathsf{T}} \boldsymbol{\mathcal{H}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \ldots, \qquad \mathcal{H}_{ij} \equiv -\frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j}\Big|_{\hat{\boldsymbol{\theta}}}. \tag{10.2}$$

(1) Taylor's theorem in m variables, to second order.

(2) At the top of the hill every first derivative vanishes, so the linear terms drop out. We collect the second derivatives, with a minus sign, into the matrix $\boldsymbol{\mathcal{H}}$; at a maximum it is positive definite [$\boldsymbol{v}^{\mathsf{T}} \boldsymbol{\mathcal{H}} \boldsymbol{v} > 0$ for every nonzero vector $\boldsymbol{v}$, so the hill falls off in every direction].

Why the curvature is a covariance. Exponentiate (10.2), dropping the higher terms: $\mathcal{L}(\boldsymbol{\theta}) \propto \exp[-\frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^{\mathsf{T}} \boldsymbol{\mathcal{H}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})]$. Take one parameter first. Then $\boldsymbol{\mathcal{H}}$ is a single positive number $\mathcal{H}$ and the likelihood is $\exp[-\mathcal{H}(\theta - \hat{\theta})^2/2]$. A Gaussian of variance σ^2 has the shape $\exp[-(\theta - \hat{\theta})^2/2\sigma^2]$; the two match when $1/\sigma^2 = \mathcal{H}$, so the variance is $1/\mathcal{H}$. With many parameters, the Gaussian of covariance matrix $\boldsymbol{\Sigma}$ has the shape $\exp[-\frac{1}{2}\boldsymbol{d}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{d}]$ with $\boldsymbol{d} = \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}$, and matching the two exponents term by term gives $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\mathcal{H}}$, that is $\boldsymbol{\Sigma} = \boldsymbol{\mathcal{H}}^{-1}$ [Sivia (3.31)–(3.32)]. This is the many-parameter version of "the curvature is the error bar". The matrix $\boldsymbol{\mathcal{H}}$ is the **observed information**: the curvature of *this* likelihood, at *this* peak.

From one data set to the average. Each new data set has its own peak and its own curvature. To forecast, we average over the data sets the experiment could produce, drawn from $p(\boldsymbol{d} \mid \boldsymbol{\theta}_0)$, and evaluate the second derivatives at the fiducial point rather than at a peak we do not have. Two facts make this a small change for large data sets. First, with n independent observations ℓ is a sum of n terms, so its curvature is a sum of n independent curvatures. By the law of large numbers that sum is n times its average plus a fluctuation whose relative size falls as $1/\sqrt{n}$. Second, the peak $\hat{\boldsymbol{\theta}}$ lies within a distance of order $1/\sqrt{n}$ of $\boldsymbol{\theta}_0$ (consistency, Section 3.13.1), so moving the point of evaluation from $\hat{\boldsymbol{\theta}}$ to $\boldsymbol{\theta}_0$ changes the curvature by a relative amount of the same order. The average curvature at the truth is the Fisher matrix [Wasserman p. 133, (9.18)], [192]:

$$\boxed{F_{ij}(\boldsymbol{\theta}_0) = -\,\mathbb{E}_{\boldsymbol{\theta}_0}\left[\frac{\partial^2 \ln p(\boldsymbol{d} \mid \boldsymbol{\theta})}{\partial \theta_i \, \partial \theta_j}\right]_{\boldsymbol{\theta} = \boldsymbol{\theta}_0}} \tag{10.3}$$

where $\mathbb{E}_{\boldsymbol{\theta}_0}$ is an average over data drawn from the model at $\boldsymbol{\theta}_0$.[1] Read in words: *how fast, on average, does the log-likelihood fall away from the truth?* The two facts above together give $\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\theta}}) = \mathbf{F}(\boldsymbol{\theta}_0)\,[1 + O(1/\sqrt{n})]$. So the covariance $\boldsymbol{\mathcal{H}}^{-1}$ that one data set would show is, up to corrections of relative size $1/\sqrt{n}$, the matrix $\mathbf{F}^{-1}$ that we can compute before any data exist. The exact statement, that no unbiased estimate can scatter less than $\mathbf{F}^{-1}$, is the Cramér–Rao bound of Section 10.2.

[1] Hobson et al. [HobsonBMC (5.14)–(5.15)] define $H_{ij} = \partial^2 \ln \mathcal{L}/\partial\theta_i \partial\theta_j$ and then write $F_{ij} = \langle H_{ij} \rangle$; with that sign $\mathbf{F}$ would be negative definite. The Hessian there must be that of $-\ln \mathcal{L}$, as in Tegmark et al. [192], eq. (4), who write $\mathcal{L} = -\ln L$.

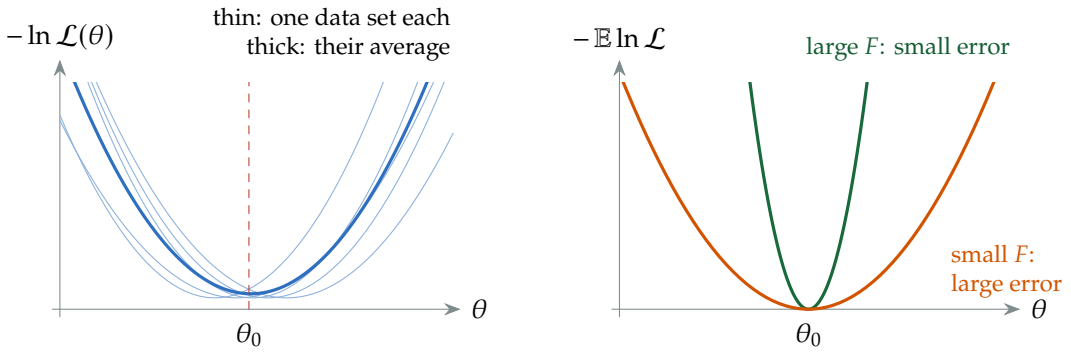


Figure 10.2: Left: every data set the experiment could produce gives its own log-likelihood (thin curves), with its own minimum of $-\ln \mathcal{L}$ near the true value θ_0 and its own curvature; the Fisher information is the curvature of their average (thick curve) at θ_0. Right: the forecast compares experiments by this average curvature, a sharp valley meaning a small error bar.

Working definition 10.1 (Fisher matrix)

For data with density $p(\boldsymbol{d} \mid \boldsymbol{\theta})$ and a fiducial point $\boldsymbol{\theta}_0$, the **Fisher matrix** is the $m \times m$ matrix $F_{ij} = -\mathbb{E}_{\boldsymbol{\theta}_0}[\partial_i \partial_j \ln p(\boldsymbol{d} \mid \boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{\theta}_0}[S_i S_j]$, where $S_i = \partial_i \ln p(\boldsymbol{d} \mid \boldsymbol{\theta})$ is the **score** [the slope of the log-likelihood in the direction of θ_i, a random quantity because the data are random], all evaluated at $\boldsymbol{\theta}_0$. *Operationally:* write the log-likelihood of the planned experiment, differentiate twice, average over data simulated at $\boldsymbol{\theta}_0$ (or do the average analytically), and change the sign. Its inverse $\mathbf{F}^{-1}$ is the forecast covariance of the parameters.

10.1.2 Two faces of one matrix: average curvature and covariance of the score

For one parameter, Section 3.12.3 proved that the score has mean zero and that its variance equals minus the average curvature. The same two facts hold for many parameters, component by component, and they are worth redoing because the second one is what makes the Fisher matrix a covariance matrix, and hence positive semidefinite.

The score has mean zero. Write $p = p(\boldsymbol{d} \mid \boldsymbol{\theta})$ and $\partial_i = \partial/\partial\theta_i$. For each i,

$$\mathbb{E}[S_i] \overset{(1)}{=} \int \frac{\partial_i p}{p}\, p \,\mathrm{d}\boldsymbol{d} \overset{(2)}{=} \partial_i \int p \,\mathrm{d}\boldsymbol{d} \overset{(3)}{=} \partial_i 1 = 0. \tag{10.4}$$

(1) The expectation is an integral over data weighted by their density at the same $\boldsymbol{\theta}$, and $\partial_i \ln p = \partial_i p / p$.
(2) The p cancels; take the derivative outside the integral. This exchange is the regularity condition: it needs the range of the data not to depend on $\boldsymbol{\theta}$ (Section 3.13.3).
(3) A density integrates to one for every value of $\boldsymbol{\theta}$.

The covariance of the score is minus the average Hessian. Differentiate the identity $\int (\partial_i \ln p)\, p \,\mathrm{d}\boldsymbol{d} = 0$ with respect to θ_j:

$$0 \overset{(1)}{=} \int \Big[(\partial_j \partial_i \ln p)\, p + (\partial_i \ln p)\, \partial_j p\Big] \mathrm{d}\boldsymbol{d} \overset{(2)}{=} \int \Big[\partial_i \partial_j \ln p + (\partial_i \ln p)(\partial_j \ln p)\Big] p \,\mathrm{d}\boldsymbol{d},$$

$$\mathbb{E}[S_i S_j] \overset{(3)}{=} -\mathbb{E}[\partial_i \partial_j \ln p] = F_{ij}. \tag{10.5}$$

(1) Product rule under the integral, again exchanging derivative and integral.

(2) $\partial_j p = (\partial_j \ln p)\, p$.
(3) Split the integral into two expectations. Since $\mathbb{E}\, S_i = 0$ by (10.4), the left side $\mathbb{E}[S_i S_j]$ is the covariance $\mathrm{Cov}(S_i, S_j)$.

So $\mathbf{F}$ is the covariance matrix of the score vector. A covariance matrix is positive semidefinite: $\boldsymbol{v}^\mathsf{T}\mathbf{F}\boldsymbol{v} = \mathrm{Var}(\boldsymbol{v}^\mathsf{T}\boldsymbol{S}) \geq 0$ for every $\boldsymbol{v}$. If it is strictly positive, $\mathbf{F}$ can be inverted. If some combination $\boldsymbol{v}^\mathsf{T}\boldsymbol{S}$ has zero variance, the data never respond to a change of the parameters along $\boldsymbol{v}$: that combination is not measured at all, $\mathbf{F}$ is singular, and no amount of data will fix it (an **exact degeneracy**).

Independent pieces of data add. If the data consist of independent parts, $p(\boldsymbol{d} \mid \boldsymbol{\theta}) = \prod_k p_k(\boldsymbol{d}_k \mid \boldsymbol{\theta})$, the log-likelihood is a sum, every second derivative is a sum, and so is its average: $\mathbf{F} = \sum_k \mathbf{F}_k$. For n independent, identically distributed observations $\mathbf{F} = n\,\mathbf{F}_1$ [Cowan (6.20)], which is where the familiar $1/\sqrt{n}$ comes from: $\mathbf{F}^{-1} \propto 1/n$, so the error bars shrink as $1/\sqrt{n}$.

Example: a Gaussian sample, mean and width both unknown. Take n independent draws $x_k \sim \mathcal{N}(\mu, \sigma^2)$ and $\boldsymbol{\theta} = (\mu, \sigma)$. Section 3.13.2 averaged the three second derivatives of $\ln f = -\ln\sigma - (x-\mu)^2/2\sigma^2$ and found [Wasserman p. 134, Ex. 9.29]

$$\mathbf{F} = n \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{pmatrix}.$$

Seen as a covariance of scores, every entry of this matrix has a plain reason, and it is worth computing them once that way. Write $u_k = x_k - \mu$ for the deviations; each has mean 0 and variance σ^2, and for a Gaussian $\mathbb{E}[u^3] = 0$ and $\mathbb{E}[u^4] = 3\sigma^4$ (the Gaussian fourth moment behind (3.34)). The two scores are $S_\mu = \sum_k u_k/\sigma^2$ and $S_\sigma = \sum_k (u_k^2 - \sigma^2)/\sigma^3$. Then

$$\begin{aligned}
\mathrm{Var}\, S_\mu &\overset{(1)}{=} \frac{n\,\mathrm{Var}\, u}{\sigma^4} = \frac{n}{\sigma^2}, \\
\mathrm{Var}\, S_\sigma &\overset{(2)}{=} \frac{n\,\mathrm{Var}(u^2)}{\sigma^6} \overset{(3)}{=} \frac{n\,(3\sigma^4 - \sigma^4)}{\sigma^6} = \frac{2n}{\sigma^2}, \\
\mathrm{Cov}(S_\mu, S_\sigma) &\overset{(4)}{=} \frac{n\;\mathbb{E}[u\,(u^2 - \sigma^2)]}{\sigma^5} = \frac{n\,(\mathbb{E}\,u^3 - \sigma^2\,\mathbb{E}\,u)}{\sigma^5} \overset{(5)}{=} 0.
\end{aligned}$$

(1) The n terms are independent, so their variances add; dividing a variable by σ^2 divides its variance by σ^4.
(2) The same, with the constant $-\sigma^2$ in each term not changing the variance.
(3) $\mathrm{Var}(u^2) = \mathbb{E}[u^4] - (\mathbb{E}\,u^2)^2 = 3\sigma^4 - \sigma^4 = 2\sigma^4$.
(4) Terms with different k are independent and have mean zero, so only the n terms with the same k survive; both scores have mean zero, so the covariance is the mean of the product.
(5) A Gaussian is symmetric about μ, so its odd moments $\mathbb{E}\,u$ and $\mathbb{E}\,u^3$ vanish.

These are the entries of $\mathbf{F}$ above. The width score has twice the variance of the mean score because u^2 fluctuates by $\sqrt{2}\sigma^2$ about its mean while u fluctuates by σ; the two scores are uncorrelated because one is odd in the deviations and the other even. Since $\mathbf{F}$ is diagonal, its inverse is found entry by entry, $\mathbf{F}^{-1} = \mathrm{diag}(\sigma^2/n,\ \sigma^2/2n)$. The forecast errors are therefore $\sigma/\sqrt{n}$ on the mean and $\sigma/\sqrt{2n}$ on the width, and the two estimates are uncorrelated.

The score cloud in Python. The covariance face of $\mathbf{F}$ can be watched directly. The script `02_score_cloud.py` draws 20000 samples of $n = 20$ points from $\mathcal{N}(0,1)$, computes the score vector (S_μ, S_σ) of each at the true parameters, and plots the cloud (Figure 10.3, left). The mean

score is $(+0.044, +0.001)$, consistent with zero: the simulation noise is $\sqrt{20/20000} \approx \pm 0.03$ on the first entry and $\sqrt{40/20000} \approx \pm 0.05$ on the second. The covariance of the cloud has diagonal $(19.95, 39.96)$ and off-diagonal -0.02; the average of minus the Hessian has diagonal $(20.00, 40.00)$ and off-diagonal $+0.09$; and the formula above gives $(20, 40)$ and 0. Three routes, one matrix.

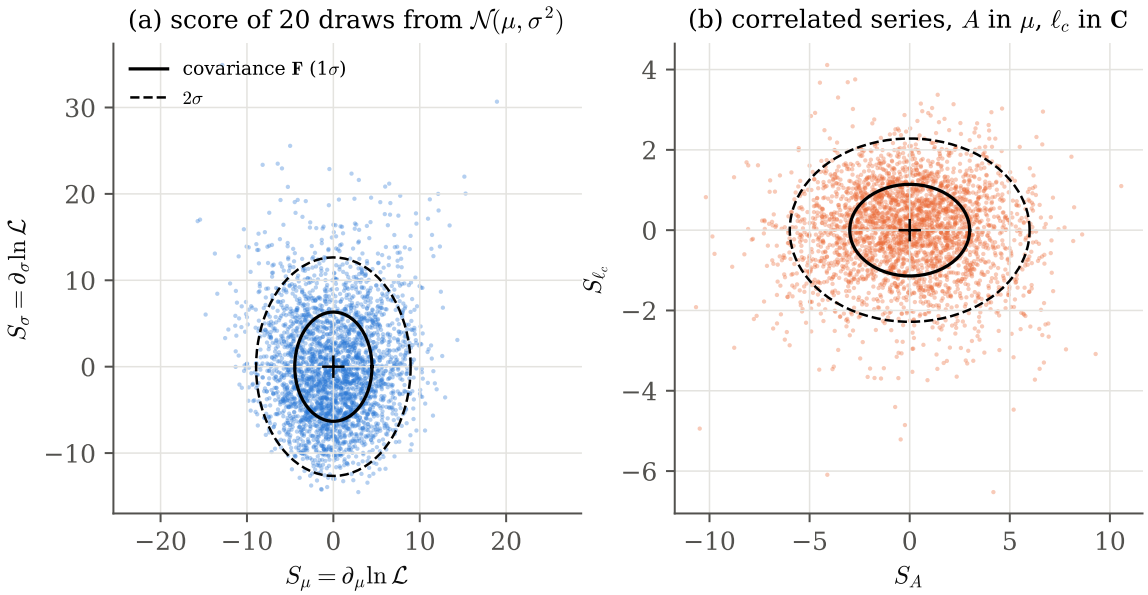


Figure 10.3: The score is a random vector whose covariance is the Fisher matrix. Each point is the slope of the log-likelihood of one simulated data set, evaluated at the true parameters; the solid and dashed ellipses are drawn from $\mathbf{F}$ alone. Left: 20 Gaussian draws, parameters (μ, σ). Right: a correlated time series in which the amplitude A enters the mean and the correlation length ℓ_c enters the covariance (Section 10.3).

10.1.3 Example: forecasting Cowan's angular distribution

Back to the two proposals. Cowan used this very density to ask a question every experimenter meets [Cowan §6.8]: *how precisely can two shape parameters be pulled out of one finite sample, and does the error bar read off a single fit agree with the scatter we would see if the experiment were repeated many times?* The second half matters because in practice there is only one data set, and the error bar quoted from it has to stand in for the scatter of experiments that will never be done.

His setup. The density (10.1), seen through the window $-0.95 \le x \le 0.95$, with true values $\alpha = \beta = 0.5$. He generated one sample of 2000 events, found the maximum of the likelihood, and read the errors and the correlation off the numerical Hessian of that one fit [Cowan (6.28)]. Then he generated 500 further samples of 2000 events, fitted each, and took the standard deviations and the correlation of the 500 estimates [Cowan (6.29)]. His numbers are in the first two rows of Table 10.1. Our question: can we predict them without generating a single event?

The log-density of one event. With $g(x) = 1 + \alpha x + \beta x^2$ and the normalisation over the window $[a, b] = [-0.95, 0.95]$,

$$\mathcal{N}(\alpha, \beta) = \int_a^b g \, \mathrm{d}x = (b - a) + \alpha \frac{b^2 - a^2}{2} + \beta \frac{b^3 - a^3}{3}, \qquad \ln f = \ln g(x) - \ln \mathcal{N}(\alpha, \beta).$$

Its score. Differentiating, $\partial_\alpha \ln f = x/g - \mathcal{N}_\alpha/\mathcal{N}$ and $\partial_\beta \ln f = x^2/g - \mathcal{N}_\beta/\mathcal{N}$, with $\mathcal{N}_\alpha = (b^2 - a^2)/2$ and $\mathcal{N}_\beta = (b^3 - a^3)/3$ the derivatives of the normalisation. *The Fisher matrix per event* is the

covariance of this score, an integral over the model:

$$I_{ij} = \mathbb{E}[\partial_i \ln f \, \partial_j \ln f] = \int_a^b \partial_i \ln f \; \partial_j \ln f \; f(x)\, \mathrm{d}x, \tag{10.6}$$

which we do by numerical quadrature:

Listing 10.1: The Fisher matrix of one event: the covariance of the score, integrated over the model (`code/ch10/01_cowan_two_param.py`)

```
41 def fisher_per_event(a, b):
42     """I_ij = E[d_i ln f d_j ln f] = int (d_i f)(d_j f)/f dx, with d_i f by hand."""
43     n = norm(a, b)
44     dn = np.array([(XMAX**2 - XMIN**2) / 2, (XMAX**3 - XMIN**3) / 3])   # d norm / d(a, b)
45
46     def dlnf(x):
47         g = 1 + a * x + b * x**2
48         return np.array([x / g - dn[0] / n, x**2 / g - dn[1] / n])
49
50     I = np.zeros((2, 2))
51     for i in range(2):
52         for j in range(2):
53             I[i, j] = integrate.quad(lambda x: dlnf(x)[i] * dlnf(x)[j] * pdf(x, a, b),
       XMIN, XMAX)[0]
54     return I
```

Lines 46–48 are the score written above; line 53 is (10.6), one quadrature per matrix element. No random number appears anywhere. It gives $I_{\alpha\alpha} = 0.2250$, $I_{\alpha\beta} = -0.0451$ and $I_{\beta\beta} = 0.0497$. *Many events.* The events are independent, so $\mathbf{F} = 2000\,\boldsymbol{I}$, with entries $F_{\alpha\alpha} = 450$, $F_{\alpha\beta} = -90$ and $F_{\beta\beta} = 99$.

The forecast is the inverse. For a symmetric 2×2 matrix the inverse is the swapped diagonal with the off-diagonal negated, divided by the determinant:

$$\mathbf{F}^{-1} = \frac{1}{\det \mathbf{F}} \begin{pmatrix} F_{\beta\beta} & -F_{\alpha\beta} \\ -F_{\alpha\beta} & F_{\alpha\alpha} \end{pmatrix}, \qquad \det \mathbf{F} = F_{\alpha\alpha}F_{\beta\beta} - F_{\alpha\beta}^2, \tag{10.7}$$

as multiplying it by $\mathbf{F}$ confirms. The diagonal entries are the variances and the off-diagonal entry the covariance, so

$$\sigma_\alpha^2 = \frac{F_{\beta\beta}}{\det \mathbf{F}}, \qquad \sigma_\beta^2 = \frac{F_{\alpha\alpha}}{\det \mathbf{F}}, \qquad \rho = \frac{(\mathbf{F}^{-1})_{\alpha\beta}}{\sigma_\alpha \sigma_\beta} = \frac{-F_{\alpha\beta}/\det \mathbf{F}}{\sqrt{F_{\alpha\alpha}F_{\beta\beta}}/\det \mathbf{F}} = -\frac{F_{\alpha\beta}}{\sqrt{F_{\alpha\alpha}F_{\beta\beta}}}.$$

With the numbers above, $\det \mathbf{F} = 450 \times 99 - (-90)^2 \approx 3.6 \times 10^4$, so $\sigma_\alpha = \sqrt{99/\det \mathbf{F}} = 0.052$, $\sigma_\beta = \sqrt{450/\det \mathbf{F}} = 0.111$, and $\rho = -(-90)/\sqrt{450 \times 99} = +0.43$. A negative entry of the Fisher matrix becomes a positive correlation: the sign flips in the inverse.

Our own repeats. To have a second set of repeats with known code, `01_cowan_two_param.py` also generates 500 experiments of 2000 events, fits each by Newton–Raphson, and records the observed Hessian of the first one (Figure 10.4).

Why the rows differ. The scatter of 500 fits is itself an estimate. The standard deviation of N Gaussian numbers has a relative error of $1/\sqrt{2(N-1)}$, which is 3% for $N = 500$, and a correlation coefficient has an error of about $(1-\rho^2)/\sqrt{N} \approx 0.04$. Cowan's repeats differ from the forecast by 2%, 0% and 0.01, all inside these limits; ours by about 4%, 3% and 0.01, about one sampling error. The single-fit rows differ for another reason. The observed Hessian $\boldsymbol{\mathcal{H}}$ is the curvature of one particular likelihood, and it fluctuates from data set to data set by a relative amount of order $1/\sqrt{n}$ (Section 10.1.1); the Fisher matrix is its average. That is why

	σ_α	σ_β	ρ
Cowan, Hessian of one fit	0.052	0.11	0.46
Cowan, scatter of 500 fits	0.051	0.111	0.42
ours, Fisher forecast (no events)	0.052	0.111	0.43
ours, Hessian of one fit	0.053	0.115	0.38
ours, scatter of 500 fits	0.050	0.108	0.44

Table 10.1: Cowan's two-parameter example [Cowan (6.28)–(6.29)], 2000 events at $\alpha = \beta = 0.5$. A standard deviation estimated from 500 fits carries a sampling error of $1/\sqrt{2 \times 499} \approx \pm 3\%$, and a correlation $\pm(1-\rho^2)/\sqrt{500} \approx \pm 0.04$.

Cowan's single fit gives $\rho = 0.46$ and ours 0.38, on either side of 0.43, while their errors stay within a few per cent of the forecast. For 2000 events the fluctuation is small, so his first row and his second row are two estimates of the same thing, and the Fisher matrix is the number both scatter around.

The positive correlation also has a plain reason. Over the window the two shapes x and x^2 both rise towards $x = +0.95$, and since the density must stay normalised, an upward fluctuation in the number of events at large positive x can be absorbed partly by α and partly by β, so the two estimates tend to move together.

The answer to the proposal question is now one line. Errors scale as $1/\sqrt{n}$, so 500 events in the same window would give $\sigma_\beta = 0.111 \times \sqrt{2000/500} \approx 0.22$. A wider window changes $\boldsymbol{I}$ itself, and the same quadrature, with a and b changed, tells us by how much.

The lesson: the forecast and the repeated experiment agree to within the error of the repeats, so the expensive part of Cowan's study, the 500 simulated experiments, could have been replaced by one integral per matrix element; what a single fit adds is a curvature that wobbles about that integral.

Key idea: The Fisher matrix

$F_{ij} = -\mathbb{E}[\partial_i \partial_j \ln \mathcal{L}] = \mathbb{E}[S_i S_j]$, evaluated at the fiducial parameters, is the average curvature of the log-likelihood and the covariance of the score. It needs a model of the experiment, not data. Its inverse is the forecast covariance of the parameters, and independent data sets add their Fisher matrices.

10.2 Nobody beats the inverse Fisher matrix: the Cramér–Rao bound

The forecast in Section 10.1.3 used the width of the likelihood. An analysis of the same experiment might use moments of the data or train a network to estimate the parameters from a histogram. Does $\mathbf{F}^{-1}$ describe the precision of a maximum likelihood fit alone, or does it limit the precision of other estimators as well? To answer that question we must specify what we require of the estimator, in particular whether it is unbiased.

For one parameter the answer was the Cramér–Rao inequality of Section 3.13.3: every unbiased estimator has $\mathrm{Var}\,\hat{\theta} \geq 1/I_n(\theta)$ [CasellaBerger Thm. 7.3.9], [Cowan (6.16)]. With several parameters, a new question appears that the one-parameter bound cannot answer.

The question in words. Can any unbiased method of turning the data into estimates $\boldsymbol{T}(\boldsymbol{d}) = (T_1, \ldots, T_m)$ have, *in some direction of parameter space*, a smaller scatter than $\mathbf{F}^{-1}$ predicts? And when we quote the error on one parameter, is the limit $1/\sqrt{F_{ii}}$, which looks only at the i-th

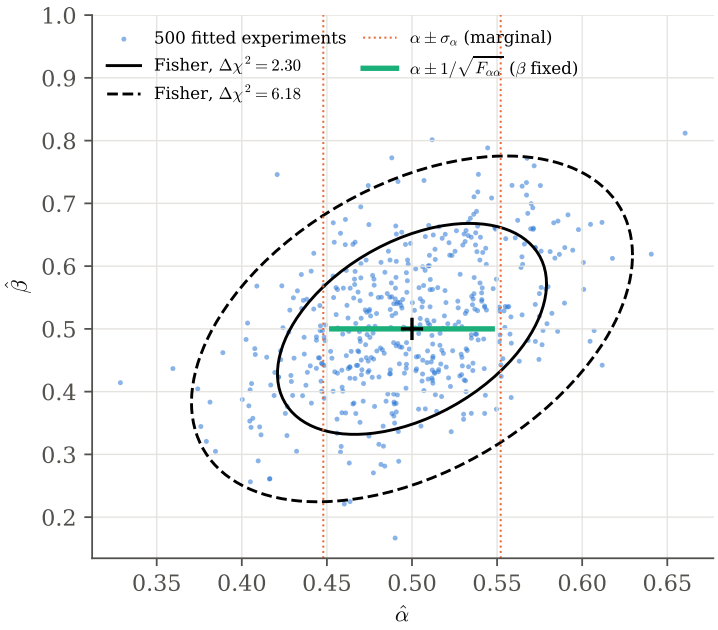


Figure 10.4: Cowan's experiment repeated: each point is the maximum-likelihood fit of one simulated experiment of 2000 events. The ellipses come from the Fisher matrix alone, computed before any event was generated: the solid one should contain 68.3% of the points and the dashed one 95.4% (Section 10.4.2). The dotted vertical lines are $\alpha \pm \sigma_\alpha$; the green bar is the shorter error on α that we would quote if β were known (Section 10.4.1).

diagonal element, or $\sqrt{(\mathbf{F}^{-1})_{ii}}$, which involves the whole matrix?

10.2.1 The bound, from one covariance matrix

Let $\boldsymbol{T}$ be unbiased, $\mathbb{E}_{\boldsymbol{\theta}}\,\boldsymbol{T} = \boldsymbol{\theta}$ for every $\boldsymbol{\theta}$, with covariance matrix $\boldsymbol{V} = \operatorname{Cov}\boldsymbol{T}$, and let $\boldsymbol{S}$ be the score of the whole data set. The proof needs one fact about how $\boldsymbol{T}$ and $\boldsymbol{S}$ are related, and one fact about covariance matrices.

How the estimator and the score are correlated. For each pair a, j,

$$\operatorname{Cov}(T_a, S_j) \overset{(1)}{=} \mathbb{E}[T_a\, S_j] \overset{(2)}{=} \int T_a(\boldsymbol{d})\,\frac{\partial_j p(\boldsymbol{d}\,|\,\boldsymbol{\theta})}{p(\boldsymbol{d}\,|\,\boldsymbol{\theta})}\,p(\boldsymbol{d}\,|\,\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{d} \overset{(3)}{=} \partial_j \int T_a(\boldsymbol{d})\,p(\boldsymbol{d}\,|\,\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{d} \overset{(4)}{=} \partial_j \theta_a = \delta_{aj}. \tag{10.8}$$

(1) $\operatorname{Cov}(T_a, S_j) = \mathbb{E}[T_a S_j] - \mathbb{E}\,T_a\ \mathbb{E}\,S_j$, and $\mathbb{E}\,S_j = 0$ by (10.4).
(2) Write the expectation as an integral; $S_j = \partial_j p / p$.
(3) Cancel p and take the derivative out of the integral. T_a depends on the data only, not on $\boldsymbol{\theta}$; the exchange is the same regularity condition as before.
(4) The integral is $\mathbb{E}\,T_a$, which equals θ_a because the estimator is unbiased.

So an unbiased estimator must be correlated with the score in a fixed way: $\operatorname{Cov}(\boldsymbol{T}, \boldsymbol{S}) = \mathbb{1}$. This is the whole content of unbiasedness for the proof. If $\boldsymbol{T}$ did not respond to the data along the directions in which the data respond to $\boldsymbol{\theta}$, it could not follow $\boldsymbol{\theta}$ on average.

Covariance matrices are positive semidefinite. Stack the $2m$ random numbers $(\boldsymbol{T}, \boldsymbol{S})$ into one vector. Its covariance matrix is

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{V} & \mathbb{1} \\ \mathbb{1} & \mathbf{F} \end{pmatrix}, \qquad (\boldsymbol{u}^{\mathsf{T}}, \boldsymbol{w}^{\mathsf{T}})\,\boldsymbol{M}\begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{w} \end{pmatrix} = \operatorname{Var}(\boldsymbol{u}^{\mathsf{T}}\boldsymbol{T} + \boldsymbol{w}^{\mathsf{T}}\boldsymbol{S}) \geq 0 \quad \text{for all } \boldsymbol{u}, \boldsymbol{w}, \tag{10.9}$$

where the off-diagonal blocks are (10.8) and the lower block is (10.5).

The bound. Pick any direction $\boldsymbol{a}$ in parameter space and set $\boldsymbol{u} = \boldsymbol{a}$, $\boldsymbol{w} = -\mathbf{F}^{-1}\boldsymbol{a}$:

$$0 \overset{(1)}{\leq} \operatorname{Var}\left(\boldsymbol{a}^{\mathsf{T}}\boldsymbol{T} - \boldsymbol{a}^{\mathsf{T}}\mathbf{F}^{-1}\boldsymbol{S}\right) \overset{(2)}{=} \boldsymbol{a}^{\mathsf{T}}\boldsymbol{V}\boldsymbol{a} - 2\,\boldsymbol{a}^{\mathsf{T}}\mathbf{F}^{-1}\boldsymbol{a} + \boldsymbol{a}^{\mathsf{T}}\mathbf{F}^{-1}\mathbf{F}\mathbf{F}^{-1}\boldsymbol{a} \overset{(3)}{=} \boldsymbol{a}^{\mathsf{T}}\left(\boldsymbol{V} - \mathbf{F}^{-1}\right)\boldsymbol{a}. \tag{10.10}$$

(1) (10.9) with this choice of $\boldsymbol{u}, \boldsymbol{w}$.
(2) Expand the quadratic form block by block: $\boldsymbol{u}^{\mathsf{T}}\boldsymbol{V}\boldsymbol{u} + 2\boldsymbol{u}^{\mathsf{T}}\mathbb{1}\,\boldsymbol{w} + \boldsymbol{w}^{\mathsf{T}}\mathbf{F}\boldsymbol{w}$, using $\mathbf{F}^{-1}$ symmetric.
(3) $\mathbf{F}^{-1}\mathbf{F} = \mathbb{1}$; collect.

Why that choice of $\boldsymbol{w}$? One parameter shows it. Then T and S are two random numbers with $\operatorname{Cov}(T, S) = 1$ by (10.8) and $\operatorname{Var} S = F$. Ask for the multiple cS of the score that comes closest to T, in the sense of the smallest variance of the difference:

$$\operatorname{Var}(T - cS) \overset{(1)}{=} \operatorname{Var} T - 2c\,\operatorname{Cov}(T, S) + c^2 \operatorname{Var} S \overset{(2)}{=} \operatorname{Var} T - 2c + c^2 F,$$

$$\frac{\mathrm{d}}{\mathrm{d}c}\operatorname{Var}(T - cS) = -2 + 2cF = 0 \overset{(3)}{=} c = \frac{1}{F}, \qquad \operatorname{Var}\left(T - \frac{S}{F}\right) \overset{(4)}{=} \operatorname{Var} T - \frac{1}{F}.$$

(1) The variance of a difference.
(2) Insert $\operatorname{Cov}(T, S) = 1$ and $\operatorname{Var} S = F$.
(3) The best coefficient is covariance over variance, as in a least-squares regression of T on S.
(4) Put $c = 1/F$ back into the first line: $\operatorname{Var} T - 2/F + 1/F$.

The residual $T - S/F$ is uncorrelated with the score, $\operatorname{Cov}(T - S/F, S) = 1 - F/F = 0$: it is the part of T the score cannot predict. Its variance cannot be negative, so $\operatorname{Var} T \geq 1/F$. The variance of any unbiased T splits into a part $1/F$ that is forced by its correlation with the score and a leftover that a good estimator makes small. With many parameters the same regression of $\boldsymbol{a}^{\mathsf{T}}\boldsymbol{T}$ on the score vector has coefficient vector $\mathbf{F}^{-1}\boldsymbol{a}$ (covariance $\boldsymbol{a}^{\mathsf{T}}\mathbb{1}$ times the inverse of the variance matrix $\mathbf{F}$), which is exactly the $\boldsymbol{w} = -\mathbf{F}^{-1}\boldsymbol{a}$ chosen above, and (10.10) is the matrix version of the last line. Since $\boldsymbol{a}$ was arbitrary,

$$\boxed{\operatorname{Cov}\boldsymbol{T} - \mathbf{F}^{-1} \geq 0 : \qquad \operatorname{Var}(\boldsymbol{a}^{\mathsf{T}}\boldsymbol{T}) \geq \boldsymbol{a}^{\mathsf{T}}\mathbf{F}^{-1}\boldsymbol{a} \text{ for every } \boldsymbol{a}, \qquad \operatorname{Var} T_i \geq (\mathbf{F}^{-1})_{ii},} \tag{10.11}$$

where ≥ 0 means positive semidefinite. The last form is the choice $\boldsymbol{a} = \boldsymbol{e}_i$, the i-th unit vector [Cowan (6.19)], [192]. Positive semidefinite ordering has a picture: the ellipse of $\operatorname{Cov}\boldsymbol{T}$ contains the ellipse of $\mathbf{F}^{-1}$, because the half-width of an ellipse $\{\boldsymbol{x} : \boldsymbol{x}^{\mathsf{T}}\boldsymbol{V}^{-1}\boldsymbol{x} \leq 1\}$ measured along a unit direction $\boldsymbol{a}$ is $\sqrt{\boldsymbol{a}^{\mathsf{T}}\boldsymbol{V}\boldsymbol{a}}$ (Figure 10.5).

10.2.2 Marginalised or conditional: the second half of the answer

The bound on T_i is $(\mathbf{F}^{-1})_{ii}$, the ii element of the *inverse*. How does it compare with $1/F_{ii}$, the inverse of the ii element? Two tools are needed. *Picking out an element.* Write $\boldsymbol{e}_i$ for the unit vector with 1 in place i and 0 elsewhere; then $\boldsymbol{M}\boldsymbol{e}_i$ is the i-th column of a matrix $\boldsymbol{M}$ and $\boldsymbol{e}_i^{\mathsf{T}}\boldsymbol{M}\boldsymbol{e}_i = M_{ii}$. *A square root of* $\mathbf{F}$. A symmetric matrix can be written as $\mathbf{F} = \boldsymbol{V}\,\operatorname{diag}(\lambda_1, \ldots, \lambda_m)\boldsymbol{V}^{\mathsf{T}}$, with the orthonormal eigenvectors as the columns of $\boldsymbol{V}$ ($\boldsymbol{V}^{\mathsf{T}}\boldsymbol{V} = \mathbb{1}$) and the eigenvalues λ_a on the diagonal; for a positive definite $\mathbf{F}$ all $\lambda_a > 0$. Then $\mathbf{F}^{1/2} \equiv \boldsymbol{V}\,\operatorname{diag}(\sqrt{\lambda_a})\boldsymbol{V}^{\mathsf{T}}$ is symmetric and squares to $\mathbf{F}$, because the inner $\boldsymbol{V}^{\mathsf{T}}\boldsymbol{V}$ is $\mathbb{1}$; likewise $\mathbf{F}^{-1/2} \equiv \boldsymbol{V}\,\operatorname{diag}(1/\sqrt{\lambda_a})\boldsymbol{V}^{\mathsf{T}}$, and $\mathbf{F}^{1/2}\mathbf{F}^{-1/2} = \mathbb{1}$. Then

$$1 \overset{(1)}{=} \left(\boldsymbol{e}_i^{\mathsf{T}}\boldsymbol{e}_i\right)^2 \overset{(2)}{=} \left[(\mathbf{F}^{1/2}\boldsymbol{e}_i)^{\mathsf{T}}(\mathbf{F}^{-1/2}\boldsymbol{e}_i)\right]^2 \overset{(3)}{\leq} \left(\boldsymbol{e}_i^{\mathsf{T}}\mathbf{F}\boldsymbol{e}_i\right)\left(\boldsymbol{e}_i^{\mathsf{T}}\mathbf{F}^{-1}\boldsymbol{e}_i\right) = F_{ii}\,(\mathbf{F}^{-1})_{ii}, \tag{10.12}$$

(1) A unit vector has length one.

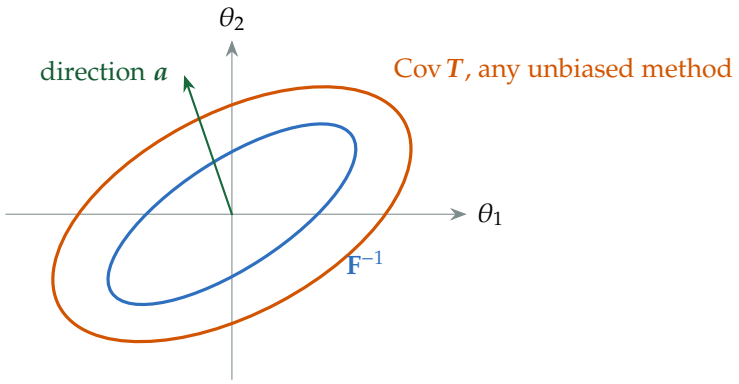


Figure 10.5: The Cramér–Rao bound as a picture. For every direction $\boldsymbol{a}$, the spread of an unbiased estimator along $\boldsymbol{a}$ is at least $\sqrt{\boldsymbol{a}^\mathsf{T}\mathbf{F}^{-1}\boldsymbol{a}}$, so the error ellipse of any method (orange) encloses the ellipse of the inverse Fisher matrix (blue). Maximum likelihood with informative data shrinks its ellipse onto the blue one.

(2) Insert $\mathbf{F}^{1/2}\mathbf{F}^{-1/2} = \mathbb{1}$ between the two factors; both matrices are symmetric.
(3) Cauchy–Schwarz for ordinary vectors, $(\boldsymbol{x}^\mathsf{T}\boldsymbol{y})^2 \le (\boldsymbol{x}^\mathsf{T}\boldsymbol{x})(\boldsymbol{y}^\mathsf{T}\boldsymbol{y})$.

so $(\mathbf{F}^{-1})_{ii} \ge 1/F_{ii}$.

When are the two equal? Cauchy–Schwarz is an equality only when the two vectors are parallel, $\mathbf{F}^{1/2}\boldsymbol{e}_i = c\,\mathbf{F}^{-1/2}\boldsymbol{e}_i$ for some number c. Multiplying both sides by $\mathbf{F}^{1/2}$ turns this into $\mathbf{F}\boldsymbol{e}_i = c\,\boldsymbol{e}_i$: the i-th column of $\mathbf{F}$ has a single nonzero entry, on the diagonal. So $F_{ji} = 0$ for every $j \ne i$, and parameter i shares no information with any other. Then $\mathbf{F}^{-1}\boldsymbol{e}_i = \boldsymbol{e}_i/c$ as well, and $(\mathbf{F}^{-1})_{ii} = 1/F_{ii}$. In every other case the inverse of the matrix gives the larger number. A small example gives the size: for $F_{11} = 2$, $F_{22} = 1$, $F_{12} = 1$ the determinant is 1 and $(\mathbf{F}^{-1})_{11} = F_{22}/\det\mathbf{F} = 1$ by (10.7), so $F_{11}(\mathbf{F}^{-1})_{11} = 2$: the error on θ_1 is $\sqrt{2}$ times larger when θ_2 must be estimated too. This is the content of Tegmark, Taylor and Heavens' statement of the bound [192]: $1/\sqrt{F_{ii}}$ is the smallest error on θ_i *if all the other parameters are known*; if they are estimated from the same data, the smallest error rises to $\sqrt{(\mathbf{F}^{-1})_{ii}}$. We come back to this distinction, the single most common error in quoting forecasts, in Section 10.4.1. In the angular-distribution experiment of Section 10.1.3 (Cowan's two-parameter example) the two numbers for α are $1/\sqrt{F_{\alpha\alpha}} = 0.047$ and $\sqrt{(\mathbf{F}^{-1})_{\alpha\alpha}} = 0.052$ (Figure 10.4, green bar against dotted lines).

10.2.3 When the bound is reached, and what it means

A bound is only useful if someone reaches it. Vallisneri lists three readings of $\mathbf{F}^{-1}$ that are often confused [193]: (i) the Cramér–Rao bound on any unbiased estimator, a statement about the scatter over imagined repetitions of the experiment; (ii) the covariance of the maximum-likelihood estimator, which reaches the bound when the model is linear in $\boldsymbol{\theta}$ over the region where the errors live; (iii) the covariance of the Bayesian posterior from *one* experiment, with a flat prior, in the same limit. The first is always true; the second and third need a condition, and the examples show what the condition means.

When is the bound an equality, and what if the estimator is biased? Equality in (10.10) needs $\mathrm{Var}(\boldsymbol{a}^\mathsf{T}\boldsymbol{T} - \boldsymbol{a}^\mathsf{T}\mathbf{F}^{-1}\boldsymbol{S}) = 0$ for every $\boldsymbol{a}$. A random variable with zero variance is a constant, and since both $\boldsymbol{T}$ and $\boldsymbol{S}$ have the right means ($\boldsymbol{\theta}$ and $\mathbf{0}$) the constant is $\boldsymbol{\theta}$ itself:

$$\boldsymbol{T}(\boldsymbol{d}) = \boldsymbol{\theta} + \mathbf{F}^{-1}\boldsymbol{S}(\boldsymbol{d}) \quad \text{for all data.}$$

So an estimator reaches the bound exactly when the score is a linear function of the estimator. This is the case for the linear-Gaussian model below, where the score is $\boldsymbol{A}^{\mathsf{T}}\mathbf{C}^{-1}(\boldsymbol{d}-\boldsymbol{A}\boldsymbol{\theta})$, and it holds only approximately in the other models of this chapter. The ratio $e = \big[(\mathbf{F}^{-1})_{ii}\big] \,/\, \operatorname{Var} T_i \le 1$ is the **efficiency** of the estimator [Cowan §6.6]: how much of the available information it uses.

A biased estimator. Suppose instead $\mathbb{E}\,\boldsymbol{T} = \boldsymbol{\theta} + \boldsymbol{b}(\boldsymbol{\theta})$, with a bias $\boldsymbol{b}$ that may depend on the parameters, and write $B_{aj} = \partial_j b_a$ for its derivatives. Only the last step of (10.8) changes: the integral is now $\mathbb{E}\,T_a = \theta_a + b_a$, so

$$\operatorname{Cov}(T_a, S_j) = \partial_j(\theta_a + b_a) = \delta_{aj} + B_{aj}, \qquad \text{that is} \qquad \operatorname{Cov}(\boldsymbol{T}, \boldsymbol{S}) = \mathbb{1} + \boldsymbol{B},$$

and the off-diagonal blocks of (10.9) become $\mathbb{1} + \boldsymbol{B}$ and $(\mathbb{1} + \boldsymbol{B})^{\mathsf{T}}$. Repeat the regression with the new covariance: $\boldsymbol{u} = \boldsymbol{a}$ and $\boldsymbol{w} = -\mathbf{F}^{-1}(\mathbb{1} + \boldsymbol{B})^{\mathsf{T}}\boldsymbol{a}$. Then

$$\begin{aligned} 0 &\overset{(1)}{\le} \boldsymbol{a}^{\mathsf{T}}\boldsymbol{V}\boldsymbol{a} + 2\,\boldsymbol{a}^{\mathsf{T}}(\mathbb{1} + \boldsymbol{B})\boldsymbol{w} + \boldsymbol{w}^{\mathsf{T}}\mathbf{F}\boldsymbol{w} \\ &\overset{(2)}{=} \boldsymbol{a}^{\mathsf{T}}\boldsymbol{V}\boldsymbol{a} - 2\,\boldsymbol{a}^{\mathsf{T}}(\mathbb{1} + \boldsymbol{B})\mathbf{F}^{-1}(\mathbb{1} + \boldsymbol{B})^{\mathsf{T}}\boldsymbol{a} + \boldsymbol{a}^{\mathsf{T}}(\mathbb{1} + \boldsymbol{B})\mathbf{F}^{-1}\mathbf{F}\mathbf{F}^{-1}(\mathbb{1} + \boldsymbol{B})^{\mathsf{T}}\boldsymbol{a} \\ &\overset{(3)}{=} \boldsymbol{a}^{\mathbb{T}}\big[\boldsymbol{V} - (\mathbb{1} + \boldsymbol{B})\mathbf{F}^{-1}(\mathbb{1} + \boldsymbol{B})^{\mathsf{T}}\big]\boldsymbol{a}. \end{aligned}$$

(1) The quadratic form of the joint covariance, expanded block by block, is a variance and so not negative.
(2) Insert $\boldsymbol{w}$; its transpose is $-\boldsymbol{a}^{\mathsf{T}}(\mathbb{1} + \boldsymbol{B})\mathbf{F}^{-1}$, since $\mathbf{F}^{-1}$ is symmetric.
(3) $\mathbf{F}^{-1}\mathbf{F} = \mathbb{1}$, so the last term is $\boldsymbol{a}^{\mathsf{T}}(\mathbb{1} + \boldsymbol{B})\mathbf{F}^{-1}(\mathbb{1} + \boldsymbol{B})^{\mathsf{T}}\boldsymbol{a}$, and the middle term is -2 times the same number; together they leave minus it once.

Since $\boldsymbol{a}$ is arbitrary,

$$\operatorname{Cov}\boldsymbol{T} \ge (\mathbb{1} + \boldsymbol{B})\,\mathbf{F}^{-1}(\mathbb{1} + \boldsymbol{B})^{\mathsf{T}}, \tag{10.13}$$

which is the information inequality of [Cowan (6.16)] for several parameters. With $\boldsymbol{B} = \mathbf{0}$ it is (10.11) again.

A bias can buy a smaller variance. Take n draws from $\mathcal{N}(\theta, \sigma^2)$ with known σ, so $F = n/\sigma^2$, and shrink the sample mean towards zero: $T = c\,\bar{x}$ with $0 < c < 1$. Its mean is $c\theta$, so the bias is $b = -(1-c)\theta$ and $B = \mathrm{d}b/\mathrm{d}\theta = -(1-c)$, giving $1 + B = c$. The bound (10.13) reads $\operatorname{Var} T \ge c^2/F = c^2\sigma^2/n$, and indeed $\operatorname{Var} T = c^2 \operatorname{Var}\bar{x} = c^2\sigma^2/n$: the shrunk estimator sits on its bound, which is below the unbiased bound $1/F = \sigma^2/n$ by the factor c^2. A bias whose derivative is negative pulls estimates together, and the variance drops below $\mathbf{F}^{-1}$; the price is that T is off by $(1-c)\theta$ on average. This is the loophole of the last example below.

Example: a linear model reaches the bound exactly. Let the data be $\boldsymbol{d} = \boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{n}$ with a known $N \times m$ matrix $\boldsymbol{A}$ (each column a template, for example the columns 1 and x of a straight-line fit) and Gaussian noise $\boldsymbol{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$ with known $\mathbf{C}$. Then $\ln \mathcal{L} = -\tfrac{1}{2}(\boldsymbol{d} - \boldsymbol{A}\boldsymbol{\theta})^{\mathsf{T}}\mathbf{C}^{-1}(\boldsymbol{d} - \boldsymbol{A}\boldsymbol{\theta}) + \text{const}$ and

$$\begin{aligned} -\partial_i\partial_j \ln \mathcal{L} &\overset{(1)}{=} (\boldsymbol{A}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{A})_{ij} \equiv F_{ij}, \\ \hat{\boldsymbol{\theta}} &\overset{(2)}{=} (\boldsymbol{A}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{d} = \mathbf{F}^{-1}\boldsymbol{A}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{d}, \\ \operatorname{Cov}\hat{\boldsymbol{\theta}} &\overset{(3)}{=} \mathbf{F}^{-1}\boldsymbol{A}^{\mathsf{T}}\mathbf{C}^{-1}\,\operatorname{Cov}(\boldsymbol{d})\,\mathbf{C}^{-1}\boldsymbol{A}\,\mathbf{F}^{-1} \overset{(4)}{=} \mathbf{F}^{-1}\big(\boldsymbol{A}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{A}\big)\mathbf{F}^{-1} = \mathbf{F}^{-1}. \end{aligned}$$

(1) Differentiate the quadratic form twice: the second derivative is constant, the same for every data set, so no average is needed.
(2) Set the gradient $\boldsymbol{A}^{\mathsf{T}}\mathbf{C}^{-1}(\boldsymbol{d} - \boldsymbol{A}\boldsymbol{\theta})$ to zero: the normal equations of Chapter 3.
(3) $\hat{\boldsymbol{\theta}}$ is a linear map $\boldsymbol{B}\boldsymbol{d}$ of the data, so $\operatorname{Cov}\hat{\boldsymbol{\theta}} = \boldsymbol{B}\operatorname{Cov}(\boldsymbol{d})\boldsymbol{B}^{\mathsf{T}}$; the term $\boldsymbol{A}\boldsymbol{\theta}$ in $\boldsymbol{d}$ is not random.

(4) $\mathrm{Cov}(\boldsymbol{d}) = \mathbf{C}$, and $\mathbf{C}^{-1}\mathbf{C}\mathbf{C}^{-1} = \mathbf{C}^{-1}$.

The bound is reached exactly, at any number of data points. The log-likelihood is an exact parabola in $\boldsymbol{\theta}$, its curvature is the same for every data set, and all three readings of $\mathbf{F}^{-1}$ coincide. Every other situation is this one only approximately: a nonlinear model looks linear when the errors are small compared with the scale on which the model bends, which is what "informative data" or "high signal-to-noise" means.

Example: how fast does Cowan's experiment reach the bound? The angular distribution is not linear in (α, β), because the parameters also sit in the denominator of $\ln g$. `01_cowan_two_param.py` repeats the experiment 2000 times at each of several event numbers n and compares the variance of the fitted $\hat{\alpha}$ and $\hat{\beta}$ with the bound $(\mathbf{F}^{-1})_{ii} = (\boldsymbol{I}^{-1})_{ii}/n$ (Figure 10.6). At $n = 2000$ the ratios are 1.03 and 1.00, one within the simulation error of about ± 0.03. At $n = 50$ they are 1.57 and 1.73, and in 1.5% of the experiments the fit does not converge at all: the likelihood keeps increasing as $\beta \to \infty$, because a handful of events all at large $|x|$ look like a pure x^2 shape. With so little information the log-likelihood is far from a parabola over the range where the estimates scatter, and a Fisher forecast for 50 events would be too optimistic by a quarter in σ_α.

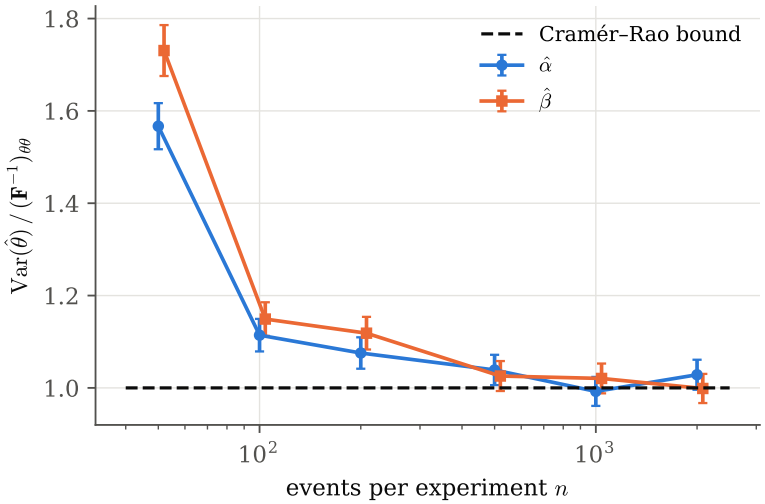


Figure 10.6: The variance of the maximum-likelihood estimates in Cowan's experiment divided by the Cramér–Rao bound, against the number of events. Below a few hundred events the fits scatter more than the bound; from about 500 events on, maximum likelihood reaches it. Error bars are the Monte Carlo uncertainty of a variance estimated from 2000 experiments.

Example: the bound is for unbiased estimators. Two loopholes are worth naming. First, a biased estimator can beat $\mathbf{F}^{-1}$. The posterior mean with an informative prior is one: it is pulled towards the prior, and its scatter is smaller, which is fair because the prior is extra information. In the Fisher language the prior simply adds its own matrix (Section 10.4.4). Second, the bound needs the regularity condition: for Uniform$(0, \theta)$ the maximum of the sample has a variance falling like $1/n^2$ (Section 3.13.3), because a sharp edge in the data is not covered by the derivation. In the other direction, the bound does not promise that a good estimator exists at small n. Take the variance of a Gaussian with unknown mean, now with (μ, σ^2) as parameters instead of (μ, σ). *The bound.* The Jacobian rule (10.27) converts the Fisher entry for σ found in Section 10.1.2 into one for σ^2: with $\sigma = \sqrt{\sigma^2}$, $\mathrm{d}\sigma/\mathrm{d}\sigma^2 = 1/(2\sigma)$, so

$$F_{\sigma^2\sigma^2} = \left(\frac{\mathrm{d}\sigma}{\mathrm{d}\sigma^2}\right)^2 F_{\sigma\sigma} = \frac{1}{4\sigma^2}\,\frac{2n}{\sigma^2} = \frac{n}{2\sigma^4},$$

while $F_{\mu\mu} = n/\sigma^2$ and the off-diagonal entry stay as they were, zero. The matrix is diagonal, so the bound on any unbiased estimator of σ^2 is the inverse of this one entry, $2\sigma^4/n$. *The best unbiased estimator.* The sample variance S^2 has $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$, and a χ^2_k variable has variance $2k$, so $\mathrm{Var}\, S^2 = \sigma^4 \cdot 2(n-1)/(n-1)^2 = 2\sigma^4/(n-1)$, as in (3.34). It exceeds the bound by the factor $n/(n-1)$, and no unbiased estimator does better [CasellaBerger Ex. 7.3.14]: the bound is approached only as n grows.

Key idea: Cramér–Rao for many parameters

For any unbiased estimator, $\mathrm{Cov}\, \boldsymbol{T} - \mathbf{F}^{-1}$ is positive semidefinite: no direction of parameter space can be measured better than $\mathbf{F}^{-1}$ allows. The error on one parameter is at least $\sqrt{(\mathbf{F}^{-1})_{ii}} \geq 1/\sqrt{F_{ii}}$, the second value applying only when all other parameters are known. Maximum likelihood reaches the bound exactly for linear-Gaussian models and approximately whenever the data are informative enough that the log-likelihood is a parabola over the scatter of the estimates.

10.3 Gaussian data: the Fisher matrix in closed form

Most data in physics reach the analysis as a Gaussian vector, or are compressed into one. The pixels of a CMB map, its harmonic coefficients $a_{\ell m}$, a binned power spectrum, the Fourier amplitudes of a detector's strain, the points of a fitted line: each is either Gaussian by nature (a Gaussian random field, thermal detector noise) or Gaussian to a good approximation by the central limit theorem (a bin average of many modes, Chapter 2). The parameters reach such data in two different ways. Either they move the *mean*: a peak position, a template amplitude, the shape of a gravitational waveform. Or they change the *covariance*: the CMB temperature has zero mean at every point, and every cosmological parameter acts only through the variance C_ℓ of the $a_{\ell m}$. A forecast formula for Gaussian data must handle both, and it turns out to be one line.

The question. If the data vector $\boldsymbol{d}$ (length N) is Gaussian with a mean $\boldsymbol{\mu}(\boldsymbol{\theta})$ and a covariance $\mathbf{C}(\boldsymbol{\theta})$ that we can compute for any parameters, what is the Fisher matrix, written with nothing but $\boldsymbol{\mu}$, $\mathbf{C}$ and their derivatives?

The inputs. One physical input: the data are Gaussian, with mean and covariance known as functions of the parameters (all the physics of the experiment, signal and noise, sits in these two functions). Two mathematical inputs: $\boldsymbol{\mu}$ and $\mathbf{C}$ are differentiable in $\boldsymbol{\theta}$, and $\mathbf{C}$ is positive definite, hence invertible. Everything else is derived. With $\boldsymbol{r} = \boldsymbol{d} - \boldsymbol{\mu}$ the residual,

$$-2\ln \mathcal{L}(\boldsymbol{\theta}) = \ln\det \mathbf{C} + \boldsymbol{r}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{r} + N\ln 2\pi \tag{10.14}$$

[HobsonBMC (5.16)], [192], eq. (7). It is the likelihood in the sense of Chapter 3: the data are the model mean plus noise, $\boldsymbol{d} = \boldsymbol{\mu}(\boldsymbol{\theta}) + \boldsymbol{n}$, and for a given $\boldsymbol{\theta}$ the likelihood asks how plausible the particular noise realisation $\boldsymbol{r} = \boldsymbol{d} - \boldsymbol{\mu}(\boldsymbol{\theta})$ is that would turn the model into the data we have. The second term is the familiar χ^2 of Chapter 3; the first is what is left of the normalisation $(2\pi)^{-N/2}(\det \mathbf{C})^{-1/2}$, and it matters as soon as $\mathbf{C}$ depends on the parameters. We write $\boldsymbol{\mu}_{,i} = \partial_i \boldsymbol{\mu}$, $\mathbf{C}_{,i} = \partial_i \mathbf{C}$, and so on.

10.3.1 Three tools

The derivative of an inverse. Differentiate $\mathbf{C}\mathbf{C}^{-1} = \mathbb{1}$: $\mathbf{C}_{,i}\mathbf{C}^{-1} + \mathbf{C}\,\partial_i(\mathbf{C}^{-1}) = 0$, so

$$\partial_i(\mathbf{C}^{-1}) = -\mathbf{C}^{-1}\mathbf{C}_{,i}\mathbf{C}^{-1}. \tag{10.15}$$

It is the matrix version of $(1/c)' = -c'/c^2$; the order matters because matrices do not commute.

The derivative of a log-determinant. For a small change $\delta\mathbf{C}$,

$$\begin{aligned}\ln\det(\mathbf{C}+\delta\mathbf{C}) &\overset{(1)}{=} \ln\det\mathbf{C} + \ln\det(\mathbb{1} + \mathbf{C}^{-1}\delta\mathbf{C}) \overset{(2)}{=} \ln\det\mathbf{C} + \ln\prod_a(1+\epsilon_a)\\ &\overset{(3)}{=} \ln\det\mathbf{C} + \mathrm{Tr}(\mathbf{C}^{-1}\delta\mathbf{C}) + O(\delta^2),\end{aligned} \tag{10.16}$$

(1) $\mathbf{C} + \delta\mathbf{C} = \mathbf{C}(\mathbb{1} + \mathbf{C}^{-1}\delta\mathbf{C})$, and the determinant of a product is the product of determinants.
(2) ϵ_a are the eigenvalues of the small matrix $\mathbf{C}^{-1}\delta\mathbf{C}$; a determinant is the product of the eigenvalues, and $\mathbb{1} + \boldsymbol{\epsilon}$ has eigenvalues $1 + \epsilon_a$.
(3) $\ln(1+\epsilon_a) = \epsilon_a + O(\epsilon^2)$, and the sum of the eigenvalues is the trace.

Hence $\partial_i \ln\det\mathbf{C} = \mathrm{Tr}(\mathbf{C}^{-1}\mathbf{C}_{,i})$, the matrix version of $(\ln c)' = c'/c$.

The average of a quadratic form. For any fixed matrix $\boldsymbol{B}$ and a residual with $\mathbb{E}\,\boldsymbol{r} = \mathbf{0}$, $\mathbb{E}\,\boldsymbol{r}\boldsymbol{r}^{\mathsf{T}} = \mathbf{C}$:

$$\mathbb{E}[\boldsymbol{r}^{\mathsf{T}}\boldsymbol{B}\boldsymbol{r}] = \mathbb{E}\,\mathrm{Tr}(\boldsymbol{B}\boldsymbol{r}\boldsymbol{r}^{\mathsf{T}}) = \mathrm{Tr}(\boldsymbol{B}\,\mathbb{E}[\boldsymbol{r}\boldsymbol{r}^{\mathsf{T}}]) = \mathrm{Tr}(\boldsymbol{B}\mathbf{C}), \tag{10.17}$$

because a number is its own trace, the trace is cyclic, and both trace and expectation are linear. Terms linear in $\boldsymbol{r}$ average to zero, and terms with three factors of $\boldsymbol{r}$ never appear below.

10.3.2 The derivation

The score. Differentiate (10.14) once. The residual depends on $\boldsymbol{\theta}$ through the mean, $\partial_i\boldsymbol{r} = -\boldsymbol{\mu}_{,i}$, so

$$\partial_i(-2\ln\mathcal{L}) \overset{(1)}{=} \mathrm{Tr}(\mathbf{C}^{-1}\mathbf{C}_{,i}) - 2\boldsymbol{\mu}_{,i}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{r} - \boldsymbol{r}^{\mathsf{T}}\mathbf{C}^{-1}\mathbf{C}_{,i}\mathbf{C}^{-1}\boldsymbol{r}. \tag{10.18}$$

(1) The first term is (10.16). In the quadratic form, differentiating either of the two $\boldsymbol{r}$ gives $-\boldsymbol{\mu}_{,i}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{r}$ (the two are equal because $\mathbf{C}^{-1}$ is symmetric), and differentiating $\mathbf{C}^{-1}$ uses (10.15).

Its average is $\mathrm{Tr}(\mathbf{C}^{-1}\mathbf{C}_{,i}) - 0 - \mathrm{Tr}(\mathbf{C}^{-1}\mathbf{C}_{,i}\mathbf{C}^{-1}\mathbf{C}) = 0$ by (10.17): the score has mean zero, as it must.

The second derivative. Differentiate each of the three terms of (10.18) with respect to θ_j, and average at once, dropping every term linear in $\boldsymbol{r}$:

$$\begin{aligned}\mathbb{E}\;\partial_j\,\mathrm{Tr}(\mathbf{C}^{-1}\mathbf{C}_{,i}) &\overset{(1)}{=} -\mathrm{Tr}(\mathbf{C}^{-1}\mathbf{C}_{,j}\mathbf{C}^{-1}\mathbf{C}_{,i}) + \mathrm{Tr}(\mathbf{C}^{-1}\mathbf{C}_{,ij}),\\ \mathbb{E}\;\partial_j\big(-2\boldsymbol{\mu}_{,i}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{r}\big) &\overset{(2)}{=} 2\,\boldsymbol{\mu}_{,i}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{\mu}_{,j},\\ \mathbb{E}\;\partial_j\big(-\boldsymbol{r}^{\mathsf{T}}\mathbf{C}^{-1}\mathbf{C}_{,i}\mathbf{C}^{-1}\boldsymbol{r}\big) &\overset{(3)}{=} \mathbb{E}\;\boldsymbol{r}^{\mathsf{T}}\big[\mathbf{C}^{-1}\mathbf{C}_{,j}\mathbf{C}^{-1}\mathbf{C}_{,i}\mathbf{C}^{-1} - \mathbf{C}^{-1}\mathbf{C}_{,ij}\mathbf{C}^{-1} + \mathbf{C}^{-1}\mathbf{C}_{,i}\mathbf{C}^{-1}\mathbf{C}_{,j}\mathbf{C}^{-1}\big]\boldsymbol{r}\\ &\overset{(4)}{=} \mathrm{Tr}(\mathbf{C}^{-1}\mathbf{C}_{,j}\mathbf{C}^{-1}\mathbf{C}_{,i}) - \mathrm{Tr}(\mathbf{C}^{-1}\mathbf{C}_{,ij}) + \mathrm{Tr}(\mathbf{C}^{-1}\mathbf{C}_{,i}\mathbf{C}^{-1}\mathbf{C}_{,j}).\end{aligned} \tag{10.19}$$

(1) Product rule on $\mathbf{C}^{-1}\mathbf{C}_{,i}$, with (10.15) for the derivative of $\mathbf{C}^{-1}$. Nothing is random here.
(2) Of the three pieces of the product rule, $-2\boldsymbol{\mu}_{,ij}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{r}$ and $+2\boldsymbol{\mu}_{,i}^{\mathsf{T}}\mathbf{C}^{-1}\mathbf{C}_{,j}\mathbf{C}^{-1}\boldsymbol{r}$ are linear in $\boldsymbol{r}$ and average to zero; the third comes from $\partial_j\boldsymbol{r} = -\boldsymbol{\mu}_{,j}$.

(3) The terms in which ∂_j hits an $\boldsymbol{r}$ are linear in $\boldsymbol{r}$ and drop out; what is left is the derivative of the middle matrix $\mathbf{C}^{-1}\mathbf{C}_{,i}\mathbf{C}^{-1}$, using (10.15) on each $\mathbf{C}^{-1}$.
(4) (10.17), then $\mathbf{C}^{-1}\mathbf{C} = \mathbb{1}$ at the end of each product.

Add the three lines. The first trace of the first line cancels the first trace of the third line (they are equal by the cyclic property), the two $\mathrm{Tr}(\mathbf{C}^{-1}\mathbf{C}_{,ij})$ cancel, and what survives is $2\boldsymbol{\mu}_{,i}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{\mu}_{,j} + \mathrm{Tr}(\mathbf{C}^{-1}\mathbf{C}_{,i}\mathbf{C}^{-1}\mathbf{C}_{,j})$. Dividing by two,

$$\boxed{F_{ij} = \boldsymbol{\mu}_{,i}^{\mathsf{T}}\,\mathbf{C}^{-1}\boldsymbol{\mu}_{,j} + \frac{1}{2}\,\mathrm{Tr}\big[\mathbf{C}^{-1}\mathbf{C}_{,i}\,\mathbf{C}^{-1}\mathbf{C}_{,j}\big]} \tag{10.20}$$

which is Tegmark, Taylor and Heavens' eq. (15), there written $\frac{1}{2}\,\mathrm{Tr}[\boldsymbol{A}_i\boldsymbol{A}_j + \mathbf{C}^{-1}\boldsymbol{M}_{ij}]$ with $\boldsymbol{A}_i = \mathbf{C}^{-1}\mathbf{C}_{,i}$ and $\boldsymbol{M}_{ij} = \boldsymbol{\mu}_{,i}\boldsymbol{\mu}_{,j}^{\mathsf{T}} + \boldsymbol{\mu}_{,j}\boldsymbol{\mu}_{,i}^{\mathsf{T}}$ [192], [HobsonBMC (5.17)].

Three things in (10.20) are worth saying in words. First, *only first derivatives appear*: the second derivatives $\boldsymbol{\mu}_{,ij}$ and $\mathbf{C}_{,ij}$ cancelled. A forecast needs to know how the model responds to each parameter, and nothing more. Second, the information splits into two independent channels (Figure 10.7). The mean part is a dot product of the response vectors $\boldsymbol{\mu}_{,i}$ and $\boldsymbol{\mu}_{,j}$ in the metric $\mathbf{C}^{-1}$: a change of θ_i is detectable if it moves the mean by an amount that is large compared with the noise, in the direction where the noise is small. The covariance part measures how much a change of the parameters reshapes the scatter of the data, in units of the scatter itself. Third, the geometric reading of Tegmark et al.: if two response vectors are nearly parallel in that metric, the two parameters do the same thing to the data, the Fisher matrix is nearly singular, and their errors explode together. That is a degeneracy, now seen at the level of the model.

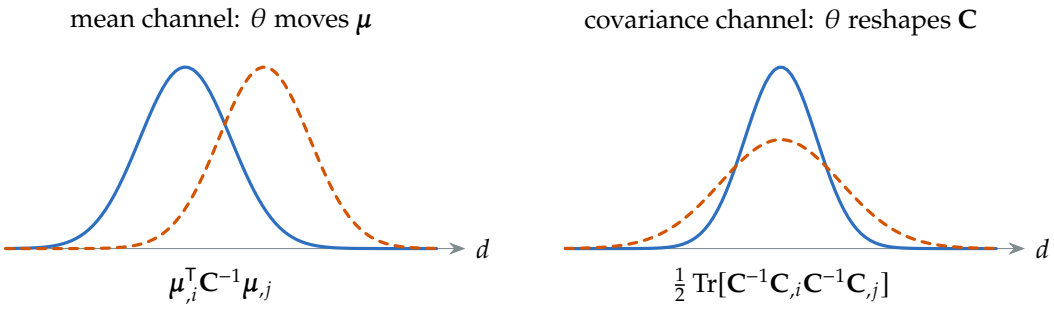


Figure 10.7: The two ways a parameter reaches Gaussian data. Left: changing θ (solid to dashed) shifts the distribution of the data; the information is the squared shift in units of the noise. Right: changing θ leaves the mean at zero and widens the distribution, as cosmological parameters do to the CMB $a_{\ell m}$; the information is the squared fractional change of the variance, with a factor one half.

For simulations it is useful to have the score itself, which is minus one half of (10.18):

$$S_i = \boldsymbol{\mu}_{,i}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{r} + \tfrac{1}{2}\big(\boldsymbol{r}^{\mathsf{T}}\mathbf{C}^{-1}\mathbf{C}_{,i}\mathbf{C}^{-1}\boldsymbol{r} - \mathrm{Tr}\,\mathbf{C}^{-1}\mathbf{C}_{,i}\big). \tag{10.21}$$

10.3.3 Examples

Example: a template of known shape (mean channel only). A detector sees noise of known covariance $\mathbf{C}$ and possibly a signal $A\boldsymbol{s}$ of known shape $\boldsymbol{s}$ and unknown amplitude A. Then $\boldsymbol{\mu}_{,A} = \boldsymbol{s}$, $\mathbf{C}_{,A} = 0$, and $F_{AA} = \boldsymbol{s}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{s}$, so $\sigma_A = (\boldsymbol{s}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{s})^{-1/2}$. The ratio of the true amplitude to its error is the signal-to-noise ratio of a matched filter, $\mathrm{SNR} = A\sqrt{\boldsymbol{s}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{s}}$. For white noise $\mathbf{C} = \sigma^2\mathbb{1}$ this is $\sqrt{\sum_k s_k^2}/\sigma$ times A: every sample adds the square of its signal-to-noise. Part 10b uses exactly this with $\mathbf{C}$ replaced by the noise power spectrum of a gravitational-wave detector.

Example: estimating a variance (covariance channel only). Let N numbers be drawn independently from $\mathcal{N}(0, v)$ and let the parameter be the variance v. Then $\boldsymbol{\mu} = \mathbf{0}$, $\mathbf{C} = v\mathbb{1}$, $\mathbf{C}^{-1}\mathbf{C}_{,v} = \mathbb{1}/v$, and

$$F_{vv} = \tfrac{1}{2}\,\mathrm{Tr}\big[(\mathbb{1}/v)^2\big] = \frac{N}{2v^2}, \qquad \sigma_v = v\sqrt{\frac{2}{N}}.$$

A variance estimated from N independent Gaussian numbers has a fractional error $\sqrt{2/N}$, however precisely each number is measured. This is the origin of cosmic variance: the $2\ell+1$ coefficients $a_{\ell m}$ at one multipole are $N = 2\ell+1$ such numbers with variance C_ℓ, so $\sigma(C_\ell)/C_\ell = \sqrt{2/(2\ell+1)}$ (Chapter 7). In the parameter $\sigma = \sqrt{v}$ the same information reads $F_{\sigma\sigma} = (\mathrm{d}v/\mathrm{d}\sigma)^2 F_{vv} = 2N/\sigma^2$, the value found directly in Section 10.1.2; this rescaling is a first instance of the Jacobian rule of Section 10.4.5.

Example: both channels at once. Suppose each of n points is $\mathcal{N}(\theta, \theta^2)$: a quantity whose noise grows in proportion to its size. The mean channel gives $1/\theta^2$ per point; in the covariance channel $\mathbf{C} = \theta^2$, $\mathbf{C}^{-1}\mathbf{C}_{,\theta} = 2/\theta$, and $\frac{1}{2}(2/\theta)^2 = 2/\theta^2$. In total $F = 3n/\theta^2$: here the spread of the data carries twice as much information as their average position. With $n = 10$ and $\theta = 2$, the variance of the score (10.21) over 200000 simulated data sets is 7.566, against $3n/\theta^2 = 7.500$.

The same example contains a warning. A Poisson count with mean λ is often approximated by $\mathcal{N}(\lambda, \lambda)$. The formula then gives $1/\lambda$ from the mean and $1/2\lambda^2$ from the variance, while the exact Poisson information is $1/\lambda$ (Section 3.12.3). The Gaussian approximation invents the extra $1/2\lambda^2$: it pretends that the observed scatter is an independent measurement of λ, which for a Poisson variable it is not. For large λ the invented part is negligible; for a few counts per bin it is not, and a forecast built that way is too optimistic.

Example: a correlated time series in Python. A more realistic case separates the two channels cleanly. Forty samples d_k at times t_k spread over two periods carry a signal $A \sin t_k$ on top of correlated noise plus white noise,

$$\mu_k = A \sin t_k, \qquad C_{kl} = s^2 e^{-|t_k - t_l|/\ell_c} + \sigma_n^2 \delta_{kl},$$

with $A = 1$, $s = 0.8$, $\ell_c = 1.5$ and $\sigma_n = 0.5$. The amplitude A is in the mean only, the correlation length ℓ_c is in the covariance only, so (10.20) has no cross term: $F_{AA} = 8.97$, $F_{\ell_c\ell_c} = 1.30$, $F_{A\ell_c} = 0.00$. The script draws 20000 series, evaluates the score (10.21) for each, and finds a covariance with diagonal $(8.85, 1.33)$ and off-diagonal -0.01, equal to the formula within the simulation noise of one to two per cent (Figure 10.3, right).

The same script asks a design question: what if we sampled the same two periods more densely? Figure 10.8 gives the answer. For white noise the error on A would fall as $n^{-1/2}$. Here it falls from 0.420 at 10 samples to only 0.312 at 320, because the correlated part of the noise is common to neighbouring samples and does not average down: once the samples are much closer than ℓ_c, a new sample repeats the noise of its neighbour. The correlation length itself keeps improving (from 1.67 to 0.45), because denser sampling resolves the shape of the correlation at short separations, which is exactly where ℓ_c shows. Before the data are taken, the Fisher matrix tells us which of the two quantities a denser sampling would actually buy.

Key idea: The Fisher matrix of Gaussian data

For $\boldsymbol{d} \sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{\theta}), \mathbf{C}(\boldsymbol{\theta}))$, $F_{ij} = \boldsymbol{\mu}_{,i}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{\mu}_{,j} + \frac{1}{2}\,\mathrm{Tr}[\mathbf{C}^{-1}\mathbf{C}_{,i}\mathbf{C}^{-1}\mathbf{C}_{,j}]$. Only first derivatives of the model enter. Parameters in the mean contribute the squared shift of the mean in units of the noise; parameters in the covariance contribute the squared fractional change of the scatter, with a factor one half.

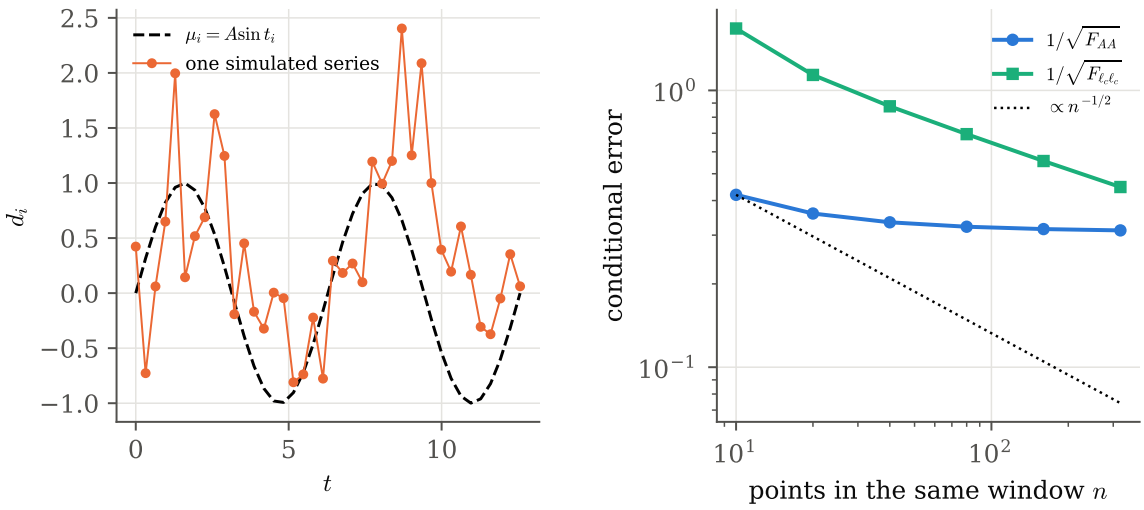


Figure 10.8: Left: one simulated series (points) and its mean $A\sin t$ (dashed). Right: the conditional errors $1/\sqrt{F_{ii}}$ on A and ℓ_c when the same window is sampled with more and more points. Against the white-noise expectation $n^{-1/2}$ (dotted), the amplitude error stalls, because correlated noise does not average down, while the correlation length keeps improving.

10.4 Reading a Fisher matrix

A Fisher matrix for six cosmological parameters is thirty-six numbers. On its own it answers nothing; the questions people actually ask are of a different kind. How well is n_s measured if everything else is also free? Which combination of parameters does the experiment really pin down? What if we add the measurement of τ that another experiment already has? What is the error on H_0, which is not one of the parameters we differentiated with respect to? Each question has a one-line answer in terms of $\mathbf{F}$, and each answer comes from treating $\mathbf{F}$ as what it approximately is: the inverse covariance of a Gaussian in parameter space, $\mathcal{L}(\boldsymbol{\theta}) \propto \exp[-\tfrac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^{\mathsf{T}}\mathbf{F}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)]$. This section collects those answers and the reasons for them, then condenses them into a recipe.

10.4.1 Marginal and conditional errors

The question. We want the error on one group of parameters $\boldsymbol{a}$ (say n_s), while the others, $\boldsymbol{b}$, are also unknown. "Also unknown" can mean two different things. Either the other parameters are estimated from the same data and we do not care about them: we *marginalise*, averaging the Gaussian over all their values, as in Chapter 1 [integrating a joint distribution over the variables we do not want, to isolate the one we do]. Or someone tells us their values: we *condition* on them, taking a slice of the Gaussian at the known values. Which width does each give?

Split $\boldsymbol{\delta} = \boldsymbol{\theta} - \boldsymbol{\theta}_0$ into $(\boldsymbol{\delta}_a, \boldsymbol{\delta}_b)$ and $\mathbf{F}$ into blocks $\mathbf{F}_{aa}$, $\mathbf{F}_{ab}$, $\mathbf{F}_{ba} = \mathbf{F}_{ab}^{\mathsf{T}}$, $\mathbf{F}_{bb}$. The exponent of the Gaussian is

$$\begin{aligned}\boldsymbol{\delta}^{\mathsf{T}}\mathbf{F}\boldsymbol{\delta} &\overset{(1)}{=} \boldsymbol{\delta}_a^{\mathsf{T}}\mathbf{F}_{aa}\boldsymbol{\delta}_a + 2\boldsymbol{\delta}_a^{\mathsf{T}}\mathbf{F}_{ab}\boldsymbol{\delta}_b + \boldsymbol{\delta}_b^{\mathsf{T}}\mathbf{F}_{bb}\boldsymbol{\delta}_b \\ &\overset{(2)}{=} \big(\boldsymbol{\delta}_b + \mathbf{F}_{bb}^{-1}\mathbf{F}_{ba}\boldsymbol{\delta}_a\big)^{\mathsf{T}}\mathbf{F}_{bb}\big(\boldsymbol{\delta}_b + \mathbf{F}_{bb}^{-1}\mathbf{F}_{ba}\boldsymbol{\delta}_a\big) + \boldsymbol{\delta}_a^{\mathsf{T}}\big(\mathbf{F}_{aa} - \mathbf{F}_{ab}\mathbf{F}_{bb}^{-1}\mathbf{F}_{ba}\big)\boldsymbol{\delta}_a. \end{aligned} \tag{10.22}$$

(1) Multiply out the block form; the two cross terms are equal because $\mathbf{F}$ is symmetric.
(2) Complete the square in $\boldsymbol{\delta}_b$: expanding the first product gives back $\boldsymbol{\delta}_b^{\mathsf{T}}\mathbf{F}_{bb}\boldsymbol{\delta}_b + 2\boldsymbol{\delta}_a^{\mathsf{T}}\mathbf{F}_{ab}\boldsymbol{\delta}_b + \boldsymbol{\delta}_a^{\mathsf{T}}\mathbf{F}_{ab}\mathbf{F}_{bb}^{-1}\mathbf{F}_{ba}\boldsymbol{\delta}_a$, and the last piece is subtracted again in the second term.

Conditioning sets $\boldsymbol{\delta}_b = \mathbf{0}$ in the original exponent: what remains is $\boldsymbol{\delta}_a^{\mathsf{T}}\mathbf{F}_{aa}\boldsymbol{\delta}_a$, so the conditional covariance of $\boldsymbol{a}$ is $\mathbf{F}_{aa}^{-1}$. *Marginalising* integrates over $\boldsymbol{\delta}_b$. In the second line of (10.22) the first term is a Gaussian in $\boldsymbol{\delta}_b$ centred on $-\mathbf{F}_{bb}^{-1}\mathbf{F}_{ba}\boldsymbol{\delta}_a$ with a width that does not depend on $\boldsymbol{\delta}_a$, so its integral is a constant. The marginal is a Gaussian in $\boldsymbol{\delta}_a$ with inverse covariance

$$\widetilde{\mathbf{F}}_{aa} = \mathbf{F}_{aa} - \mathbf{F}_{ab}\mathbf{F}_{bb}^{-1}\mathbf{F}_{ba}, \tag{10.23}$$

the **Schur complement** of $\mathbf{F}_{bb}$ [what is left of the information on $\boldsymbol{a}$ after the part that could be confused with changes of $\boldsymbol{b}$ is removed]. The same block also appears when we invert $\mathbf{F}$: multiplying out shows that

$$\mathbf{F}\begin{pmatrix} \widetilde{\mathbf{F}}_{aa}^{-1} \\ -\mathbf{F}_{bb}^{-1}\mathbf{F}_{ba}\widetilde{\mathbf{F}}_{aa}^{-1} \end{pmatrix} = \begin{pmatrix} (\mathbf{F}_{aa} - \mathbf{F}_{ab}\mathbf{F}_{bb}^{-1}\mathbf{F}_{ba})\widetilde{\mathbf{F}}_{aa}^{-1} \\ (\mathbf{F}_{ba} - \mathbf{F}_{bb}\mathbf{F}_{bb}^{-1}\mathbf{F}_{ba})\widetilde{\mathbf{F}}_{aa}^{-1} \end{pmatrix} = \begin{pmatrix} \mathbb{1} \\ \mathbf{0} \end{pmatrix},$$

so the aa block of $\mathbf{F}^{-1}$ is $\widetilde{\mathbf{F}}_{aa}^{-1}$. In words: *to marginalise, invert the whole Fisher matrix and read off the block you want* [HobsonBMC p. 107], [194]. The same block-inverse formula gave conditional and marginal variances of a multivariate Gaussian in Chapter 2; here the roles of covariance and inverse covariance are exchanged.

Working definition 10.2 (marginalised and conditional errors)

The **marginalised error** on θ_i is $\sigma_i = \sqrt{(\mathbf{F}^{-1})_{ii}}$, the error when all other parameters are estimated from the same data. The **conditional error** is $1/\sqrt{F_{ii}}$, the error if all other parameters were known exactly. *Operationally:* invert the full matrix and take the square root of the diagonal; never take the inverse square root of the diagonal of $\mathbf{F}$ itself unless the other parameters really are fixed by something outside the experiment.

Since $\mathbf{F}_{ab}\mathbf{F}_{bb}^{-1}\mathbf{F}_{ba}$ is positive semidefinite, $\widetilde{\mathbf{F}}_{aa} \preceq \mathbf{F}_{aa}$: marginalising can only lose information, which is the matrix form of (10.12). For two parameters the loss has a simple form. The 2×2 inverse (10.7), written with indices $1,2$, has $(\mathbf{F}^{-1})_{11} = F_{22}/\det\mathbf{F}$, $(\mathbf{F}^{-1})_{22} = F_{11}/\det\mathbf{F}$ and $(\mathbf{F}^{-1})_{12} = -F_{12}/\det\mathbf{F}$, with $\det\mathbf{F} = F_{11}F_{22} - F_{12}^2$. Then

$$\begin{aligned}
(\mathbf{F}^{-1})_{11} &\overset{(1)}{=} \frac{F_{22}}{F_{11}F_{22} - F_{12}^2} = \frac{1}{F_{11} - F_{12}^2/F_{22}}, \\
\rho &\overset{(2)}{=} \frac{(\mathbf{F}^{-1})_{12}}{\sqrt{(\mathbf{F}^{-1})_{11}(\mathbf{F}^{-1})_{22}}} = \frac{-F_{12}/\det\mathbf{F}}{\sqrt{F_{22}F_{11}}/\det\mathbf{F}} = -\frac{F_{12}}{\sqrt{F_{11}F_{22}}}, \\
F_{11} - \frac{F_{12}^2}{F_{22}} &\overset{(3)}{=} F_{11}\left(1 - \frac{F_{12}^2}{F_{11}F_{22}}\right) = F_{11}\,(1-\rho^2).
\end{aligned}$$

(1) Divide numerator and denominator by F_{22}. The result is the Schur complement (10.23) with $\boldsymbol{a} = \theta_1$ and $\boldsymbol{b} = \theta_2$, as it must be.

(2) The correlation coefficient is the covariance over the product of the two standard deviations; the determinant cancels.

(3) Take F_{11} out, and recognise $F_{12}^2/(F_{11}F_{22}) = \rho^2$ from the line above.

With the conditional error $\sigma_1^{\text{cond}} = 1/\sqrt{F_{11}}$, the marginalised error is therefore

$$\sigma_1^{\text{marg}} = \frac{1}{\sqrt{F_{11} - F_{12}^2/F_{22}}} = \frac{\sigma_1^{\text{cond}}}{\sqrt{1-\rho^2}}, \tag{10.24}$$

Sivia's (3.21) in Fisher notation [Sivia (3.21)–(3.22)]. A correlation of 0.9 more than doubles the error, 0.99 multiplies it by seven.

Figure 10.9 shows the geometry, which Cowan describes in words [Cowan p. 83]. Take the ellipse $\boldsymbol{\delta}^{\mathsf{T}}\mathbf{F}\boldsymbol{\delta} = 1$. *The chord.* On the horizontal line through the centre, $\delta_2 = 0$, the equation is $F_{11}\delta_1^2 = 1$, so the chord ends at $\delta_1 = \pm 1/\sqrt{F_{11}} = \pm\sigma_1^{\text{cond}}$. *The tangents.* The completed square (10.22), with one parameter in each group, writes the same equation as

$$F_{22}\left(\delta_2 + \frac{F_{12}}{F_{22}}\delta_1\right)^2 + \left(F_{11} - \frac{F_{12}^2}{F_{22}}\right)\delta_1^2 = 1.$$

The first term is never negative, so every point of the ellipse has $(F_{11} - F_{12}^2/F_{22})\,\delta_1^2 \le 1$, that is $|\delta_1| \le \sigma_1^{\text{marg}}$; the bound is reached where the first term vanishes, $\delta_2 = -F_{12}\delta_1/F_{22}$. So the ellipse reaches out to $\delta_1 = \pm\sigma_1^{\text{marg}}$ and no further: its vertical tangents sit there. In the angular-distribution experiment of Section 10.1.3 the two errors on α were 0.052 and 0.047 (Figure 10.4).

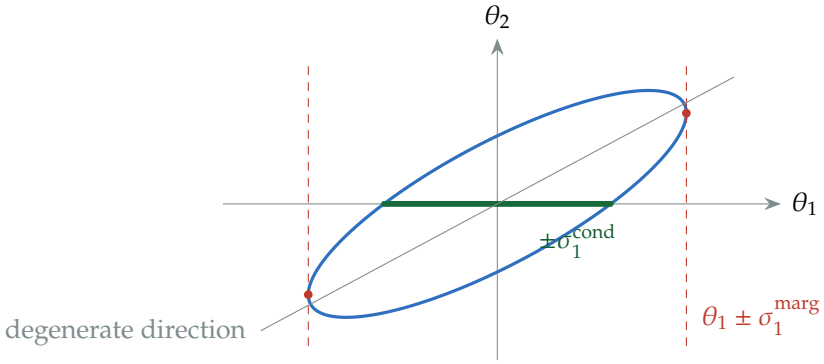


Figure 10.9: The ellipse $\boldsymbol{\delta}^{\mathsf{T}}\mathbf{F}\boldsymbol{\delta} = 1$ for two parameters with correlation $\rho = 0.8$. If θ_2 is known, the error on θ_1 is the half-chord at $\theta_2 = \theta_2^0$ (green). If θ_2 is free, the error is the half-width of the shadow of the ellipse on the θ_1 axis (red tangents), larger by $1/\sqrt{1-\rho^2}$. The long axis (grey) is the combination the data cannot pin down.

Watch out: Quoting the conditional error

$1/\sqrt{F_{ii}}$ is the error on θ_i with every other parameter fixed at its true value. Quoting it as "the error" of a forecast is the classic way to overstate an experiment. In the CMB forecast of Section 10.6, the conditional errors on τ and $\ln A_s$ are fifteen times smaller than the marginalised ones, and the error on ω_c three times smaller.

Nuisance parameters. A calibration factor, a foreground amplitude, a detector gain: parameters we must fit but do not care about [195]. In a forecast they are included in $\mathbf{F}$ like any other parameter and then marginalised by inverting. Fixing them instead ("assume the calibration is perfect") gives the conditional error, which is wrong unless the nuisance really is known. For a Gaussian likelihood, marginalising and profiling agree (Section 3.14.4), so the Fisher answer serves both schools. Section 10.6.3 shows a single nuisance parameter, Planck's calibration, accounting for most of the gap between a naive forecast and the published error on $A_s e^{-2\tau}$.

10.4.2 Confidence ellipses: where 2.30 and 6.18 come from

A two-parameter forecast is usually drawn as an ellipse, and a reader needs to know how much probability it encloses. For a pair of parameters, marginalise the rest. By (10.23) and the inversion rule that follows it, the marginal distribution of the pair is a Gaussian whose covariance is the corresponding 2×2 block of $\mathbf{F}^{-1}$; call that block $\boldsymbol{\Sigma}$. Its inverse defines $\Delta\chi^2 = \boldsymbol{\delta}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$ for the two-component $\boldsymbol{\delta}$ [Cowan (9.31)–(9.32)].

The question. What fraction p of the Gaussian lies inside the ellipse $\Delta\chi^2 \le q$? *Whiten the variables.* Factor the covariance as $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^{\mathsf{T}}$ with $\boldsymbol{L}$ lower triangular (the Cholesky factor) and set $\boldsymbol{z} = \boldsymbol{L}^{-1}\boldsymbol{\delta}$, so $\boldsymbol{\delta} = \boldsymbol{L}\boldsymbol{z}$. A linear map transforms a covariance by sandwiching it, so

$$\operatorname{Cov}\boldsymbol{z} = \boldsymbol{L}^{-1}\boldsymbol{\Sigma}\,\boldsymbol{L}^{-\mathsf{T}} = \boldsymbol{L}^{-1}\boldsymbol{L}\boldsymbol{L}^{\mathsf{T}}\boldsymbol{L}^{-\mathsf{T}} = \mathbb{1}.$$

The two components of $\boldsymbol{z}$ are Gaussian, have unit variance and are uncorrelated. For Gaussians that is enough for independence: with covariance $\mathbb{1}$ the density $\propto e^{-(z_1^2+z_2^2)/2}$ has no cross term, so it factorises into a function of z_1 times a function of z_2. The quadratic form becomes a sum of squares,

$$\Delta\chi^2 = (\boldsymbol{L}\boldsymbol{z})^{\mathsf{T}}(\boldsymbol{L}\boldsymbol{L}^{\mathsf{T}})^{-1}(\boldsymbol{L}\boldsymbol{z}) = \boldsymbol{z}^{\mathsf{T}}\boldsymbol{L}^{\mathsf{T}}\boldsymbol{L}^{-\mathsf{T}}\boldsymbol{L}^{-1}\boldsymbol{L}\boldsymbol{z} = \boldsymbol{z}^{\mathsf{T}}\boldsymbol{z} = z_1^2 + z_2^2 \equiv r^2,$$

using $(\boldsymbol{L}\boldsymbol{L}^{\mathsf{T}})^{-1} = \boldsymbol{L}^{-\mathsf{T}}\boldsymbol{L}^{-1}$: a squared radius in the $\boldsymbol{z}$ plane. Then

$$\mathrm{P}(\Delta\chi^2 \le q) \overset{(1)}{=} \int_{r^2\le q} \frac{1}{2\pi} e^{-(z_1^2+z_2^2)/2}\,\mathrm{d}z_1\mathrm{d}z_2 \overset{(2)}{=} \int_0^{2\pi}\int_0^{\sqrt{q}} \frac{1}{2\pi} e^{-r^2/2}\, r\,\mathrm{d}r\,\mathrm{d}\phi \overset{(3)}{=} 1 - e^{-q/2}, \qquad (10.25)$$

(1) The joint density of two independent standard normals.
(2) Polar coordinates, $\mathrm{d}z_1\mathrm{d}z_2 = r\,\mathrm{d}r\,\mathrm{d}\phi$.
(3) The ϕ integral gives 2π; $\int_0^{\sqrt{q}} re^{-r^2/2}\mathrm{d}r = [-e^{-r^2/2}]_0^{\sqrt{q}}$.

So $q = -2\ln(1-p)$. For the probability inside $\pm1\sigma$ of a one-dimensional Gaussian, $p = 0.6827$, this gives $q = 2.30$; for $\pm2\sigma$, $p = 0.9545$, $q = 6.18$; for $\pm3\sigma$, $p = 0.9973$, $q = 11.8$.[2] These are the levels in Figure 10.4, and the simulation is a test of (10.25): 0.704 of Cowan's 500 fits fall inside the 2.30 ellipse and 0.974 inside the 6.18 one, against 0.683 and 0.954, with binomial uncertainties ±0.021 and ±0.009.

Two ellipses for two questions. The ellipse $\Delta\chi^2 = 1$ is the one whose tangents mark $\pm\sigma_i^{\mathrm{marg}}$ (Figure 10.9); it is Cowan's contour $\ln\mathcal{L} = \ln\mathcal{L}_{\max} - \frac{1}{2}$ [Cowan (6.25), (6.31)]. But it contains only $1 - e^{-1/2} = 39\%$ of the two-dimensional probability. The ellipse that contains 68.3% is the larger one, $\Delta\chi^2 = 2.30$. How far does it reach along each axis? The tangent argument of Section 10.4.1 applies to any ellipse: on $\boldsymbol{\delta}^{\mathsf{T}}\boldsymbol{M}\boldsymbol{\delta} = q$ the largest $|\delta_1|$ is $\sqrt{q}$ divided by the square root of the Schur complement of $\boldsymbol{M}$, and by the inversion rule after (10.23) the inverse of that Schur complement is $(\boldsymbol{M}^{-1})_{11}$. Here $\boldsymbol{M} = \boldsymbol{\Sigma}^{-1}$, so $(\boldsymbol{M}^{-1})_{11} = \Sigma_{11} = \sigma_1^2$, and the ellipse $\Delta\chi^2 = q$ casts a shadow $\pm\sqrt{q}\,\sigma_i$ on each axis. For $q = 1$ that is $\pm\sigma_i$, as stated; for $q = 2.30$ it is $\pm\sqrt{2.30}\,\sigma_i = \pm1.52\,\sigma_i$. Both are correct; they answer different questions (one parameter at a time, or the pair jointly), and a figure must say which it shows. With m parameters jointly the levels come from χ^2_m [Cowan Table 9.4]; for $m = 6$ the 68.3% region is $\Delta\chi^2 \le 7.04$.

Drawing the ellipse. To draw it we need its axes, their lengths and its tilt. Write the pair covariance as $\boldsymbol{\Sigma} = \left(\begin{smallmatrix} a & b \\ b & c \end{smallmatrix}\right)$ with $a = \sigma_x^2$, $c = \sigma_y^2$, $b = \sigma_{xy}$. *The lengths.* The eigenvalues λ of $\boldsymbol{\Sigma}$ solve $\det(\boldsymbol{\Sigma} - \lambda\mathbb{1}) = (a-\lambda)(c-\lambda) - b^2 = 0$, that is $\lambda^2 - (a+c)\lambda + (ac - b^2) = 0$, the trace and the determinant as coefficients. The quadratic formula gives

$$\lambda_\pm = \frac{a+c}{2} \pm \sqrt{\left(\frac{a+c}{2}\right)^2 - ac + b^2} = \frac{\sigma_x^2+\sigma_y^2}{2} \pm \sqrt{\frac{(\sigma_x^2-\sigma_y^2)^2}{4} + \sigma_{xy}^2}.$$

Let $\boldsymbol{R}$ be the rotation whose columns are the two unit eigenvectors, so $\boldsymbol{\Sigma} = \boldsymbol{R}\,\mathrm{diag}(\lambda_+,\lambda_-)\boldsymbol{R}^{\mathsf{T}}$ and $\boldsymbol{\Sigma}^{-1} = \boldsymbol{R}\,\mathrm{diag}(1/\lambda_+, 1/\lambda_-)\boldsymbol{R}^{\mathsf{T}}$. In the rotated coordinates $(u,v) = \boldsymbol{R}^{\mathsf{T}}\boldsymbol{\delta}$ the ellipse is

[2] Coe [194] and others print 6.17; $-2\ln(1-0.9545) = 6.180$.

$u^2/\lambda_+ + v^2/\lambda_- = q$, an upright ellipse with half-axes $\sqrt{q\lambda_+}$ and $\sqrt{q\lambda_-}$. *The tilt.* With $\boldsymbol{R}$ the rotation by the angle ϑ, the matrix $\boldsymbol{R}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{R}$ must be diagonal. Its off-diagonal entry is $(c-a)\sin\vartheta\cos\vartheta + b(\cos^2\vartheta - \sin^2\vartheta) = \frac{c-a}{2}\sin 2\vartheta + b\cos 2\vartheta$, and setting it to zero gives

$$\tan 2\vartheta = \frac{2b}{a-c} = \frac{2\sigma_{xy}}{\sigma_x^2 - \sigma_y^2}$$

for the angle ϑ of the long axis [194], [Cowan (6.32)]. In code it is simpler to map a unit circle through the Cholesky factor, $\boldsymbol{\delta} = \sqrt{q}\,\boldsymbol{L}(\cos t, \sin t)^\mathsf{T}$: by the whitening above this is exactly the set $\boldsymbol{z}^\mathsf{T}\boldsymbol{z} = q$, and it is how every ellipse in the figures is drawn.

10.4.3 Degeneracy directions

The long axis of the ellipse is a combination of parameters the data barely constrain; the short axis is a combination they pin down. For many parameters the same information is in the eigenvectors of $\mathbf{F}$. Write $\mathbf{F} = \boldsymbol{V}\,\mathrm{diag}(\lambda_1,\dots,\lambda_m)\boldsymbol{V}^\mathsf{T}$ with orthonormal eigenvectors $\boldsymbol{v}_a$ as the columns of $\boldsymbol{V}$, as in Section 10.2.2. Since $\boldsymbol{V}^\mathsf{T}\boldsymbol{V} = \mathbb{1}$, the inverse is $\mathbf{F}^{-1} = \boldsymbol{V}\,\mathrm{diag}(1/\lambda_1,\dots,1/\lambda_m)\boldsymbol{V}^\mathsf{T}$, and $\boldsymbol{V}^\mathsf{T}\boldsymbol{v}_a$ is the unit vector $\boldsymbol{e}_a$. The variance of the combination $\boldsymbol{v}_a^\mathsf{T}\boldsymbol{\theta}$, and its covariance with another such combination, follow from the rule $\mathrm{Cov}(\boldsymbol{u}^\mathsf{T}\boldsymbol{\theta}, \boldsymbol{w}^\mathsf{T}\boldsymbol{\theta}) = \boldsymbol{u}^\mathsf{T}\mathbf{F}^{-1}\boldsymbol{w}$:

$$\mathrm{Var}(\boldsymbol{v}_a^\mathsf{T}\boldsymbol{\theta}) = \boldsymbol{v}_a^\mathsf{T}\mathbf{F}^{-1}\boldsymbol{v}_a = \boldsymbol{e}_a^\mathsf{T}\,\mathrm{diag}(1/\lambda)\,\boldsymbol{e}_a = \frac{1}{\lambda_a},$$

$$\mathrm{Cov}(\boldsymbol{v}_a^\mathsf{T}\boldsymbol{\theta}, \boldsymbol{v}_b^\mathsf{T}\boldsymbol{\theta}) = \boldsymbol{e}_a^\mathsf{T}\,\mathrm{diag}(1/\lambda)\,\boldsymbol{e}_b = 0 \quad (a \neq b).$$

So each eigen-combination is measured with error $1/\sqrt{\lambda_a}$, the combinations are mutually uncorrelated [Sivia (3.20), Fig. 3.7], [192], and because the calculation used $\mathbf{F}^{-1}$, this is the marginalised error, with every other combination free. A very small eigenvalue is a **degeneracy**: a direction along which the model hardly changes the data. One caution: the eigenvectors depend on the units of the parameters (rescale θ_1 by 1000 and the directions turn), so they are only meaningful for parameters of comparable kind, or after each parameter is scaled by its expected error.

Example: the straight line and its pivot. Experiment A measures $y = a + bx$ at ten points evenly spread over $0 \le x \le 1$, each with Gaussian noise $\sigma = 0.1$. With $\boldsymbol{\mu}_{,a} = (1,\dots,1)$ and $\boldsymbol{\mu}_{,b} = (x_1,\dots,x_{10})$, the formula (10.20) gives

$$\mathbf{F} = \frac{1}{\sigma^2}\begin{pmatrix} \sum_k 1 & \sum_k x_k \\ \sum_k x_k & \sum_k x_k^2 \end{pmatrix},$$

the matrix of the normal equations (Section 9.17.2). The intercept and slope come out strongly anticorrelated, $\rho = -0.84$, with $\sigma_a = 0.059$, $\sigma_b = 0.099$. The reason is a lever arm: all data sit to the right of $x = 0$, so a steeper line is compensated by a lower intercept. The combination the data really fix is the height of the line in the middle of the data, $y_p = a + bx_p$, for a suitable point x_p. *Which point?* The one at which y_p and b are uncorrelated. Covariance is linear in each argument and x_p is a fixed number, so

$$0 = \mathrm{Cov}(y_p, b) \overset{(1)}{=} \mathrm{Cov}(a,b) + x_p\,\mathrm{Var}(b) \quad\Rightarrow\quad x_p = -\frac{(\mathbf{F}^{-1})_{ab}}{(\mathbf{F}^{-1})_{bb}} \overset{(2)}{=} \frac{F_{ab}}{F_{aa}} \overset{(3)}{=} \frac{\sum_k x_k}{n} = 0.50.$$

(1) The covariances come from $\mathbf{F}^{-1}$: $\mathrm{Cov}(a,b) = (\mathbf{F}^{-1})_{ab}$ and $\mathrm{Var}(b) = (\mathbf{F}^{-1})_{bb}$.

(2) The 2×2 inverse (10.7): $(\mathbf{F}^{-1})_{ab} = -F_{ab}/\det\mathbf{F}$ and $(\mathbf{F}^{-1})_{bb} = F_{aa}/\det\mathbf{F}$; the determinant cancels.

(3) Read the entries off the matrix above, with $n = 10$ points; the common factor $1/\sigma^2$ cancels.

So the pivot is the mean of the x_k. Each point carries the same information here, because all have the same noise; with unequal noise the same steps give the mean of the x_k weighted by $1/\sigma_k^2$, the information each point carries. That is the general statement: the pivot sits at the information-weighted mean of the variable along which the line is drawn. Section 8.7.3 found the same statement for the CMB, with $\ln k$ in place of x and the information of each multipole as the weight, and Section 10.6.2 uses it again.

The pivot as new parameters. Solving $y_p = a + bx_p$ for the old parameter gives $a = y_p - bx_p$, while b stays b. These are the functions $\boldsymbol{\theta}(\boldsymbol{p})$ that Section 10.4.5 turns into a Jacobian, and in the variables (y_p, b) the errors are uncorrelated. This is the reason the CMB amplitude A_s is quoted at a pivot wavenumber.

10.4.4 Priors and independent experiments add

Two independent experiments A and B measuring the same parameters have a joint likelihood $\mathcal{L} = \mathcal{L}_A \mathcal{L}_B$. The log-likelihoods add, so their second derivatives add, so

$$\mathbf{F}_{A+B} = \mathbf{F}_A + \mathbf{F}_B. \tag{10.26}$$

A Gaussian prior $p(\boldsymbol{\theta}) \propto \exp[-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\mathsf{T}} \boldsymbol{P} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)]$ multiplies the likelihood in Bayes' theorem, so in the exponent it adds: the posterior has inverse covariance $\mathbf{F} + \boldsymbol{P}$ [HobsonBMC p. 107]. A prior is one more experiment, and a prior of width σ_p on one parameter adds $1/\sigma_p^2$ to one diagonal element. Two rules make this work. Add the matrices *before* inverting, and write them for the same set of parameters in the same units (padding with zeros a parameter an experiment does not see).

Example: two experiments on one line. Experiment B measures the same line at ten points over $2 \le x \le 3$. On its own it is even more degenerate, $\rho = -0.99$, because its lever arm from $x = 0$ is longer, and its ellipse is tilted at a different angle from A's (Figure 10.10, left). The sum of the two Fisher matrices gives $\sigma_a = 0.039$ and $\sigma_b = 0.021$: the slope is 4.7 times better than either experiment alone. A naive combination of the two *marginalised* slope errors, $1/\sqrt{1/0.099^2 + 1/0.099^2} = 0.070$, would have missed most of the gain, because the long axes of the two ellipses point in different directions and each experiment measures well what the other measures badly.

Example: a prior on the intercept. Suppose a calibration measurement fixes the intercept to ± 0.02. Adding $1/0.02^2$ to F_{aa} of experiment A gives $\sigma_b = 0.060$ instead of 0.099 (Figure 10.10, right): a prior on one parameter improves another through their correlation. This is how a measurement of τ from large-scale polarization improves the CMB amplitude in Section 10.6.

Why the matrices, not the errors, are combined. Two forecasts quoted as marginalised errors cannot be combined by adding inverse variances: the correlations are gone, and with them the gain from crossing degeneracies. The full Fisher matrices are kept, published and added, and marginalisation comes last.

10.4.5 Changing parameters: the Jacobian sandwich

The Fisher matrix is computed in the parameters the derivatives were taken with respect to, but we may want the answer in others: H_0 instead of an angle, $\ln A_s$ instead of A_s, the pivot

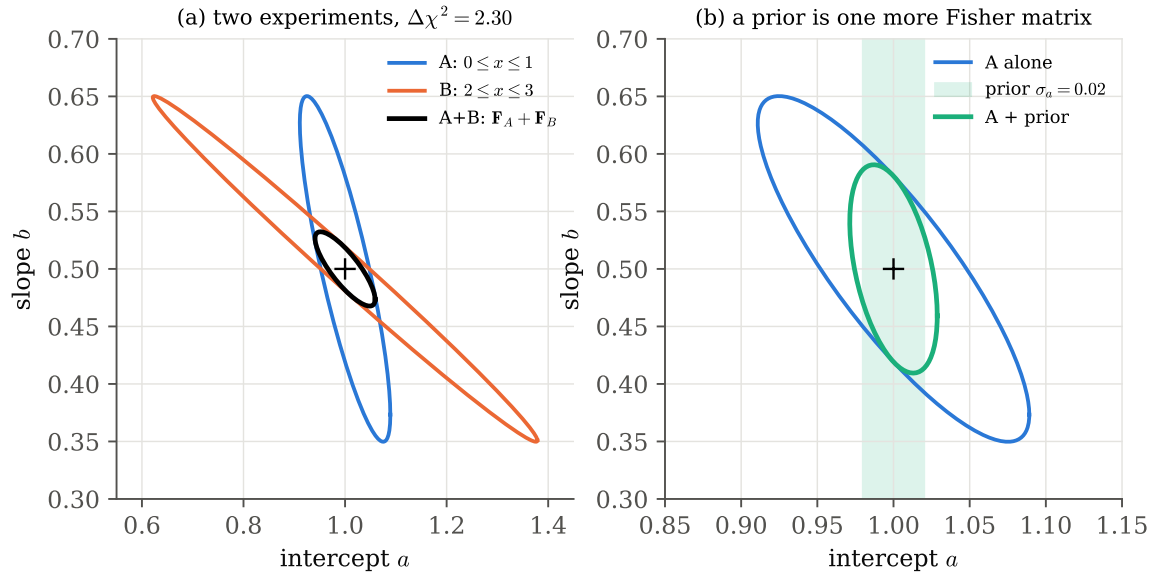


Figure 10.10: Left: two experiments measure the same line over different ranges of x; each has an elongated 68% ellipse, tilted in its own direction. The sum of their Fisher matrices gives the small black ellipse, much smaller than either. Right: a Gaussian prior on the intercept (shaded band) is one more Fisher matrix, and through the correlation it also shrinks the slope.

amplitude instead of the intercept. Let $\boldsymbol{p}$ be the new parameters and $\boldsymbol{\theta}(\boldsymbol{p})$ the old ones as functions of the new, with Jacobian $J_{ia} = \partial\theta_i/\partial p_a$. The score transforms by the chain rule, and $\mathbf{F}$ is the covariance of the score:

$$S'_a = \frac{\partial \ln \mathcal{L}}{\partial p_a} \overset{(1)}{=} \sum_i J_{ia} S_i, \qquad F'_{ab} = \mathbb{E}[S'_a S'_b] \overset{(2)}{=} \sum_{ij} J_{ia}\, \mathbb{E}[S_i S_j]\, J_{jb}, \qquad \mathbf{F}' = \boldsymbol{J}^{\mathsf{T}} \mathbf{F} \boldsymbol{J}. \tag{10.27}$$

(1) The likelihood depends on $\boldsymbol{p}$ only through $\boldsymbol{\theta}(\boldsymbol{p})$.
(2) $\boldsymbol{J}$ is not random; take it out of the expectation.

[194], eq. (18). *The covariance transforms the other way.* Inverting $\mathbf{F}' = \boldsymbol{J}^{\mathsf{T}}\mathbf{F}\boldsymbol{J}$ reverses the order of the factors: $(\mathbf{F}')^{-1} = \boldsymbol{J}^{-1}\mathbf{F}^{-1}\boldsymbol{J}^{-\mathsf{T}}$, with $\boldsymbol{J}^{-1} = \partial\boldsymbol{p}/\partial\boldsymbol{\theta}$ the Jacobian of the new parameters with respect to the old. Read as covariances,

$$\mathrm{Cov}(\boldsymbol{p}) = \frac{\partial \boldsymbol{p}}{\partial \boldsymbol{\theta}}\, \mathbf{F}^{-1} \left(\frac{\partial \boldsymbol{p}}{\partial \boldsymbol{\theta}}\right)^{\mathsf{T}},$$

which is the rule $\mathrm{Cov}(\boldsymbol{A}\boldsymbol{\theta}) = \boldsymbol{A}\,\mathrm{Cov}(\boldsymbol{\theta})\boldsymbol{A}^{\mathsf{T}}$ for a linear map, applied to the linear part of $\boldsymbol{p}(\boldsymbol{\theta})$ near $\boldsymbol{\theta}_0$. *A single derived quantity.* Now let one component of $\boldsymbol{p}$ be a quantity $g(\boldsymbol{\theta})$ that is not one of the parameters, such as H_0 or $A_s e^{-2\tau}$ (the other components can be any quantities that complete the set; they do not enter its own variance). Its row of $\partial\boldsymbol{p}/\partial\boldsymbol{\theta}$ is the gradient $\nabla g^{\mathsf{T}} = (\partial_1 g, \ldots, \partial_m g)$, and the corresponding diagonal entry of $\mathrm{Cov}(\boldsymbol{p})$ is

$$\sigma_g^2 = \nabla g^{\mathsf{T}}\, \mathbf{F}^{-1}\, \nabla g = \sum_{ij} \partial_i g\, (\mathbf{F}^{-1})_{ij}\, \partial_j g,$$

the multiparameter delta method [Wasserman p. 134, Thm. 9.28] of Section 3.13.2. A check with two parameters and $g = \theta_1 + \theta_2$: $\nabla g = (1,1)$, so $\sigma_g^2 = (\mathbf{F}^{-1})_{11} + (\mathbf{F}^{-1})_{22} + 2(\mathbf{F}^{-1})_{12}$, the familiar variance of a sum, $\mathrm{Var}\,\theta_1 + \mathrm{Var}\,\theta_2 + 2\,\mathrm{Cov}(\theta_1, \theta_2)$. Two derived quantities g and g' have covariance $\nabla g^{\mathsf{T}}\mathbf{F}^{-1}\nabla g'$, an off-diagonal entry of the same matrix.[3] The transformation is exact

[3] In Wasserman's Example 9.29 the result for $\tau = \sigma/\mu$ is printed as $\frac{1}{\sqrt{n}}\sqrt{1/\hat\mu^4 + \hat\sigma^2/2\hat\mu^2}$; the first term should be $\hat\sigma^4/\hat\mu^4$, as the product of his gradient $(-\sigma/\mu^2, 1/\mu)$ and his J_n shows.

for the Fisher matrix, which is a local quantity at $\boldsymbol{\theta}_0$; whether the transformed likelihood is still close to Gaussian is a separate question, and a nonlinear change of variables can make it better or worse (Section 9.17.6).

Example: the pivot as a change of variables. For experiment A, $a = y_p - bx_p$ and $b = b$, so $\boldsymbol{J} = \left(\begin{smallmatrix} 1 & -x_p \\ 0 & 1 \end{smallmatrix}\right)$. The new matrix $\boldsymbol{J}^{\mathsf{T}}\mathbf{F}\boldsymbol{J}$ has errors $\sigma_{y_p} = 0.032$, $\sigma_b = 0.099$ and correlation +0.000. The slope error has not changed (the slope is the same parameter), but the amplitude is now quoted where it is measured, almost twice better than the intercept at $x = 0$.

10.4.6 One number for an experiment: figures of merit

To rank designs, a single number helps. Hobson et al. list the common choices [HobsonBMC pp. 108–109]. *The determinant* $\det \mathbf{F}$ ("D-optimality"): the ellipsoid $\boldsymbol{\delta}^{\mathsf{T}}\mathbf{F}\boldsymbol{\delta} = q$ has half-axes $\sqrt{q/\lambda_a}$ along the eigenvectors of $\mathbf{F}$, so its volume is proportional to $\prod_a \lambda_a^{-1/2} = (\det \mathbf{F})^{-1/2}$, and a large determinant means a small volume. *The trace of* $\mathbf{F}^{-1}$ ("A-optimality"): the sum of the marginalised variances.

The information gain measures how much the data narrow the prior, as the Kullback–Leibler divergence (3.100) of the posterior from the prior. Take a Gaussian prior with inverse covariance $\boldsymbol{P}$, so the posterior has inverse covariance $\mathbf{F} + \boldsymbol{P}$ (Section 10.4.4), and centre both at $\boldsymbol{\theta}_0$, as a forecast does. With $\boldsymbol{\delta} = \boldsymbol{\theta} - \boldsymbol{\theta}_0$ and $\mathbb{E}$ the average over the posterior,

$$
\begin{aligned}
D(\text{post}, \text{prior}) &\overset{(1)}{=} \mathbb{E}\big[\ln p_{\text{post}}(\boldsymbol{\delta}) - \ln p_{\text{prior}}(\boldsymbol{\delta})\big] \\
&\overset{(2)}{=} \tfrac{1}{2}\ln\det(\mathbf{F}+\boldsymbol{P}) - \tfrac{1}{2}\ln\det\boldsymbol{P} - \tfrac{1}{2}\,\mathbb{E}\big[\boldsymbol{\delta}^{\mathsf{T}}(\mathbf{F}+\boldsymbol{P})\boldsymbol{\delta}\big] + \tfrac{1}{2}\,\mathbb{E}\big[\boldsymbol{\delta}^{\mathsf{T}}\boldsymbol{P}\boldsymbol{\delta}\big] \\
&\overset{(3)}{=} \tfrac{1}{2}\big[\ln\det(\mathbf{F}+\boldsymbol{P}) - \ln\det\boldsymbol{P} - \operatorname{Tr}\big(\mathbb{1} - \boldsymbol{P}(\mathbf{F}+\boldsymbol{P})^{-1}\big)\big].
\end{aligned}
$$

(1) The divergence is the average, under the first density, of the log of the ratio of the two.
(2) A Gaussian with inverse covariance $\boldsymbol{M}$ has $\ln p = \frac{1}{2}\ln\det\boldsymbol{M} - \frac{1}{2}\boldsymbol{\delta}^{\mathsf{T}}\boldsymbol{M}\boldsymbol{\delta}$ plus a constant that is the same for both and cancels.
(3) For any matrix $\boldsymbol{M}$, $\mathbb{E}[\boldsymbol{\delta}^{\mathsf{T}}\boldsymbol{M}\boldsymbol{\delta}] = \mathbb{E}\operatorname{Tr}(\boldsymbol{M}\boldsymbol{\delta}\boldsymbol{\delta}^{\mathsf{T}}) = \operatorname{Tr}(\boldsymbol{M}\boldsymbol{\Sigma})$ with the posterior covariance $\boldsymbol{\Sigma} = (\mathbf{F}+\boldsymbol{P})^{-1}$; so the third term is $-\frac{1}{2}\operatorname{Tr}\mathbb{1}$ and the fourth $+\frac{1}{2}\operatorname{Tr}(\boldsymbol{P}(\mathbf{F}+\boldsymbol{P})^{-1})$.

If the data say nothing, $\mathbf{F} = \mathbf{0}$ and every term vanishes; the more the data shrink the prior volume, the larger the gain, mostly through the first two terms, which compare the two volumes.

The dark-energy figure of merit. The equation of state of dark energy is often written $w(a) = w_0 + w_a(1-a)$ with a the scale factor (cosmology notes): w_0 is its value today and w_a how fast it changes. The figure of merit is $1/\sqrt{\det \boldsymbol{\Sigma}_{w_0 w_a}}$, with $\boldsymbol{\Sigma}_{w_0 w_a}$ the marginalised 2×2 block of $\mathbf{F}^{-1}$. By the half-axes of Section 10.4.2, the ellipse $\Delta\chi^2 = q$ has area $\pi\sqrt{q\lambda_+}\sqrt{q\lambda_-} = \pi q\sqrt{\det\boldsymbol{\Sigma}}$, so the figure of merit is inversely proportional to the area of the (w_0, w_a) ellipse: it is the determinant criterion applied to that one pair.

Each number rewards a different shape: a determinant can be large for a long thin ellipse whose long axis is useless, which is why the full matrix, not a single number, is what should be compared.

10.4.7 The recipe

Working definition 10.3 (Fisher forecast)

1. Choose the parameters and a fiducial point $\boldsymbol{\theta}_0$ (the forecast depends on both).
2. Model the data: mean $\boldsymbol{\mu}(\boldsymbol{\theta})$ and covariance $\mathbf{C}(\boldsymbol{\theta})$ of the summary the analysis will use, including the instrument (noise, beam, sky fraction).
3. Differentiate the model at $\boldsymbol{\theta}_0$ (analytically, or by finite differences checked for stability, Section 10.5).
4. Assemble $\mathbf{F}$ by (10.20), or by (10.3) for non-Gaussian data.
5. Add the Fisher matrices of priors and of independent experiments.
6. Transform to the parameters you want to report, $\mathbf{F}' = \boldsymbol{J}^{\mathsf{T}}\mathbf{F}\boldsymbol{J}$.
7. Invert. Marginalised errors are $\sqrt{(\mathbf{F}^{-1})_{ii}}$; 2×2 blocks of $\mathbf{F}^{-1}$ give ellipses at $\Delta\chi^2 = 2.30,\ 6.18$; eigenvectors give degeneracies; $\nabla g^{\mathsf{T}}\mathbf{F}^{-1}\nabla g$ gives derived parameters.
8. Check: is the predicted error small compared with the scale on which the model is nonlinear? If not, validate with simulations or a chain (Section 10.6.6).

10.5 Derivatives from a code that cannot differentiate

Equation (10.20) needs the derivative of the model with respect to each parameter. For the CMB the model is a Boltzmann code, CAMB [8]: tens of thousands of lines that integrate the coupled photon, baryon, neutrino and dark-matter perturbations from the early universe to today and return $C_\ell(\boldsymbol{\theta})$ (the physics is in the cosmology notes; Chapter T1 maps the objects). It returns numbers, not derivatives. The only practical route is a finite difference, $[C_\ell(\theta+h)-C_\ell(\theta-h)]/2h$, and the whole forecast then hangs on a number we choose by hand, the step h.

The question. How large should the step be, and how do we know the forecast does not depend on it? Two errors pull in opposite directions. With a large step, the model is no longer linear between $\theta-h$ and $\theta+h$, and the difference quotient is not the derivative: a *truncation error*. With a small step, the two spectra are nearly equal, and whatever numerical noise the code has (from its finite grids in time and wavenumber, its interpolations, its convergence tolerances) is divided by the small number $2h$: a *noise error*.

10.5.1 Truncation against noise

Expand $f(\theta\pm h) = f \pm hf' + \frac{1}{2}h^2f'' \pm \frac{1}{6}h^3f''' + \frac{1}{24}h^4f'''' \pm \frac{1}{120}h^5f^{(5)} + \dots$ (all derivatives at θ). Then the two-point and four-point central differences are

$$
\begin{aligned}
D_2(h) &\equiv \frac{f(\theta+h)-f(\theta-h)}{2h} \overset{(1)}{=} f' + \frac{h^2}{6}f''' + \frac{h^4}{120}f^{(5)} + \dots, \\
D_4(h) &\equiv \frac{4D_2(h)-D_2(2h)}{3} \overset{(2)}{=} \frac{-f(\theta+2h)+8f(\theta+h)-8f(\theta-h)+f(\theta-2h)}{12h} \\
&\overset{(3)}{=} f' - \frac{h^4}{30}f^{(5)} + \dots
\end{aligned}
\tag{10.28}
$$

(1) Subtract the two series: the even powers cancel; divide by $2h$.

(2) Write out $D_2(h)$ and $D_2(2h)$ and put everything over $12h$.

(3) From line (1), $D_2(2h) = f' + \frac{4h^2}{6}f''' + \frac{16h^4}{120}f^{(5)}$; in $4D_2(h) - D_2(2h)$ the h^2 terms cancel ($4\cdot\frac{1}{6} - \frac{4}{6} = 0$) and the h^4 terms leave $\frac{4-16}{120}h^4 f^{(5)}$, which divided by 3 is $-h^4 f^{(5)}/30$.

The four-point rule is the two-point rule with its leading error extrapolated away (Richardson extrapolation), at the price of two more model evaluations. Chapter 9 used the two-point rule to build its emulator (Section 9.20.2).

Now add noise: suppose each evaluation of f carries an independent relative error of size ϵ, an absolute error of about $\epsilon\,|f|$. *The noise term.* In D_2 the two errors enter as a difference; being independent, they add in quadrature to about $\sqrt{2}\,\epsilon\,|f|$, and dividing by $2h$ leaves $\epsilon\,|f|\,/(\sqrt{2}\,h)$. We keep only its order, $\epsilon\,|f|\,/h$: a factor of order one changes no scaling. *The truncation term* is the first neglected term in line (1) of (10.28), $h^2\,|f'''|\,/6$. The total error is about

$$E(h) \approx \frac{\epsilon\,|f|}{h} + \frac{h^2}{6}\,|f'''|\,, \qquad h_{\rm opt} = \left(\frac{3\epsilon\,|f|}{|f'''|}\right)^{1/3}. \tag{10.29}$$

The first term falls with h and the second rises, so there is a best step in between, where the slope of E is zero:

$$\frac{\mathrm{d}E}{\mathrm{d}h} \overset{(1)}{=} -\frac{\epsilon\,|f|}{h^2} + \frac{h\,|f'''|}{3} = 0 \quad\Rightarrow\quad h^3 = \frac{3\epsilon\,|f|}{|f'''|},$$
$$\frac{\epsilon\,|f|}{h_{\rm opt}} \overset{(2)}{=} \frac{\epsilon^{2/3}\,|f|^{2/3}\,|f'''|^{1/3}}{3^{1/3}}, \qquad \frac{h_{\rm opt}^2}{6}\,|f'''| \overset{(3)}{=} \frac{3^{2/3}\epsilon^{2/3}\,|f|^{2/3}\,|f'''|^{1/3}}{6}.$$

(1) Differentiate the two terms of (10.29), set the sum to zero, and multiply by $3h^2/|f'''|$.

(2) Insert $1/h_{\rm opt} = (|f'''|\,/3\epsilon\,|f|)^{1/3}$ into the noise term.

(3) Insert $h_{\rm opt}^2 = (3\epsilon\,|f|\,/|f'''|)^{2/3}$ into the truncation term.

At the optimum both terms are of order $\epsilon^{2/3}$ (the noise term is exactly twice the truncation term): a code accurate to one part in 10^6 gives derivatives accurate to about one part in 10^4, not 10^6.

The four-point rule. Its noise term is again of order $\epsilon\,|f|\,/h$ (the weights $1, 8, 8, 1$ over $12h$ add in quadrature to $\sqrt{130}/12 \approx 0.95$ times $\epsilon\,|f|\,/h$), and its truncation term is $h^4\left|f^{(5)}\right|/30$ from line (3) of (10.28). So $E_4(h) \approx \epsilon\,|f|\,/h + h^4\left|f^{(5)}\right|/30$; setting $\mathrm{d}E_4/\mathrm{d}h = -\epsilon\,|f|\,/h^2 + \frac{4}{30}h^3\left|f^{(5)}\right| = 0$ gives $h^5 = \frac{30}{4}\,\epsilon\,|f|\,/\left|f^{(5)}\right|$, so $h_{\rm opt} \propto \epsilon^{1/5}$, and the error $\epsilon\,|f|\,/h_{\rm opt}$ is of order $\epsilon/\epsilon^{1/5} = \epsilon^{4/5}$: larger steps, and better derivatives.

Example: a function we can differentiate by hand. Take $f(x) = e^x \sin 3x$ at $x = 0.7$, whose derivative is known exactly, and give it a simulated "code noise" of relative size $\epsilon = 10^{-6}$. Figure 10.11 shows the error of the two rules against h, as the root-mean-square over 200 noise realisations. Without noise (dotted) the errors fall as h^2 and h^4 until double-precision round-off takes over near $h \sim 10^{-5}$ to 10^{-8}. With noise, the two-point error is smallest at $h = 6.1\times10^{-3}$, where (10.29) predicts 5.8×10^{-3} and an error of 3×10^{-4} (measured: 2×10^{-4}). The four-point rule is best at the larger step 3.3×10^{-2}, with error 4×10^{-5}. Below the optimum the error climbs as $1/h$ along the dashed line: there, making the step smaller makes the derivative worse.

10.5.2 The stability test for CAMB

For CAMB we do not know ϵ or f''', so we cannot compute $h_{\rm opt}$; we test instead. The steps h_i we use are roughly a third to a whole of the expected error of each parameter: 0.0001 in ω_b,

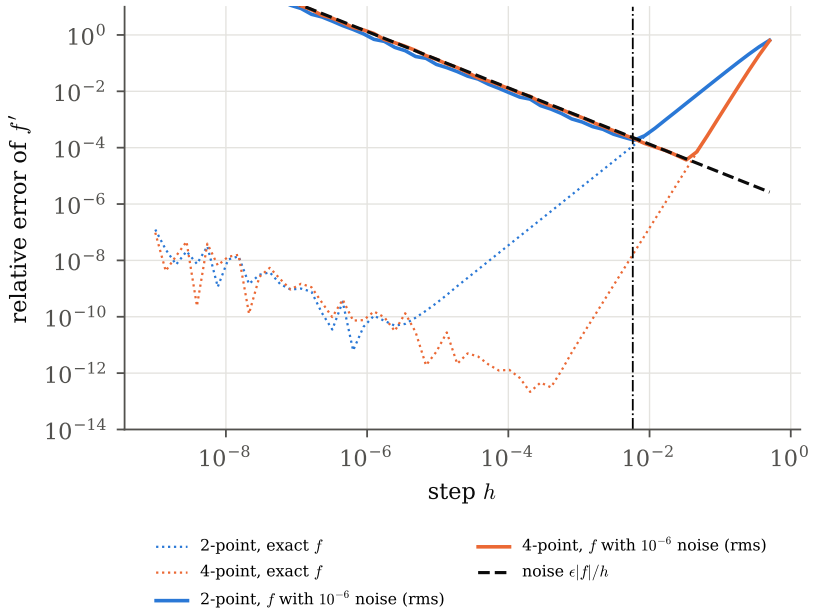


Figure 10.11: The error of a numerical derivative of $e^x \sin 3x$ against the step. Dotted: exact function values, where only truncation and double-precision round-off act. Solid: function values with a relative noise of 10^{-6}; the error is the sum of a noise term rising as $1/h$ (dashed) and a truncation term falling as h^2 (two-point) or h^4 (four-point). The vertical line is the predicted optimum of the two-point rule.

0.001 in ω_c, 0.0002 in $100\theta_{\mathrm{MC}}$, 0.005 in τ, 0.01 in $\ln(10^{10}A_s)$ and 0.004 in n_s. The reason for this scale is the truncation side of (10.29): the forecast is only meaningful over a region about one standard deviation wide, so the derivative should describe the model over that region, not over a much smaller or much larger one.

The test is to multiply every step by a factor m, recompute the derivatives with both rules, rebuild the forecast, and watch the marginalised errors. `04_derivatives.py` does this for $m = \frac{1}{64}, \frac{1}{16}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4$, which takes 84 CAMB runs, cached so that every later script reuses them:

Listing 10.2: Two- and four-point derivatives of the CMB spectra, from cached CAMB runs (`code/ch10/04_derivatives.py`)

```
def derivs(m, stencil):
    out = {s: np.zeros((6, L.LMAX + 1)) for s in ("TT", "EE", "TE")}
    for i in range(6):
        h = m * L.STEPS[i]
        p1, m1 = spectra_at(i, m), spectra_at(i, -m)
        if stencil == 4:
            p2, m2 = spectra_at(i, 2 * m), spectra_at(i, -2 * m)
        for s in out:
            if stencil == 2:
                out[s][i] = (p1[s] - m1[s]) / (2 * h)
            else:
                out[s][i] = (-p2[s] + 8 * p1[s] - 8 * m1[s] + m2[s]) / (12 * h)
```

The forecast is the Planck-like temperature-and-polarization experiment of Section 10.6.3. Figure 10.12 shows the result. For $\frac{1}{4} \le m \le 4$ every marginalised error stays within 0.07% of the four-point reference, with either rule: the forecast does not care about the step. Below that, it does. At $m = \frac{1}{16}$ the largest change is 0.41%; at $m = \frac{1}{64}$ it is 2.49% (two-point) and 3.00% (four-point), and the most affected parameter is ω_c.

Why the noise makes the forecast too good. The errors at tiny steps are *smaller*, not larger. The reason is the geometric reading of Section 10.3.2, and two parameters are enough to see it. *The Fisher matrix as dot products.* By (10.31), F_{ij} is a weighted sum over multipoles of $\partial_i C_\ell \, \partial_j C_\ell$. Call the two derivative vectors $\boldsymbol{u} = (\partial_1 C_\ell)_\ell$ and $\boldsymbol{v} = (\partial_2 C_\ell)_\ell$, and write $\boldsymbol{a} \cdot \boldsymbol{b}$ for the weighted sum $\sum_\ell \frac{2\ell+1}{2} a_\ell b_\ell / (C_\ell + N_\ell)^2$. Then $F_{11} = |\boldsymbol{u}|^2$, $F_{22} = |\boldsymbol{v}|^2$, $F_{12} = \boldsymbol{u} \cdot \boldsymbol{v}$, and by the 2×2 inverse of Section 10.4.1 the correlation of the two estimates has size $|\rho| = |\boldsymbol{u} \cdot \boldsymbol{v}| \,/ (|\boldsymbol{u}|\,|\boldsymbol{v}|)$, the cosine of the angle between the two vectors. A degeneracy is two vectors pointing in nearly the same direction, $|\rho|$ close to one.

Add noise. The computed vectors are $\boldsymbol{u}' = \boldsymbol{u} + \boldsymbol{n}_1$ and $\boldsymbol{v}' = \boldsymbol{v} + \boldsymbol{n}_2$, where $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$ fluctuate rapidly and randomly from one multipole to the next. A dot product of such a vector with a smooth one, or with another independent noise vector, is a sum of terms of random sign that nearly cancel; the dot product of a noise vector with itself is a sum of squares, which does not. So

$$F'_{11} \approx |\boldsymbol{u}|^2 + |\boldsymbol{n}_1|^2 \,, \qquad F'_{22} \approx |\boldsymbol{v}|^2 + |\boldsymbol{n}_2|^2 \,, \qquad F'_{12} \approx \boldsymbol{u} \cdot \boldsymbol{v} \,,$$

$$|\rho'| = \frac{|\boldsymbol{u} \cdot \boldsymbol{v}|}{\sqrt{(|\boldsymbol{u}|^2 + |\boldsymbol{n}_1|^2)(|\boldsymbol{v}|^2 + |\boldsymbol{n}_2|^2)}} < \frac{|\boldsymbol{u} \cdot \boldsymbol{v}|}{|\boldsymbol{u}|\,|\boldsymbol{v}|} = |\rho| \,.$$

The diagonal has gained $|\boldsymbol{n}|^2$ of spurious "information" and the off-diagonal has not. Both effects lower the marginalised error $\sigma_1^{\text{marg}} = 1/\sqrt{F_{11}(1-\rho^2)}$ of (10.24): F_{11} grows and $1-\rho^2$ grows. The second effect is the strong one near a degeneracy. With $|\rho| = 0.95$ and noise carrying 1% of the power of each vector, $|\rho'| = 0.95/1.01 = 0.941$, and $1/\sqrt{1-\rho^2}$ falls from 3.20 to 2.95; together with the factor $1/\sqrt{1.01}$ the marginalised error drops by 8%, eight times the 1% by which the noise changed the vectors. So the symptom of noisy derivatives is an optimistic forecast, and it strikes first the parameters involved in the strongest degeneracy; for us that is ω_c, whose correlation with n_s is the largest in the matrix (Section 10.6.4).

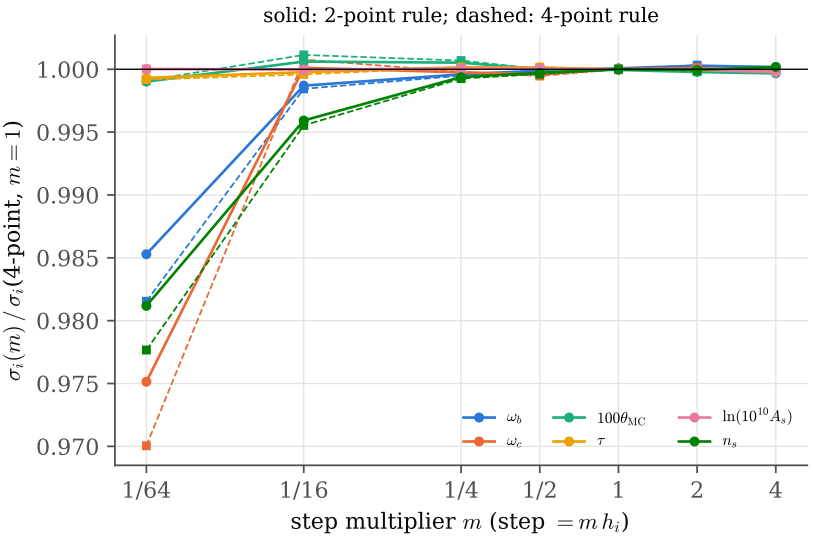


Figure 10.12: Marginalised errors of the six ΛCDM parameters for a Planck-like forecast, recomputed with every derivative step multiplied by m, relative to the four-point rule at $m = 1$. Between $m = \frac{1}{4}$ and 4 nothing moves by more than about two tenths of a per cent. At very small steps the numerical noise of CAMB enters the derivatives and the errors shrink by up to three per cent: noise looks like information.

Rules that follow. The two tests above turn into a short list of checks that every Fisher forecast should pass before its numbers are believed.

1. Choose steps of order a fraction of the expected error, not "as small as possible".
2. Halve and double the steps; quote the forecast only if it does not move.
3. Compare the two-point and four-point rules at the same step; agreement means truncation is negligible.
4. If the errors drift *downwards* as the steps shrink, the derivatives are noisy; raise the code's accuracy settings or increase the step.
5. Differentiate quantities that are nearly linear in the parameters, such as $\ln C_\ell$ in $\ln A_s$ and τ (Section 9.20.1); the truncation error is then small at any reasonable step.

10.6 A CMB forecast for ΛCDM

In 1995 the Planck satellite was a proposal. Knox [166], Jungman et al. [196] and Tegmark, Taylor and Heavens [192] used the machinery of the last five sections to tell the community what such a satellite would measure, and those forecasts helped decide its design. Twenty-three years later Planck published its final parameters [25]. We can now do what is rarely possible with a forecast: redo it, and compare it with what the experiment actually delivered.

The question. The CMB temperature and polarization are Gaussian random fields on the sphere whose statistics are set by the angular power spectra C_ℓ^{TT}, C_ℓ^{EE}, C_ℓ^{TE}, which a Boltzmann code computes from the six parameters of ΛCDM (Chapter T1; the physics is in the cosmology notes). An experiment with a given beam, noise and sky fraction will measure the $a_{\ell m}$. How well will it measure the six parameters?

10.6.1 From the $a_{\ell m}$ to the CMB Fisher matrix

The data. Expand the measured temperature map in spherical harmonics (Chapter 7). The coefficients $a_{\ell m}$ for $2 \le \ell \le \ell_{\max}$, $-\ell \le m \le \ell$, are the data vector; written as real numbers there are $2\ell+1$ of them at each ℓ. *Their mean is zero*: the parameters do not move the map, they set its variance. *Their covariance is diagonal*: statistical isotropy makes different (ℓ, m) uncorrelated, and the variance of each is the sky signal plus the instrument noise,

$$\langle a_{\ell m} a_{\ell' m'} \rangle = \big(C_\ell + N_\ell\big)\delta_{\ell\ell'}\delta_{mm'}, \qquad N_\ell = \Delta^2 e^{\ell(\ell+1)\sigma_b^2}, \quad \sigma_b = \frac{\mathrm{FWHM}}{\sqrt{8\ln 2}}. \tag{10.30}$$

Here N_ℓ is the white noise of depth Δ (in μK rad) after the Gaussian beam has been divided out of the map: the beam suppresses the signal at $\ell \gtrsim 1/\sigma_b$ but not the noise, so dividing by it makes the noise grow exponentially (Chapter 8; the cosmology notes write it as N_ℓ/W_ℓ with W_ℓ the beam window). Tegmark et al. write the same quantity as $4\pi\sigma^2/N\, e^{\theta_b^2 \ell(\ell+1)}$ for N pixels of noise σ, since $\Delta^2 = \sigma^2 \Omega_{\mathrm{pix}} = 4\pi\sigma^2/N$ [192], eq. (16).

Apply the Gaussian formula. With $\boldsymbol{\mu} = \mathbf{0}$ only the covariance channel of (10.20) survives, and $\mathbf{C}$ is diagonal, so every matrix product is a product of diagonal entries:

$$F_{ij} \overset{(1)}{=} \frac{1}{2}\,\mathrm{Tr}\big[\mathbf{C}^{-1}\mathbf{C}_{,i}\mathbf{C}^{-1}\mathbf{C}_{,j}\big] \overset{(2)}{=} \frac{1}{2}\sum_{\ell=2}^{\ell_{\max}}\sum_{m=-\ell}^{\ell} \frac{\partial_i C_\ell}{C_\ell + N_\ell}\,\frac{\partial_j C_\ell}{C_\ell + N_\ell} \overset{(3)}{=} \sum_{\ell=2}^{\ell_{\max}} \frac{2\ell+1}{2}\,\frac{\partial_i C_\ell\, \partial_j C_\ell}{(C_\ell + N_\ell)^2}. \tag{10.31}$$

(1) (10.20) with $\boldsymbol{\mu}_{,i} = \mathbf{0}$.

(2) $\mathbf{C}^{-1}\mathbf{C}_{,i}$ is diagonal with entry $\partial_i C_\ell/(C_\ell + N_\ell)$ at every (ℓ, m); the noise does not depend on the cosmological parameters.
(3) Nothing depends on m, so the sum over m gives $2\ell + 1$ equal terms.

This is Tegmark, Taylor and Heavens' eq. (18) [192], the result Section 3.16.7 found for one amplitude and Section 9.20.3 found by expanding the exact likelihood. It has a plain reading: since $\mathrm{Var}\,\widehat{C}_\ell = 2(C_\ell + N_\ell)^2/(2\ell + 1)$ (cosmic variance plus noise), each multipole contributes $(\partial_i C_\ell)(\partial_j C_\ell)/\mathrm{Var}\,\widehat{C}_\ell$, the squared response of the spectrum in units of its own error bar. Multipoles where the spectrum is precisely measured and responds strongly to θ_i dominate.

Part of the sky. Divide the full sky into many small patches of equal area, each carrying its own independent share of the information on the parameters. A survey that sees a fraction $f_{\rm sky}$ of the sky sees the fraction $f_{\rm sky}$ of those patches. Independent pieces of data add their Fisher matrices (Section 10.1.2), so the observed part carries $f_{\rm sky}$ times the Fisher matrix of the full sky: every term of (10.31) is multiplied by $f_{\rm sky}$, or, in the language of modes, each multipole has about $(2\ell+1)f_{\rm sky}$ independent modes instead of $2\ell+1$. The assumption is that the patches really are independent, that is, that the edge of the mask does not couple neighbouring multipoles. A mask does couple them (Chapter 8), which matters at low ℓ, where a mode is not small compared with the patch, and for sharp masks.

Temperature and polarization. At each (ℓ, m) the data are now the pair $(a^T_{\ell m}, a^E_{\ell m})$, correlated with each other but not with other (ℓ, m), with the 2×2 covariance $\mathbf{C}_\ell = \begin{pmatrix} C_\ell^{TT}+N_\ell^T & C_\ell^{TE} \\ C_\ell^{TE} & C_\ell^{EE}+N_\ell^P \end{pmatrix}$. The same three steps give

$$F_{ij} = \sum_\ell \frac{(2\ell+1)f_{\rm sky}}{2}\,\mathrm{Tr}\big[\mathbf{C}_\ell^{-1}\partial_i\mathbf{C}_\ell\,\mathbf{C}_\ell^{-1}\partial_j\mathbf{C}_\ell\big], \tag{10.32}$$

now with a 2×2 trace at each ℓ. Verde [197], eqs. (26)–(28), writes the same thing as $\sum_\ell \partial_i \boldsymbol{C}_\ell^{\mathsf{T}}\,\boldsymbol{\Sigma}_\ell^{-1}\,\partial_j \boldsymbol{C}_\ell$, with $\boldsymbol{C}_\ell = (C_\ell^{TT}, C_\ell^{EE}, C_\ell^{TE})$ and $\boldsymbol{\Sigma}_\ell$ the 3×3 covariance of the three estimated spectra. Our code implements both, and they agree to 4×10^{-15} in relative terms, as they must, since they are the same Gaussian likelihood seen through two different data vectors (the $a_{\ell m}$ and the spectra built from them):

Listing 10.3: The temperature-and-polarization Fisher matrix, one 2×2 trace per multipole (`code/ch10/lib_fisher10.py`)

```
163     a = C["TT"][l] + NT[l]; b = C["TE"][l]; e = C["EE"][l] + NP[l]
164     det = a * e - b * b
165     # inverse of [[a, b], [b, e]] is [[e, -b], [-b, a]] / det
166     ia, ib, ie = e / det, -b / det, a / det
167     F = np.zeros((6, 6))
168     for i in range(6):
169         Ma = ia * dC["TT"][i, l] + ib * dC["TE"][i, l]      # (Cinv dC_i), elements 11, 12,
        21, 22
170         Mb = ia * dC["TE"][i, l] + ib * dC["EE"][i, l]
171         Mc = ib * dC["TT"][i, l] + ie * dC["TE"][i, l]
172         Md = ib * dC["TE"][i, l] + ie * dC["EE"][i, l]
173         for j in range(i + 1):
174             Na = ia * dC["TT"][j, l] + ib * dC["TE"][j, l]
175             Nb = ia * dC["TE"][j, l] + ib * dC["EE"][j, l]
176             Nc = ib * dC["TT"][j, l] + ie * dC["TE"][j, l]
177             Nd = ib * dC["TE"][j, l] + ie * dC["EE"][j, l]
178             tr = Ma * Na + Mb * Nc + Mc * Nb + Md * Nd
179             F[i, j] = F[j, i] = np.sum(nu / 2 * tr)
180     return F
```

Lines 169–172 form the four entries of $\mathbf{C}_\ell^{-1}\partial_i\mathbf{C}_\ell$ for all ℓ at once, and line 178 is the trace of the product of two such 2×2 matrices.

Several frequency channels. Planck observed the CMB in several channels; each channel ν gives an independent noisy measurement of the same $a_{\ell m}$, with noise $N_{\ell,\nu}$. For the single "parameter" $a_{\ell m}$ these are independent experiments, so their Fisher informations $1/N_{\ell,\nu}$ add (Section 10.4.4): the combined noise is $1/N_\ell = \sum_\nu 1/N_{\ell,\nu}$, the inverse-variance combination.

10.6.2 Knox (1995) reproduced

His question. Inflation, the brief early epoch of accelerated expansion, predicts two kinds of primordial ripple: density fluctuations, which seed galaxies, and gravitational waves. In 1995 the anisotropy of the CMB had just been detected on large scales, and a satellite to map it on smaller scales was being argued for. Knox asked whether such a satellite could separate the three numbers that tell inflation models apart: the *amplitude* of the density fluctuations, their *tilt* (how their strength changes with scale), and the *share of gravitational waves* in the observed anisotropy [166]. If the three could not be separated, the satellite could not test inflation, however good its maps.

His method. Knox did not compute a Fisher matrix. He drew the estimated spectrum $\widehat{C}_\ell$ from its $\chi^2_{2\ell+1}$ distribution, found the maximum of the exact likelihood by simulated annealing, repeated this 100 times per experiment, and quoted the standard deviation of the maxima [Knox1995 pp. 11–14]. His Table 1 is therefore a Monte Carlo result, and a Fisher forecast should reproduce it.

His setup

The background cosmology is standard cold dark matter: a flat universe made of matter only, with no cosmological constant and no reionisation. It is held fixed; only the three primordial numbers are fitted.

The density (scalar) part. Its strength is quoted through the quadrupole it produces. Knox's Q is the root-mean-square quadrupole, defined by $Q^2 = 5C_2/4\pi$, so the scalar part alone has $C_2^S = 4\pi Q_S^2/5$. Its shape across multipoles, $s_\ell = C_\ell^S/C_2^S$, is fixed by the background cosmology and by the tilt n_S; we compute it with CAMB.

The gravitational-wave (tensor) part adds its own spectrum, with quadrupole $C_2^T = 4\pi Q_T^2/5$ and shape $t_\ell = C_\ell^T/C_2^T$. Gravitational waves decay once they come inside the horizon, so t_ℓ falls steeply above $\ell \sim 100$ and the tensor part only shows on large scales. Its strength is measured against the scalar part by the ratio of the two quadrupoles, $r = Q_T^2/Q_S^2 = C_2^T/C_2^S$. The tensor tilt n_T is held at its fiducial value.

quantity	meaning	value
h	Hubble constant/(100 km s^{-1} Mpc^{-1})	0.5
Ω	total density/critical; no Λ	1
$\Omega_b h^2$	physical baryon density	0.0125
n_S	scalar tilt (fitted)	0.94
n_T	tensor tilt (fixed)	−0.04
r	tensor/scalar quadrupole ratio (fitted)	0.28
Q_S	scalar rms quadrupole (fitted)	20 μK

These are Knox's fiducial values, from chaotic inflation [Knox1995 pp. 3–5].

His key equations, derived

The model spectrum. The two parts are independent, so their spectra add, and each is its quadrupole times its shape:

$$C_\ell \overset{(1)}{=} C_\ell^S + C_\ell^T \overset{(2)}{=} C_2^S\, s_\ell + C_2^T\, t_\ell \overset{(3)}{=} C_2^S\big[s_\ell(n_S) + r\, t_\ell\big]$$
$$\overset{(4)}{=} \frac{4\pi}{5} Q_S^2\big[s_\ell(n_S) + r\, t_\ell\big], \qquad s_\ell = \frac{C_\ell^S}{C_2^S}, \quad t_\ell = \frac{C_\ell^T}{C_2^T}. \tag{10.33}$$

(1) Uncorrelated contributions to the $a_{\ell m}$ add their variances.
(2) The definitions of the two shapes.
(3) $C_2^T = r\, C_2^S$, the definition of r.
(4) $C_2^S = 4\pi Q_S^2/5$, from $Q^2 = 5C_2/4\pi$.

The derivatives follow at once. C_ℓ is proportional to Q_S^2, so $\partial_{Q_S} C_\ell = 2C_\ell/Q_S$; it is linear in r, so $\partial_r C_\ell = \frac{4\pi}{5} Q_S^2 t_\ell$; and n_S enters only through the shape, so $\partial_{n_S} C_\ell = \frac{4\pi}{5} Q_S^2\, \partial_{n_S} s_\ell$, which we take from a cubic spline in n_S of $\ln s_\ell$ computed with CAMB on a grid of tilts. As a check of the shapes, our t_{55}/s_{55} is 0.36, against Knox's "$C_{55}^T/C_{55}^S \simeq r/3$".

His experiments. Full sky, Gaussian beams of FWHM 20′, 30′ and 40′, and noise described by a weight per solid angle w, with $w^{-1/2} = 15$ or 30 μK deg. His equation (4) gives the variance of the estimated spectrum,

$$\operatorname{Var} \widehat{C}_\ell = \frac{2}{2\ell+1}\big(C_\ell + w^{-1} e^{\ell^2 \sigma_b^2}\big)^2.$$

We derive it. At each ℓ the estimate is the average power of the $2\ell+1$ real coefficients, $\widehat{C}_\ell = \frac{1}{2\ell+1}\sum_m |a_{\ell m}|^2$, and by (10.30) each coefficient is an independent Gaussian of mean zero and variance $C_\ell + N_\ell$. Then

$$\widehat{C}_\ell \overset{(1)}{=} \frac{C_\ell + N_\ell}{2\ell+1} \sum_m \left(\frac{a_{\ell m}}{\sqrt{C_\ell + N_\ell}}\right)^2 \overset{(2)}{=} \frac{C_\ell + N_\ell}{2\ell+1}\, \chi^2_{2\ell+1},$$
$$\operatorname{Var} \widehat{C}_\ell \overset{(3)}{=} \frac{(C_\ell + N_\ell)^2}{(2\ell+1)^2}\, 2(2\ell+1) = \frac{2(C_\ell + N_\ell)^2}{2\ell+1}.$$

(1) Divide and multiply each term by the variance.
(2) The bracket is a standard normal, and the sum of $2\ell+1$ independent squared standard normals is $\chi^2_{2\ell+1}$ (Section 3.16.7).
(3) A χ^2_k variable has variance $2k$; a constant factor comes out squared.

This is his (4) with $N_\ell = w^{-1} e^{\ell^2 \sigma_b^2}$. Comparing with (10.30), $N_\ell = \Delta^2 e^{\ell(\ell+1)\sigma_b^2}$, his w^{-1} plays the part of our Δ^2, and he writes ℓ^2 for $\ell(\ell+1)$, which differ by less than 1% where the beam matters. The two noise levels are the same physical quantity, the pixel noise variance times the pixel solid angle, in different units. Knox quotes $w^{-1/2}$ in μK deg; we need Δ in μK rad. One degree is $\pi/180$ radians, so $w^{-1/2} = 15\,\mu$K deg is $\Delta = 15 \times \pi/180 = 0.262\,\mu$K rad and $\Delta^2 = 0.0685\,\mu\text{K}^2$ sr; with that conversion $\Delta^2 = w^{-1}$. His Fisher matrix is therefore (10.31) with $f_{\text{sky}} = 1$.

Symbol map. Knox's symbols and ours:

Knox	ours	meaning
$C_\ell, C_\ell^S, C_\ell^T$	$C_\ell, C_\ell^S, C_\ell^T$	total, scalar, tensor spectra
Q_S, Q_T	$(5C_2^S/4\pi)^{1/2}, (5C_2^T/4\pi)^{1/2}$	rms quadrupoles
w^{-1}	Δ^2	white-noise level, μK^2 sr
σ_b	$\sigma_b = \text{FWHM}/\sqrt{8\ln 2}$	beam width
$\ell_*, \bar{Q}_S$	$\ell_*, \bar{Q}_S$	pivot multipole, amplitude at the pivot

The pivot

The question. Knox's amplitude and tilt came out almost perfectly correlated, and he explained why: the spectrum is measured best around some multipole ℓ_*, and a change of tilt can be compensated by a change of amplitude that keeps the power at ℓ_* fixed. So he quoted the amplitude at the pivot, $\bar{Q}_S = Q_S(\ell_*/2)^{(n_S-0.94)/2}$, chosen so that it is uncorrelated with n_S [Knox1995 pp. 12–13]. Where does that formula come from, and which ℓ_* is it?

Why the amplitude at a pivot absorbs the tilt. A tilt n_S different from the fiducial 0.94 multiplies the primordial power at wavenumber k by $(k/k_2)^{n_S-0.94}$, where k_2 is the wavenumber the quadrupole sees and where the shape is normalised. Multipole ℓ sees mostly $k \approx \ell/D_*$ (Section 8.7.3, Step 2), so $k/k_2 \approx \ell/2$ and, to that approximation, $s_\ell(n_S) = s_\ell(0.94)\,(\ell/2)^{n_S-0.94}$. The scalar spectrum is then

$$\begin{aligned} C_\ell^S &\overset{(1)}{=} \frac{4\pi}{5} Q_S^2 \Big(\frac{\ell}{2}\Big)^{n_S-0.94} s_\ell(0.94) \\ &\overset{(2)}{=} \frac{4\pi}{5}\Big[Q_S^2\Big(\frac{\ell_*}{2}\Big)^{n_S-0.94}\Big]\Big(\frac{\ell}{\ell_*}\Big)^{n_S-0.94} s_\ell(0.94) \\ &\overset{(3)}{=} \frac{4\pi}{5}\,\bar{Q}_S^2\Big(\frac{\ell}{\ell_*}\Big)^{n_S-0.94} s_\ell(0.94). \end{aligned}$$

(1) The scalar part of (10.33) with the tilted shape.
(2) Split $(\ell/2) = (\ell_*/2)(\ell/\ell_*)$ and group the factor that does not depend on ℓ with Q_S^2.
(3) Name the bracket $\bar{Q}_S^2$; its square root is Knox's $\bar{Q}_S = Q_S(\ell_*/2)^{(n_S-0.94)/2}$.

Written this way, a change of tilt rotates the spectrum about $\ell = \ell_*$, where $(\ell/\ell_*)^{n_S-0.94} = 1$, and leaves the power there equal to $\frac{4\pi}{5}\bar{Q}_S^2 s_{\ell_*}$. So $\bar{Q}_S$ is the amplitude referred to ℓ_* instead of to the quadrupole, the move of the reference point that Section 8.7.3 made for A_s with (8.45).

Which ℓ_*. Taking logarithms, $\ln\bar{Q}_S = \ln Q_S + \frac{1}{2}(n_S - 0.94)\ln(\ell_*/2)$, and requiring zero covariance with n_S,

$$\begin{aligned} 0 = \operatorname{Cov}(\ln\bar{Q}_S, n_S) &\overset{(1)}{=} \operatorname{Cov}(\ln Q_S, n_S) + \tfrac{1}{2}\ln(\ell_*/2)\operatorname{Var}(n_S) \;\Rightarrow\; \ln\frac{\ell_*}{2} = -\frac{2\operatorname{Cov}(\ln Q_S, n_S)}{\operatorname{Var}(n_S)}, \\ \operatorname{Var}(\ln\bar{Q}_S) &\overset{(2)}{=} \operatorname{Var}(\ln Q_S) - \frac{\operatorname{Cov}(\ln Q_S, n_S)^2}{\operatorname{Var}(n_S)}, \end{aligned} \tag{10.34}$$

(1) Covariance is linear in each argument; the term -0.94 is a constant.
(2) Expand $\operatorname{Var}(\ln Q_S + c\,n_S)$ with $c = \frac{1}{2}\ln(\ell_*/2)$ and insert the value of c: this is the Schur complement (10.23) once more, the variance of $\ln Q_S$ with the part explained by n_S removed.

with all variances and covariances taken from $\mathbf{F}^{-1}$. This is (8.46) in Knox's variables.

The pivot is an information-weighted mean. Leave aside the tensor part, which is small where most of the information lies, and use the tilted shape above. Then $\partial C_\ell/\partial\ln Q_S = 2C_\ell$ and $\partial C_\ell/\partial n_S = C_\ell\ln(\ell/2)$. Each multipole enters the Fisher sum (10.31) with the factor $\frac{2\ell+1}{2}(C_\ell+N_\ell)^{-2}$; call $\mathcal{I}_\ell = \frac{2\ell+1}{2}\big[C_\ell/(C_\ell+N_\ell)\big]^2$ the information of multipole ℓ and $g_\ell = \ln(\ell/2)$. Then $F_{QQ} = 4\sum\mathcal{I}_\ell$, $F_{Qn} = 2\sum\mathcal{I}_\ell g_\ell$, $F_{nn} = \sum\mathcal{I}_\ell g_\ell^2$, and

$$\ln\frac{\ell_*}{2} \overset{(1)}{=} -\frac{2\operatorname{Cov}(\ln Q_S, n_S)}{\operatorname{Var}(n_S)} \overset{(2)}{=} \frac{2F_{Qn}}{F_{QQ}} \overset{(3)}{=} \frac{\sum_\ell \mathcal{I}_\ell \ln(\ell/2)}{\sum_\ell \mathcal{I}_\ell},$$

(1) (10.34).
(2) The 2×2 inverse (10.7): $\operatorname{Cov} = -F_{Qn}/\det\mathbf{F}$, $\operatorname{Var}(n_S) = F_{QQ}/\det\mathbf{F}$; the determinant cancels.

(3) Insert the two sums; the factors $2 \cdot 2/4$ cancel.

Adding $\ln 2$ to both sides, $\ln \ell_* = \sum_\ell \mathcal{I}_\ell \ln \ell \,/\, \sum_\ell \mathcal{I}_\ell$: $\ln \ell_*$ is the mean of $\ln \ell$ over the multipoles, each weighted by the information it carries, as Section 8.7.3 found for Planck's k_*. The straight-line pivot of Section 10.4.3 is the same statement with x in place of $\ln \ell$. Our numbers below keep the tensor part and the exact shapes, so the pivots in Table 10.2 come from (10.34) with the full $\mathbf{F}^{-1}$.

What we compute, side by side, and why the numbers differ

What we compute. `05_knox1995.py` builds s_ℓ and t_ℓ with CAMB for Knox's cosmology, forms the three derivatives above, sums (10.31) with $f_{\rm sky} = 1$ and Knox's noise for each of his six experiments, inverts, and applies (10.34). To test the Fisher matrix against his own procedure, it also repeats that procedure for the first experiment with 2000 simulations instead of 100: draw $\widehat{C}_\ell$, maximise the exact likelihood, take the scatter of the maxima.

Side by side. Table 10.2 and Figure 10.13 put the numbers together. The pivot of the first experiment is at $\ell_* = 210$; Knox reports $\ell_* = 210$ for it [Knox1995 p. 13]. Our 2000 simulations give $\sigma(n_S) = 0.014$, $\sigma(r) = 0.09$, $\sigma(\bar{Q}_S)/\bar{Q}_S = 0.0028$, with pivot 212, all equal to the Fisher values within the ±1.6% simulation error, and no bias (mean $\hat{n}_S = 0.9397$, $\hat{r} = 0.283$). So for this model the Fisher matrix is an accurate description of maximum likelihood.

Why they differ: the tilt and the amplitude. Knox's numbers are standard deviations of $N = 100$ values, and a standard deviation estimated from N Gaussian values has a relative error of $1/\sqrt{2(N-1)} = 1/\sqrt{198} = 7\%$. Of the twelve tilt and amplitude entries, nine agree with ours within that 7% and the other three (the tilt of the first experiment, 0.016 against 0.014, and two amplitudes) within 14%, two sampling errors: about what twelve comparisons with 7% noise should produce. Nothing beyond sampling noise is needed to explain them.

Why they differ: the tensor ratio. Here our Fisher errors equal Knox's own analytic estimates (0.10, 0.14, 0.12, 0.19, 0.16, 0.22) but sit below his simulations by factors of 1.22 to 1.29 for five experiments and 1.83 for the last, far more than the 7% sampling error. One candidate is a constraint $r \geq 0$ on the fits. Its effect can be computed. If the fits that want $r < 0$ are set to $r = 0$, the scatter of $\hat{r}$ is that of a Gaussian cut off at zero. For a Gaussian of mean μ and unit width, the cut-off variable $Y = \max(X, 0)$ has $\mathbb{E}\,Y = \mu\Phi(\mu) + \varphi(\mu)$ and $\mathbb{E}\,Y^2 = (\mu^2+1)\Phi(\mu) + \mu\varphi(\mu)$, with Φ and φ the standard normal distribution and density (integrate y and y^2 against the Gaussian over $y > 0$). For $r/\sigma_r = \mu = 1$ this gives a standard deviation $0.87\,\sigma_r$, for $\mu = 2$ it gives $0.98\,\sigma_r$; discarding those fits instead gives 0.79 and 0.94. A constraint $r \geq 0$ therefore makes the scatter *smaller*, and cannot by itself produce an excess of 20 to 80%. The cause is not stated in Knox's paper, and we have not found it. A maximiser that sometimes stops short of the maximum would add scatter of the right sign, but that is a guess we have not tested.

Lesson. Where the likelihood is close to Gaussian, as for the tilt and the pivot amplitude, a few lines of Fisher algebra reproduce a hundred simulated annealings per experiment; where a forecast and a simulation disagree, rerunning the simulation more carefully, as our 2000 fits did, decides which one to trust.

10.6.3 A Planck-like forecast against Planck 2018

Their question. The final Planck release set out to give the definitive values of the six parameters of ΛCDM from the full mission's temperature and polarization maps [25]. Several of those numbers carried questions of their own. Whether the tilt n_s is below one, and by how much, tests inflation, which predicts a slightly red spectrum. The optical depth τ dates reionisation, and through the degeneracy described in Section 10.6.4 it limits how well the primordial

		Knox 1995, 100 simulations			Fisher (ours)			
FWHM	$w^{-1/2}$	$\sigma(\bar{Q}_S)/\bar{Q}_S$	$\sigma(n_S)$	$\sigma(r)$	$\sigma(\bar{Q}_S)/\bar{Q}_S$	$\sigma(n_S)$	$\sigma(r)$	ℓ_*
20′	15	0.0027	0.016	0.11	0.0026	0.014	0.09	210
20′	30	0.0041	0.029	0.18	0.0041	0.029	0.14	175
30′	15	0.0028	0.022	0.14	0.0031	0.023	0.11	181
30′	30	0.0047	0.044	0.23	0.0047	0.042	0.18	161
40′	15	0.0035	0.034	0.19	0.0036	0.034	0.15	165
40′	30	0.0050	0.058	0.42	0.0056	0.058	0.23	152
ours, 2000 sims, 20′/15					0.0028	0.014	0.09	212

Table 10.2: Knox's Table 1 ("numerical" columns, $w^{-1/2}$ in μK deg) against the Fisher forecast for the same model and experiments, and against our own repetition of his simulation for the first experiment. ℓ_* is the pivot multipole of (10.34).

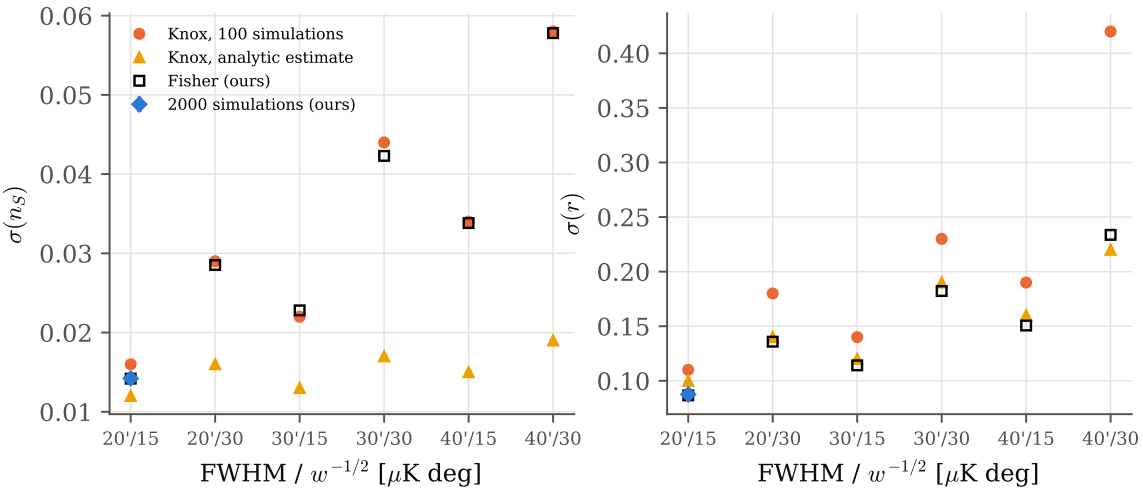


Figure 10.13: Knox's forecasts redone. Circles: Knox's 100-simulation errors; triangles: his analytic estimates; open squares: our Fisher matrix; diamond: our 2000 simulated experiments for the first configuration. The Fisher matrix reproduces his simulations for the tilt (left) and his analytic estimates for the tensor ratio (right).

amplitude is known. And the Hubble constant derived from the CMB, 67.36 ± 0.54 once the CMB lensing spectrum is added (the 0.60 of Table 10.3 is without it), is set against the 73.04 ± 1.04 of the supernova distance ladder: the Hubble tension of Section T2.12.3. Each question needs an error bar that can be trusted, which is what makes the comparison with a forecast interesting.

Their parameters, one at a time. Planck samples in a basis chosen to be close to what the spectrum measures directly, $\boldsymbol{\theta} = (\omega_b, \omega_c, 100\theta_{\mathrm{MC}}, \tau, \ln(10^{10}A_s), n_s)$. The *physical baryon density* $\omega_b = \Omega_b h^2$ sets how strongly the photon–baryon fluid is loaded, which shows as the alternating heights of odd and even acoustic peaks. The *physical cold-dark-matter density* $\omega_c = \Omega_c h^2$ sets the epoch of matter–radiation equality and so the overall heights of the first peaks. The *angle* $100\theta_{\mathrm{MC}}$ is an approximation to the angular size of the sound horizon at last scattering ($\theta_* = r_*/D_M$ in the cosmology notes); it fixes the positions of the peaks. The *optical depth* τ to reionisation suppresses all but the largest scales by $e^{-2\tau}$. The *amplitude* $\ln(10^{10}A_s)$ and the *tilt* n_s describe the primordial spectrum at the pivot $k = 0.05\,\mathrm{Mpc}^{-1}$ (Section 8.7.3): its overall height, and how its height changes from large scales to the damping tail. The fiducial model is the

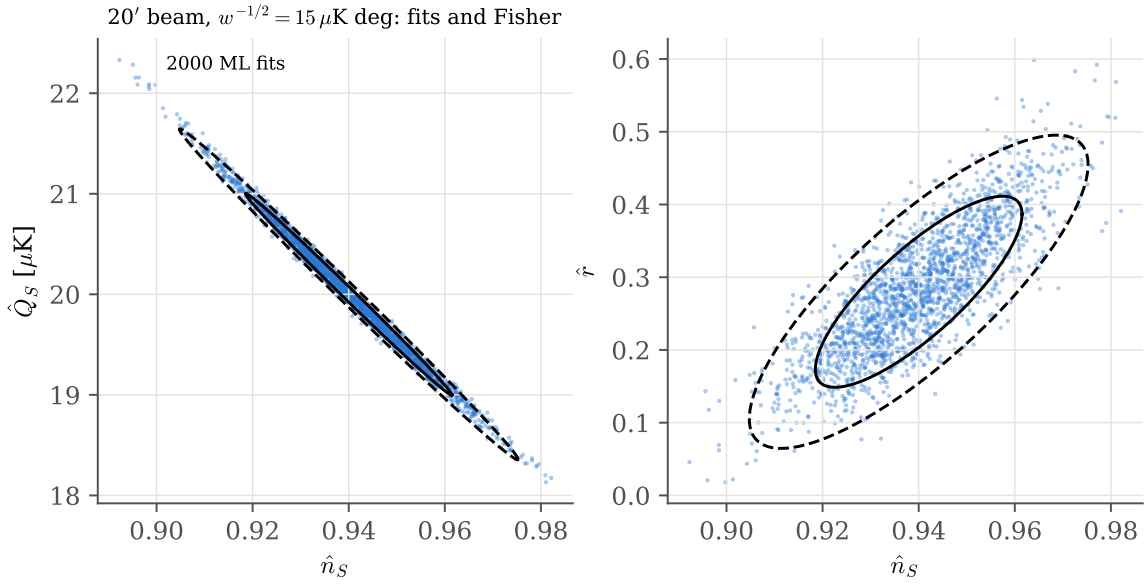


Figure 10.14: Our 2000 maximum-likelihood fits for Knox's 20′, 15 μK deg experiment, with the Fisher ellipses ($\Delta\chi^2 = 2.30$ solid, 6.18 dashed). Left: amplitude and tilt are almost perfectly anticorrelated, the lever-arm degeneracy that the pivot removes. Right: the tilt and the tensor ratio.

Planck 2018 best fit used throughout the book:

ω_b	ω_c	$100\theta_{\rm MC}$	τ	$\ln(10^{10}A_s)$	n_s	H_0 (derived)
0.02237	0.1200	≈ 1.041	0.0544	3.044	0.9649	67.36

H_0 is not in the basis; it follows from the other parameters.

Where we have met H_0 before. In Section 3.17 H_0 was fitted to mock supernovae made with $H_0 = 70$, its error was the Cramér–Rao bound from a Fisher matrix, and that matrix turned singular once the candle's absolute magnitude M was left free. The inverse ladder of Section T2.15 took the sound horizon $r_d = 147.09$ Mpc from the CMB and turned DESI's hr_d into $H_0 = 69.2 \pm 0.9$. The chain of Section 9.22 sampled H_0 directly from one simulated temperature spectrum and found 67.06 ± 0.93. What is the same here is the Fisher matrix as the error of an efficient estimator, and the experiment class of that chain: the forecast below for the chain's own one-channel experiment gives 0.93 again. What is different is that no data are fitted at all. The centre is the fiducial 67.36 by construction, and only the width is predicted; H_0 is not a parameter but a function of ω_b, ω_c and $100\theta_{\rm MC}$, its error carried by the delta method of Section 10.4.5; and the scale is set by the sound horizon at last scattering, a cousin of the r_d of the inverse ladder, not by a candle calibrated nearby. The forecast for Planck's three channels is tighter than that chain, 0.82 against 0.93, because three channels carry less noise than one, and it is tighter than Planck's own 0.92 because it leaves out the foregrounds and calibrations whose cost is traced below.

Their instrument, in our model. Planck's cosmological information at high ℓ comes from three frequency channels. We describe each by its beam and its white-noise level [186], Table 4, and combine them by inverse variance as in Section 10.6.1:

channel	beam FWHM	temperature noise	polarization noise
100 GHz	9.66′	1.29 μK deg	1.96 μK deg
143 GHz	7.22′	0.55 μK deg	1.17 μK deg
217 GHz	4.90′	0.78 μK deg	1.75 μK deg

Two forecasts are made. "PTT" uses the temperature spectrum at $2 \le \ell \le 2500$ on $f_{\rm sky} = 0.57$. "PTE" adds E-mode polarization at $30 \le \ell \le 2000$ through (10.32). Both have a Gaussian prior $\sigma_\tau = 0.0086$ standing in for Planck's low-ℓ polarization likelihood (lowE), which on its own measures τ to that precision [25]. The derivatives are the four-point ones of Section 10.5.

What Planck fitted, and how it relates to our formula. Above $\ell = 30$ Planck's likelihood is a Gaussian in the binned cross-spectra $\hat{\boldsymbol{C}}$ of the frequency maps, with a covariance $\boldsymbol{\Sigma}$ computed once for a fiducial model (Section 9.19.4):

$$-2\ln\mathcal{L} = \boldsymbol{\Delta}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta} + \text{const}, \qquad \boldsymbol{\Delta} = \hat{\boldsymbol{C}} - \boldsymbol{C}(\boldsymbol{\theta}, \boldsymbol{\nu}).$$

The model $\boldsymbol{C}(\boldsymbol{\theta}, \boldsymbol{\nu})$ is the CMB spectrum plus foregrounds and calibrations, controlled by about twenty nuisance parameters $\boldsymbol{\nu}$; below $\ell = 30$ separate likelihoods handle temperature and lowE. This is a Gaussian with fixed covariance, so only the mean channel of (10.20) survives, and its Fisher matrix is $F_{ij} = \partial_i \boldsymbol{C}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\partial_j \boldsymbol{C}$. Leave out the nuisances, bins and cross-spectra, take $\boldsymbol{\Sigma}$ to be the Gaussian covariance of the estimated spectra, and this is Verde's form, which agrees with (10.32). So our forecast is Planck's likelihood without its nuisance parameters. Putting them back adds rows and columns to $\mathbf{F}$, and marginalising over them can only enlarge the errors (Section 10.4.1); we should expect our numbers to be optimistic, and the question is by how much.

	TT + lowE			TT,TE,EE + lowE			S4-like
	Planck 2018	ours	ratio	Planck 2018	ours	ratio	ours
ω_b	0.00022	0.00017	0.77	0.00015	0.00013	0.84	3.4×10^{-5}
ω_c	0.0021	0.0019	0.92	0.0014	0.0012	0.87	0.00071
$100\theta_{\rm MC}$	0.00047	0.00042	0.89	0.00031	0.00028	0.90	9.0×10^{-5}
τ	0.0080	0.0082	1.03	0.0076	0.0080	1.05	0.0056
$\ln(10^{10}A_s)$	0.016	0.016	1.01	0.016	0.016	0.98	0.0096
n_s	0.0057	0.0047	0.82	0.0044	0.0033	0.76	0.0022
H_0	0.92	0.82		0.60	0.53		0.27
Ω_m	0.013	0.0117		0.0084	0.0075		0.0040
$\Omega_m h^3$	0.00046	0.00040		0.00029	0.00026		0.00012
$10^9 A_s e^{-2\tau}$	0.014	0.0095		0.012	0.0064		0.0045

Table 10.3: Marginalised 1σ errors: Planck 2018 [25], Table 2, against our Fisher forecast for the same data combination ("ratio" is ours over Planck's). The last four rows are derived parameters, computed with the delta method $\nabla g^{\mathsf{T}}\mathbf{F}^{-1}\nabla g$. The last column is the futuristic experiment of Section 10.6.5.

Side by side (Table 10.3, Figure 10.15). With temperature alone, the forecast reproduces Planck's errors on τ and $\ln A_s$ to within a few per cent, ω_c and $\theta_{\rm MC}$ to about ten per cent, and is too optimistic by about twenty per cent for ω_b and n_s. Adding polarization keeps the pattern. The correlation matrix (Figure 10.16) agrees entry by entry with that of the official Planck chains to within 0.28; the largest difference is the correlation of ω_b with n_s, +0.23 in our forecast and +0.52 in Planck's chains.

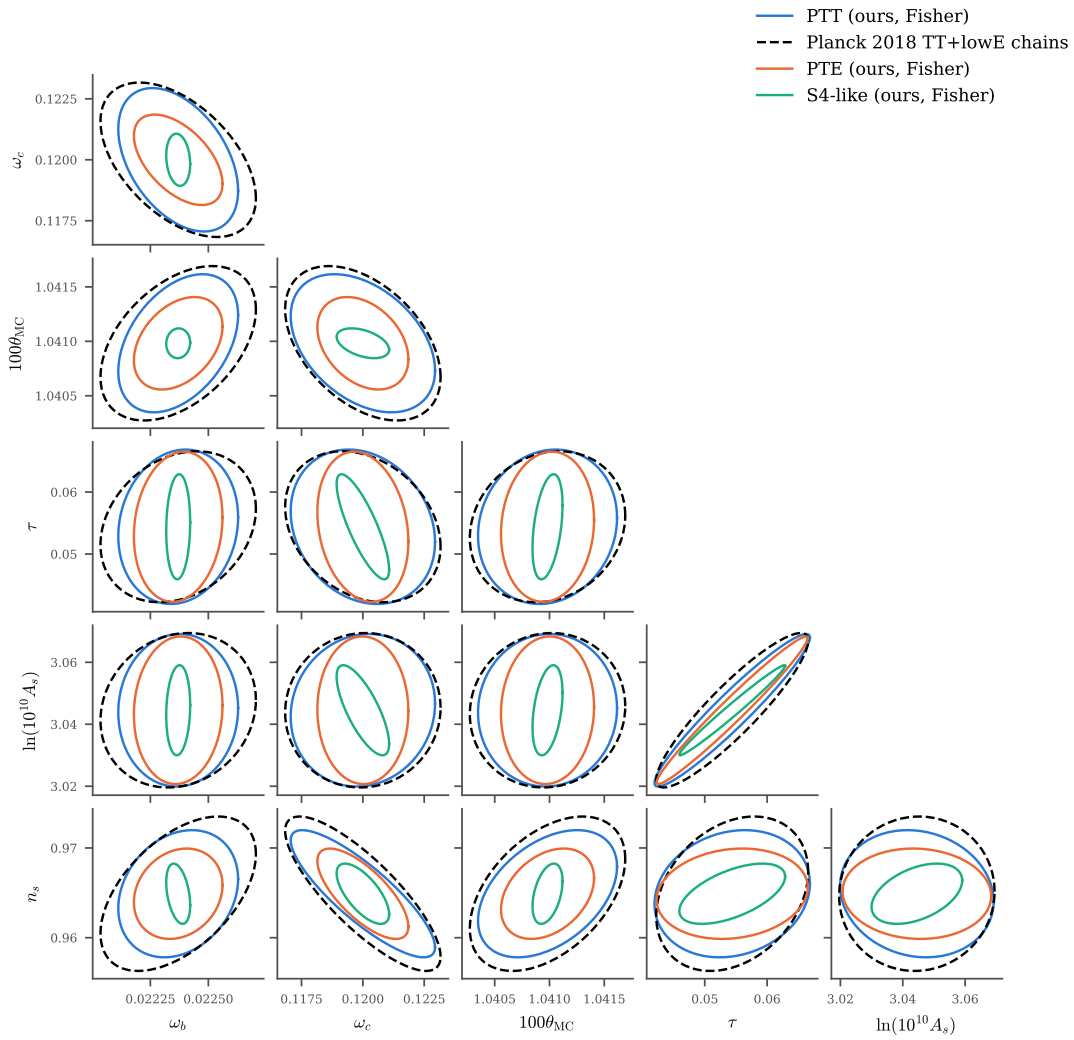


Figure 10.15: Forecast against outcome. Each panel shows the 68% ($\Delta\chi^2 = 2.30$) ellipse of a pair of ΛCDM parameters, all others marginalised, centred on the fiducial model. Black dashed: the covariance of the official Planck 2018 TT+lowE chains. Blue: our Fisher forecast for the same data. Orange: with polarization added. Green: a futuristic ground survey. The forecast reproduces the sizes and orientations of the Planck ellipses; the residual differences are in the parameters that share the high-ℓ damping tail with foregrounds.

Why the forecast is optimistic, and by how much. The idealised experiment knows its noise exactly and sees nothing but CMB. Planck's likelihood fits, in addition to the six cosmological parameters, some twenty nuisance parameters: the amplitudes of the cosmic infrared background, of unresolved point sources and of the Sunyaev–Zel'dovich effects in each frequency, Galactic dust residuals, and calibration factors (they are the extra columns of the chain covariance file). Their spectra rise towards high ℓ, where the damping tail of the CMB measures ω_b and n_s, so marginalising over them costs precision exactly in the two parameters our forecast overstates. The same mechanism explains the larger ω_b–n_s correlation in the real chains.

One nuisance can be added in a few lines, and it shows the mechanism at work. Planck compares the data with C_ℓ/y_P^2, where the overall map calibration y_P has a prior of 0.25% [25]. As a seventh parameter with $\partial C_\ell/\partial \ln y_P = -2C_\ell$ and prior Fisher element $1/0.0025^2$, it is marginalised by inverting the 7×7 matrix. The error on $10^9 A_s e^{-2\tau}$, which no other parameter can absorb, rises from 0.50% to 0.69%; Planck's TT+lowE value is 0.014/1.884 = 0.74%. The

error on $\ln A_s$ goes from 0.0162 to 0.0168. One honest nuisance parameter turns a forecast that was a third too optimistic for the best-measured amplitude into one that is within ten per cent of the published error.

Lesson. The forecast was right about everything the idealised sky contains, the sizes and the correlations, and wrong by about the amount that the missing nuisance parameters cost; adding even one of them shows where the gap comes from.

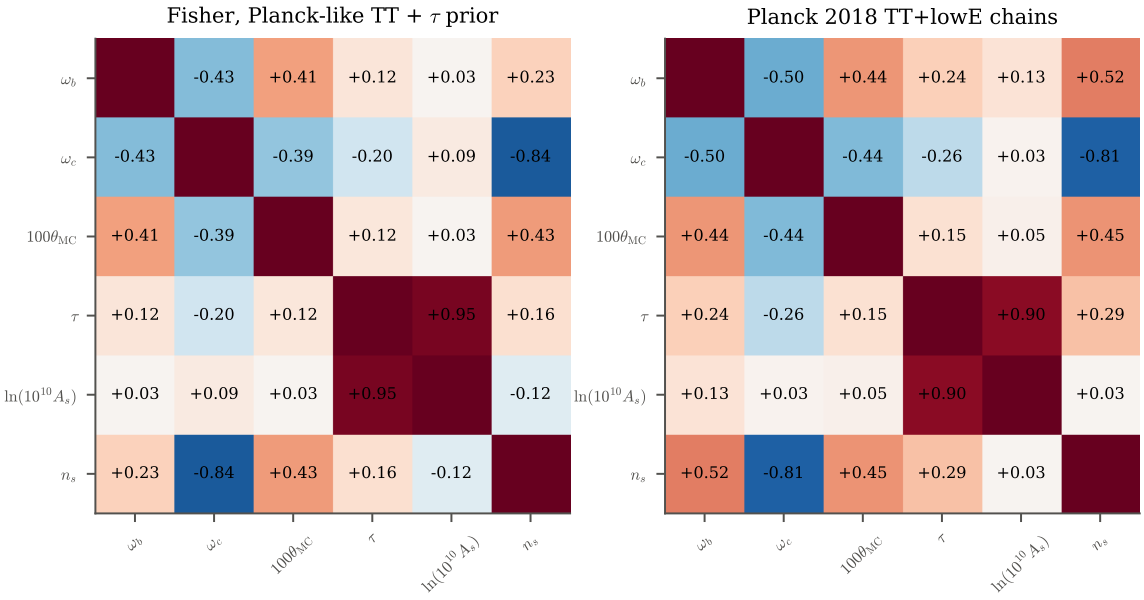


Figure 10.16: Correlation coefficients of the six parameters: our Fisher forecast for Planck-like temperature data with a τ prior (left) and the covariance of the official Planck 2018 TT+lowE chains (right). The two strongest correlations, τ with $\ln A_s$ and ω_c with n_s, are reproduced; the largest difference is ω_b with n_s.

10.6.4 The classic degeneracies

Amplitude and optical depth. Free electrons after reionisation scatter CMB photons out of the line of sight and suppress the anisotropy on all scales smaller than the horizon at reionisation by $e^{-2\tau}$, so for $\ell \gtrsim 10$ the temperature spectrum is proportional to $A_s e^{-2\tau}$ (cosmology notes). *Why this is a degeneracy.* Write that as $C_\ell \propto \exp(\ln A_s - 2\tau)$. Then

$$\frac{\partial C_\ell}{\partial \tau} = -2C_\ell, \qquad \frac{\partial C_\ell}{\partial \ln A_s} = C_\ell :$$

the two derivative vectors are parallel, both proportional to C_ℓ. In (10.31) every term of the $(\tau, \ln A_s)$ block is then C_ℓ^2 times a product of the factors -2 and 1, so the block is

$$f \begin{pmatrix} 4 & -2 \\ -2 & 1 \end{pmatrix} = f\, \boldsymbol{v}\boldsymbol{v}^{\mathsf{T}}, \qquad \boldsymbol{v} = \begin{pmatrix} -2 \\ 1 \end{pmatrix}, \qquad f = \sum_\ell \frac{2\ell+1}{2} f_{\mathrm{sky}} \left(\frac{C_\ell}{C_\ell + N_\ell} \right)^2 .$$

Its determinant is $f^2(4-4) = 0$: the matrix is singular, an exact degeneracy in the sense of Section 10.1.2. The unmeasured direction is the vector $\boldsymbol{u}$ with $\boldsymbol{v}^{\mathsf{T}}\boldsymbol{u} = 0$, $\boldsymbol{u} \propto (1, 2)$: raising τ by δ and $\ln A_s$ by 2δ leaves $\ln A_s - 2\tau$, and so the spectrum, unchanged. This is the line $\ln A_s - 2\tau = \text{const}$, of slope $\mathrm{d}\ln A_s/\mathrm{d}\tau = 2$. What the data do measure is the combination $y = \ln A_s - 2\tau$, with variance $1/f$.

What the real Fisher matrix finds. Our temperature-only Fisher matrix, *without* the τ prior, finds almost exactly this: the long axis of the $(\tau, \ln A_s)$ ellipse has slope $\mathrm{d}\ln A_s/\mathrm{d}\tau = 1.90$, the axes differ by a factor 26, the marginalised errors are $\sigma_\tau = 0.028$ and $\sigma_{\ln A_s} = 0.053$, yet the combination is measured to 0.59%. The slope is slightly below the exact null direction 2 because the lowest multipoles and the lensing of the spectrum (whose strength grows with A_s but not with τ) carry a little information on each factor separately, which turns the singular block into a merely ill-conditioned one.

Breaking it from outside. The low-ℓ polarization measures τ on its own, with error σ_τ; as a prior it adds $1/\sigma_\tau^2$ to $F_{\tau\tau}$ (Section 10.4.4). Now $\ln A_s = y + 2\tau$, with y fixed by the temperature data and τ by the prior, two independent measurements, so

$$\mathrm{Var}(\ln A_s) = \mathrm{Var}\, y + 4\,\sigma_\tau^2 \approx 4\,\sigma_\tau^2$$

whenever the combination is measured much better than $2\sigma_\tau$. That is why the amplitude error in Table 10.3 is set almost entirely by the τ prior: $\sigma_{\ln A_s} \approx 2\sigma_\tau$.

The horizon-angle degeneracy: $\Omega_m h^3$. The positions of the acoustic peaks fix the angle θ_* very precisely, and the peak heights fix the physical matter density $\omega_m = \Omega_m h^2 = \omega_b + \omega_c$ (the small neutrino contribution neglected), where $h = H_0/(100\,\mathrm{km\,s^{-1}\,Mpc^{-1}})$. The Hubble constant enters only through the distance to last scattering, so changes of h and Ω_m that keep θ_* fixed leave the spectrum nearly unchanged; Percival et al. found that this degeneracy follows $\Omega_m h^{3.4} = \text{const}$ for the data of 2002 [198], and Planck quotes $\Omega_m h^3$ as the best-measured combination [25], eq. (12). Our forecast can find the exponent itself. *The idea.* Among all the combinations $\Omega_m h^p$, the best-measured one is the one with the smallest variance: the direction in the $(\ln\Omega_m, \ln h)$ plane along which the data fix the parameters most tightly. So we write its variance as a function of p and minimise. Since $\Omega_m = \omega_m/h^2$, $\ln(\Omega_m h^p) = \ln\omega_m + (p-2)\ln h$, and its variance is

$$\begin{aligned} \mathrm{Var}\ln(\Omega_m h^p) &\overset{(1)}{=} \mathrm{Var}\ln\omega_m + 2(p-2)\,\mathrm{Cov}(\ln\omega_m, \ln h) + (p-2)^2\,\mathrm{Var}\ln h, \\ p_{\mathrm{best}} &\overset{(2)}{=} 2 - \frac{\mathrm{Cov}(\ln\omega_m, \ln h)}{\mathrm{Var}\ln h}, \end{aligned} \tag{10.35}$$

(1) $\Omega_m h^p = \omega_m h^{p-2}$; the variance of a sum.
(2) Set the derivative with respect to p to zero.

The ingredients. Both $\ln\omega_m$ and $\ln h$ are derived parameters, so their variances and covariance come from the delta method of Section 10.4.5, $\mathrm{Cov}(g, g') = \nabla g^{\mathsf{T}}\mathbf{F}^{-1}\nabla g'$. The gradient of $\ln\omega_m$ is $(1,1,0,0,0,0)/\omega_m$, nonzero only in ω_b and ω_c. The gradient of $\ln h$ is $\nabla H_0/H_0$, because $\ln h = \ln H_0 - \ln 100$ and $\mathrm{d}\ln H_0 = \mathrm{d}H_0/H_0$; CAMB's background code gives $\partial H_0/\partial(\omega_b, \omega_c, 100\theta_{\mathrm{MC}}) = (826, -350, 334)$, the other three entries being zero. The forecast gives $p_{\mathrm{best}} = 3.02$ for temperature and 2.97 with polarization: Planck's $\Omega_m h^3$, derived here from a Fisher matrix. With the much smaller errors of the futuristic survey it moves to 3.22, which is the point of Percival's different exponent: the best-measured combination depends on which multipoles carry the information. The error on $\Omega_m h^3$ itself is 0.00040 (0.42%) against Planck's 0.00046 for TT+lowE, and 0.00026 against 0.00029 with polarization.

Dark matter and tilt. The strongest correlation not broken by an external prior is ω_c with n_s, -0.84 in our forecast and -0.81 in Planck's chains. More dark matter brings matter–radiation equality earlier, which reduces the boost the first acoustic peaks receive from decaying potentials (cosmology notes); the first peaks drop relative to the damping tail, which a slightly redder tilt (smaller n_s) also does. Section 9.22.4 met the same correlation in the chain.

10.6.5 A futuristic experiment, the reach in ℓ and the sky fraction

A forecast is most useful for designing the next experiment. Take a ground-based survey with 1 μK arcmin in temperature ($\sqrt{2}$ in polarization), a 1.4′ beam, $f_{\rm sky} = 0.4$ and temperature and polarization at $30 \le \ell \le 3000$ (the atmosphere prevents a ground telescope from measuring the largest scales), plus the same τ prior. Its errors are in the last column of Table 10.3: ω_b improves by a factor 3.7 over the Planck-like experiment, $\theta_{\rm MC}$ by 3.1, ω_c by 1.7, n_s by 1.5, and even τ by 1.4, although the survey sees no multipole below 30. That last gain comes from lensing: the lensed spectra we use depend on A_s through the lensing strength but not on τ, so a precise damping tail separates the two factors. The same code with more parameters gives the larger forecasts of Jungman et al. and of Tegmark et al. (eleven parameters, [192], eqs. (19)–(20)): each new parameter costs one more derivative of the spectra and one more row and column of $\mathbf{F}$. A real forecast at this noise level must include the non-Gaussian covariance that lensing induces between multipoles, which reduces this gain (Section 10.7).

Which multipoles matter? Figure 10.17 recomputes the forecast for maximum multipoles from 500 to 3000. For the Planck-like experiment, stopping at $\ell_{\rm max} = 1000$ would multiply the error on n_s by 1.7 and on ω_b and $\theta_{\rm MC}$ by 1.5, while τ and $\ln A_s$ barely care (1.1): the third and higher peaks and the damping tail are where the densities and the tilt are measured.

How much sky? Every term of (10.31) is proportional to $f_{\rm sky}$, so the data part of $\mathbf{F}$ is too, and every error that is not dominated by a prior scales as $f_{\rm sky}^{-1/2}$. For the Planck-like experiment the error on n_s is 0.0046, 0.0033 and 0.0028 at $f_{\rm sky} = 0.3$, 0.57 and 0.8, and $0.0033\times\sqrt{0.57/0.3} \approx 0.0046$ and $0.0033 \times \sqrt{0.57/0.8} \approx 0.0028$; the error on τ, held by its prior, moves only from 0.0083 to 0.0078.

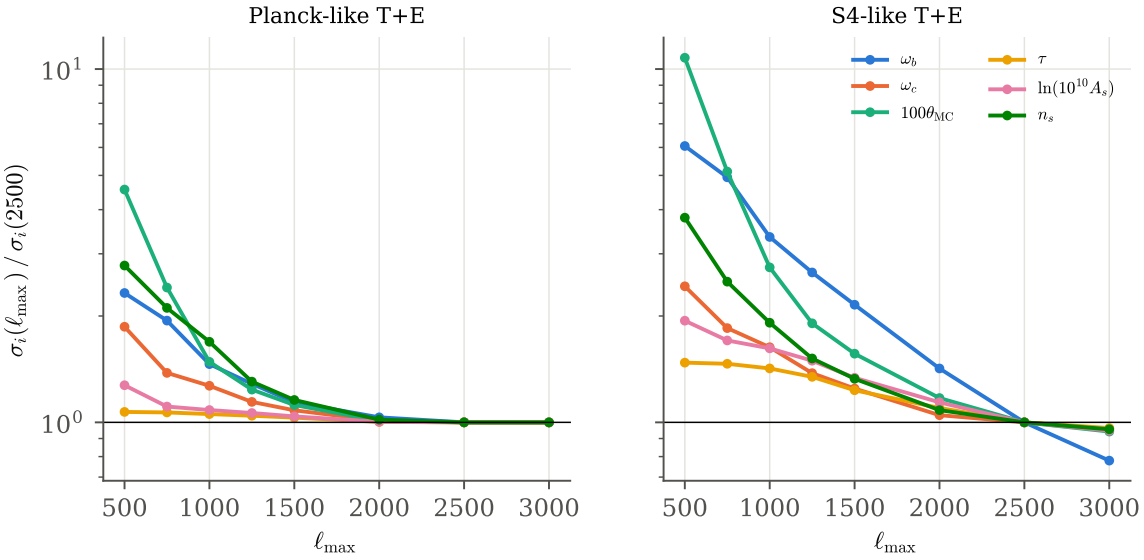


Figure 10.17: How the forecast errors depend on the highest multipole used, relative to $\ell_{\rm max} = 2500$. Left: the Planck-like experiment (polarization stops at $\ell = 2000$). Right: the futuristic survey. The densities, the angle and the tilt keep improving into the damping tail; the amplitude and τ are fixed early by the τ prior.

10.6.6 Is the forecast right? The chain and a thousand fits

A forecast is a statement about an idealised analysis of idealised data. Two independent checks are possible with tools the book already has: the Markov chain of Chapter 9, which maps the

actual posterior for one simulated data set, and a Monte Carlo of maximum-likelihood fits, which measures the actual scatter over many.

The chain. Run P6 of Section 9.22 sampled the six ΛCDM parameters from one simulated temperature spectrum: one 143 GHz-like channel (7.22′, 33 μK arcmin), $f_{\rm sky} = 0.57$, $2 \le \ell \le 2500$, the exact likelihood, and the same τ prior. It sampled in a different basis, $\boldsymbol{p} = (\ln 10^{10} A_s, n_s, H_0, \omega_b, \omega_c, \tau)$, with H_0 in place of $\theta_{\rm MC}$. Our Fisher matrix for the same experiment is computed in the Planck basis and moved to the chain's basis with the Jacobian rule (10.27). All entries of $\boldsymbol{J} = \partial\boldsymbol{\theta}/\partial\boldsymbol{p}$ are zeros and ones except the row of $100\theta_{\rm MC}$, which depends on H_0, ω_b and ω_c; CAMB's background code gives $\partial(100\theta_{\rm MC})/\partial(H_0, \omega_b, \omega_c) = (0.00299, -2.47, 1.046)$. Then:

	$\ln(10^{10}A_s)$	n_s	H_0	ω_b	ω_c	τ
Fisher, $\boldsymbol{J}^{\mathsf{T}}\mathbf{F}\boldsymbol{J}$ (CAMB derivatives)	0.0167	0.00527	0.932	0.000201	0.00215	0.00835
chain P6 (Chapter 9)	0.0166	0.00528	0.93	0.000201	0.00217	0.00833
Fisher of Chapter 9 (emulator derivatives)	0.0167	0.00529	0.936	0.000201	0.00216	0.00835
1000 maximum-likelihood fits	0.0169	0.00537	0.941	0.000204	0.00217	0.00841

The Fisher errors agree with the widths of the chain to within one per cent for every parameter, and every correlation coefficient agrees to within 0.02; the strongest, H_0 with ω_c, is -0.97 in both. Two quite different computations stand behind this: our derivatives were taken in another basis by the four-point rule and transformed with a Jacobian, chapter 9's came from an emulator of $\ln C_\ell$, and the chain never used a derivative at all.

A thousand fits. The chain describes the posterior of one data set; the Cramér–Rao reading of $\mathbf{F}^{-1}$ is about the scatter of estimates over many. `07_fisher_vs_mcmc.py` draws 1000 new spectra and τ measurements from the same experiment and finds the maximum of the likelihood of each, a thousand six-parameter maximisations. How should each be done?

One parameter first: a Newton step. Near the current guess θ_n the score, the slope of $\ln\mathcal{L}$, is nearly a straight line in θ, and its slope is minus the curvature: by (10.2), $S(\theta) \approx S(\theta_n) - \mathcal{H}\,(\theta - \theta_n)$. The maximum is where the slope of the hill vanishes, $S = 0$; solving the straight line for that point gives $\theta = \theta_n + S(\theta_n)/\mathcal{H}$, slope over curvature. If $\ln\mathcal{L}$ were an exact parabola, one step would land on its top; for a nearly parabolic hill each step lands much closer. *Many parameters.* The same expansion of the score vector is $\boldsymbol{S}(\boldsymbol{\theta}) \approx \boldsymbol{S}(\boldsymbol{\theta}_n) - \boldsymbol{\mathcal{H}}(\boldsymbol{\theta} - \boldsymbol{\theta}_n)$, and $\boldsymbol{S} = \mathbf{0}$ gives the step $\boldsymbol{\theta}_n + \boldsymbol{\mathcal{H}}^{-1}\boldsymbol{S}(\boldsymbol{\theta}_n)$. *Replace the curvature by its average.* The observed curvature $\boldsymbol{\mathcal{H}}$ of one noisy data set needs second derivatives of the model and, far from the peak, need not be positive definite, in which case the step can go downhill. Its average over data, the Fisher matrix (10.3), is known in closed form, (10.31), and is always positive definite, so the step $\mathbf{F}^{-1}\boldsymbol{S}$ always has a positive component along the uphill direction: $\boldsymbol{S}^{\mathsf{T}}\mathbf{F}^{-1}\boldsymbol{S} > 0$ whenever $\boldsymbol{S} \neq \mathbf{0}$. Near the peak the two curvatures differ only by the $1/\sqrt{n}$ fluctuation of Section 10.1.1, so little speed is lost. This iteration is called **Fisher scoring**:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \mathbf{F}(\boldsymbol{\theta})^{-1}\boldsymbol{S}(\boldsymbol{\theta}), \qquad S_i = \sum_\ell \frac{\nu_\ell}{2}\,\partial_i C_\ell\,\frac{\widehat{C}_\ell - C_\ell - N_\ell}{(C_\ell + N_\ell)^2}, \quad \nu_\ell = (2\ell+1) f_{\rm sky}, \tag{10.36}$$

with $\mathbf{F}(\boldsymbol{\theta})$ (10.31) evaluated at the current point.

The score of the exact likelihood. The score comes from the exact $-2\ln\mathcal{L} = \sum_\ell \nu_\ell[\widehat{C}_\ell/(C_\ell +$

$N_\ell) + \ln(C_\ell + N_\ell)]$ (Section 9.19.2). With $X_\ell = C_\ell + N_\ell$, which depends on θ_i only through C_ℓ,

$$S_i = \partial_i \ln \mathcal{L} \overset{(1)}{=} -\frac{1}{2}\sum_\ell \nu_\ell \Big[-\frac{\widehat{C}_\ell}{X_\ell^2} + \frac{1}{X_\ell}\Big]\partial_i C_\ell \overset{(2)}{=} \sum_\ell \frac{\nu_\ell}{2}\,\partial_i C_\ell\,\frac{\widehat{C}_\ell - X_\ell}{X_\ell^2},$$

$$\mathbb{E}[S_i] \overset{(3)}{=} \sum_\ell \frac{\nu_\ell}{2}\,\partial_i C_\ell\,\frac{\mathbb{E}\,\widehat{C}_\ell - X_\ell}{X_\ell^2} = 0.$$

(1) $\ln\mathcal{L}$ is $-\frac{1}{2}$ of the sum; by the chain rule $\partial_i(1/X) = -\partial_i C_\ell/X^2$ and $\partial_i \ln X = \partial_i C_\ell / X$.
(2) Put the bracket over the common denominator X_ℓ^2; this is the S_i of (10.36).
(3) The estimate is unbiased, $\mathbb{E}\,\widehat{C}_\ell = C_\ell + N_\ell = X_\ell$: the score has mean zero at the truth, as (10.4) requires.

The model spectrum is chapter 9's second-order emulator, so each step costs a fraction of a millisecond:

Listing 10.4: Fisher scoring: Newton's method with the expected curvature (`code/ch10/07_fisher_vs_mcmc.py`)

```
    for it in range(30):
        c, g = cl_and_grad(th)
        tot = c + N
        r = (chat - tot) / tot**2
        S = 0.5 * (g[:, use] * (nu * r)[use]).sum(1)             # score of the exact
    likelihood
        S[5] -= (th[5] - tau_obs) / SIG_TAU**2                   # ... plus the tau prior
        w = 0.5 * nu / tot**2
        Fk = (g[:, use] * w[use]) @ g[:, use].T                  # Fisher matrix at the
    current point
        Fk[5, 5] += 1 / SIG_TAU**2
        step = np.linalg.solve(Fk, S)
        th = th + step
        if np.max(np.abs(step) / sF) < 1e-4:
            break
```

The median fit converges in 4 iterations. The scatter of the 1000 estimates is in the last row of the table: between 1.01 and 1.02 times the Fisher errors, with a Monte Carlo uncertainty of ±2.2%; the mean offsets from the truth are a few hundredths of a standard deviation at most (−0.03 for $\ln A_s$, −0.03 for τ). The fraction of estimates inside the two-parameter 68.3% ellipse of $(\ln A_s, n_s)$ is 0.663, and inside the six-dimensional 68.3% ellipsoid ($\Delta\chi^2 \le 7.04$) it is 0.651, against 0.683 ± 0.015 (Figure 10.18). For this experiment maximum likelihood reaches the Cramér–Rao bound and the posterior has the Fisher width. The ellipses enclose close to what they claim; both fractions come out slightly low, by one to two Monte Carlo errors.

When the Fisher matrix fails. Everything above worked because the CMB spectrum is nearly linear in the parameters over a one-sigma region: the errors are tiny compared with the scale on which the spectrum bends. Vallisneri lists what goes wrong when that is not true [193], and the book has examples of each. At low signal-to-noise the likelihood is not a parabola over the scatter of the estimates (Cowan's experiment with 50 events, Section 10.2.3; the skewed decay-time posterior of Section 9.18). Near a physical boundary the Gaussian approximation puts probability where the parameter cannot go (a tensor ratio or a neutrino mass near zero). A quasi-singular Fisher matrix, with a direction the data barely constrain, makes $\mathbf{F}^{-1}$ meaningless in that direction unless a prior is added. In all three cases the Fisher matrix is still the right first calculation, and a chain on simulated data is the second.

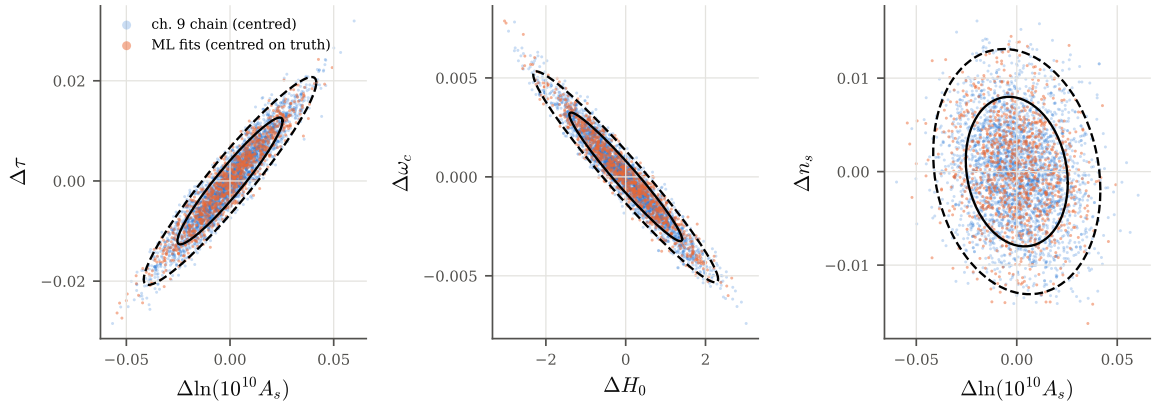


Figure 10.18: Three views of the same six-parameter problem in three planes: samples of the chapter-9 posterior for one simulated sky (blue, centred on their mean), the maximum-likelihood estimates from 1000 simulated skies (orange, centred on the truth), and the Fisher ellipses ($\Delta\chi^2 = 2.30, 6.18$) computed before any sky was simulated.

How this is used in real research: Verification and research

Here the Monte Carlo and the chain *verified* a forecast that analytic theory already gave, as the simulations of Chapter 8 verified the cosmic-variance formula. In research the roles reverse. Survey teams produce Fisher forecasts for every design option because they are instantaneous, and then run full chains on simulated data for the few designs that matter, because only the chain sees non-Gaussian posteriors, physical boundaries, nuisance parameters with non-Gaussian priors and likelihoods that are not Gaussian in the data. A forecast that survives that test is the one quoted in the proposal.

10.7 How it is done in research

A survey team writing a proposal, or a collaboration choosing between two telescope designs, runs the same calculation as Section 10.6, but every box of it is larger. A real forecast passes through those boxes in a fixed order (Figure 10.19); at each one the research version adds physics that the laptop version replaced by something simpler, and each replacement has a measurable cost.

1. Parameters, basis and fiducial. The model is ΛCDM plus the extensions the survey targets (the effective number of relativistic species $N_{\rm eff}$, which counts the light particles such as neutrinos that behaved like radiation in the early universe and is about three in the standard model; the sum of neutrino masses; the tensor ratio r), at a fiducial point taken from the best current fit. The basis is chosen close to what the data measure ($\theta_{\rm MC}$ rather than H_0, $\ln A_s$ at a pivot), so that the likelihood is nearly Gaussian in it and the finite differences behave; the quantities people quote (H_0; σ_8, the root-mean-square matter fluctuation in spheres of $8\,h^{-1}\,$Mpc; $\Omega_m h^3$) are derived afterwards with the Jacobian or delta-method rule (Section 10.4.5). *Here:* the six-parameter Planck basis at the Planck 2018 best fit.

2. Theory and derivatives. A Boltzmann code such as CAMB is run at raised accuracy settings and to higher multipoles than the survey will use, producing lensed TT, TE, EE, BB spectra and the spectrum of the lensing potential $C_\ell^{\phi\phi}$ (the projected gravitational potential that deflects CMB photons on their way to us). Derivatives are taken by finite differences and checked as in

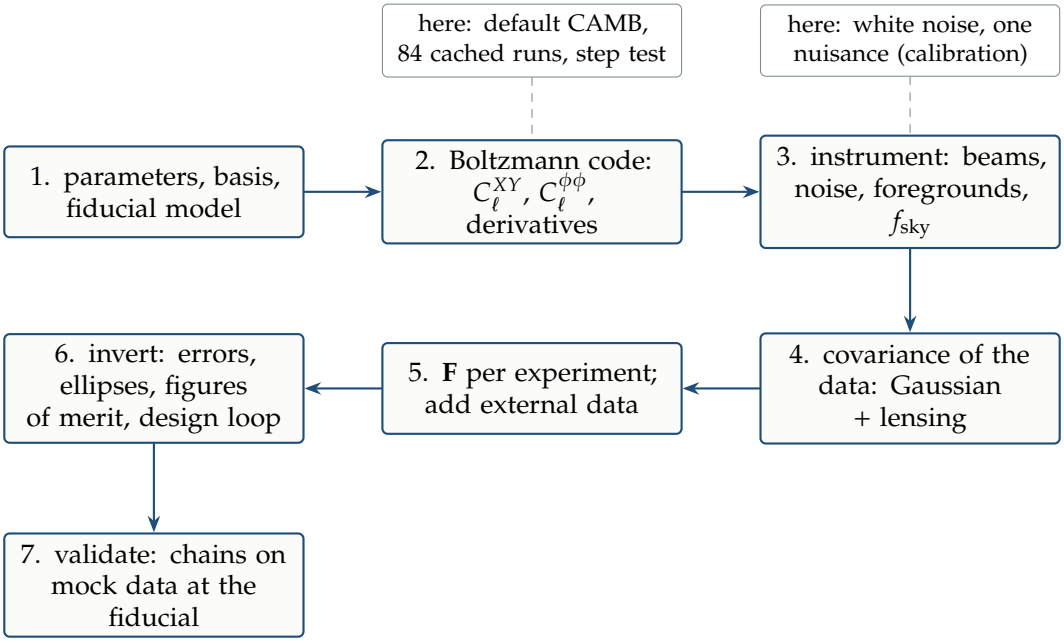


Figure 10.19: The stages of a research CMB forecast (blue boxes, in order), with two of the simplifications made by the laptop version (grey), each attached to the stage it simplifies.

Section 10.5: steps halved and doubled, two- and four-point rules compared, and the code's accuracy raised until the forecast stops moving. *Here:* default accuracy, $\ell \leq 3000$, 84 cached runs; the step test moved no error by more than a tenth of a per cent. The cost is nothing for a Planck-like experiment and possibly a few per cent for a survey that uses $\ell > 3000$.

3. The instrument. For each frequency channel: beam; white-noise depth; for a ground telescope the "$1/f$" noise, slow drifts of the atmospheric emission and of the detector gain whose power rises at low frequencies and therefore adds noise at low ℓ; the hit map, how many times each patch of sky was observed, which sets the noise of each pixel; and the sky mask. Foregrounds (Galactic dust and synchrotron, the cosmic infrared background, Sunyaev–Zel'dovich effects, point sources) are removed by a simulated component separation, which uses the several frequencies to subtract emission whose dependence on frequency differs from that of the CMB; what is left after it is added to the noise as a residual power spectrum. Lensing reconstruction, the estimate of the deflection field from the distortions it leaves in the small-scale CMB, has a noise of its own, a noise spectrum that plays for $C_\ell^{\phi\phi}$ the part N_ℓ plays for C_ℓ. *Here:* white noise, combined by inverse variance, no foregrounds, and a single nuisance parameter. The cost was measured in Section 10.6.3: errors on ω_b and n_s about twenty per cent too small for Planck, most of the gap on $A_s e^{-2\tau}$ closed by the calibration nuisance alone.

4. The covariance of the data. The Gaussian, $f_{\rm sky}$-scaled covariance (10.32) is the starting point. Research forecasts add the covariance that lensing induces between multipoles of the lensed spectra and between them and $C_\ell^{\phi\phi}$: all modes are deflected by the same large-scale field, so a fluctuation of that field moves power at many multipoles together and correlates them, which is why lensing is a non-Gaussian, mode-coupling operation. For small or irregular sky patches they add the mode coupling of the mask (cutting the sky mixes neighbouring multipoles, Chapter 8), estimated with simulations. *Here:* Gaussian covariance with $(2\ell+1)f_{\rm sky}$ modes. The cost is small for Planck-like noise and large for the futuristic survey, whose τ improvement of Section 10.6.5 comes from lensing information this approximation overstates.

5. Fisher matrices and external data. One matrix per experiment and per data product, in a common parameter basis (Jacobians where the bases differ), added together: the CMB, a τ measurement from large-scale polarization, baryon acoustic oscillations, supernovae (Section 10.4.4). Overlapping sky areas are not independent and must not simply be added; the overlap is treated as one experiment with a joint covariance. *Here:* one CMB matrix plus a τ prior.

6. Reading and design. Marginalised errors, triangle plots of ellipses, figures of merit, and the same numbers recomputed as the design changes (noise level, beam, sky fraction, observing time): a forecast is usually a curve or a grid, not a number, as in Figure 10.17. One more product is routine: the shift of the parameters that a small unmodelled systematic $\delta\boldsymbol{\mu}$ in the data would cause. *The question.* If the data have mean $\boldsymbol{\mu}(\boldsymbol{\theta}_0) + \delta\boldsymbol{\mu}$ instead of $\boldsymbol{\mu}(\boldsymbol{\theta}_0)$, where does the fit land on average? The fit lands where the score vanishes, so we ask where the *average* score vanishes. Name the vector $\boldsymbol{b}$ with components $b_i = \boldsymbol{\mu}_{,i}^{\mathsf{T}}\mathbf{C}^{-1}\delta\boldsymbol{\mu}$, the overlap of the systematic with the response of the data to θ_i; this is what $\left[\boldsymbol{\mu}_{,i}^{\mathsf{T}}\mathbf{C}^{-1}\delta\boldsymbol{\mu}\right]_i$ below denotes. Then, for Gaussian data,

$$
\begin{aligned}
\mathbb{E}\big[S_i(\boldsymbol{\theta}_0)\big] &\overset{(1)}{=} \boldsymbol{\mu}_{,i}^{\mathsf{T}}\mathbf{C}^{-1}\,\mathbb{E}[\boldsymbol{r}] + \tfrac{1}{2}\big(\mathbb{E}[\boldsymbol{r}^{\mathsf{T}}\mathbf{C}^{-1}\mathbf{C}_{,i}\mathbf{C}^{-1}\boldsymbol{r}] - \operatorname{Tr}\mathbf{C}^{-1}\mathbf{C}_{,i}\big) \overset{(2)}{=} \boldsymbol{\mu}_{,i}^{\mathsf{T}}\mathbf{C}^{-1}\delta\boldsymbol{\mu} = b_i,\\
\mathbb{E}\big[S_i(\boldsymbol{\theta}_0+\delta\boldsymbol{\theta})\big] &\overset{(3)}{=} b_i + \sum_j \delta\theta_j\;\mathbb{E}\big[\partial_j S_i\big] \overset{(4)}{=} b_i - \sum_j F_{ij}\,\delta\theta_j,\\
\mathbf{0} = \boldsymbol{b} - \mathbf{F}\,\delta\boldsymbol{\theta} &\overset{(5)}{=} \quad\Rightarrow\quad \delta\boldsymbol{\theta} = \mathbf{F}^{-1}\left[\boldsymbol{\mu}_{,i}^{\mathsf{T}}\mathbf{C}^{-1}\delta\boldsymbol{\mu}\right]_i,
\end{aligned}
\tag{10.37}
$$

(1) The score (10.21), with residual $\boldsymbol{r} = \boldsymbol{d} - \boldsymbol{\mu}(\boldsymbol{\theta}_0)$, averaged term by term.
(2) The residual is now $\boldsymbol{r} = \delta\boldsymbol{\mu} + \boldsymbol{n}$ with noise $\boldsymbol{n}$ of mean zero and covariance $\mathbf{C}$, so $\mathbb{E}\,\boldsymbol{r} = \delta\boldsymbol{\mu}$. In the second term, $\mathbb{E}[\boldsymbol{r}^{\mathsf{T}}\boldsymbol{M}\boldsymbol{r}] = \operatorname{Tr}(\boldsymbol{M}\mathbf{C}) + \delta\boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{M}\delta\boldsymbol{\mu}$; the trace equals $\operatorname{Tr}\mathbf{C}^{-1}\mathbf{C}_{,i}$ and cancels, and the rest is second order in $\delta\boldsymbol{\mu}$, which we drop.
(3) Expand the average score to first order in the shift $\delta\boldsymbol{\theta}$.
(4) The average derivative of the score is the average second derivative of $\ln\mathcal{L}$, which is $-F_{ij}$ by (10.3) (to zeroth order in $\delta\boldsymbol{\mu}$).
(5) Set the average score to zero and solve for $\delta\boldsymbol{\theta}$.

This is the bias a systematic produces, in units that can be compared with $\mathbf{F}^{-1}$ directly. *A check with one parameter.* Let the data be a template of known shape with amplitude A, $\boldsymbol{\mu} = A\boldsymbol{s}$, so $\boldsymbol{\mu}_{,A} = \boldsymbol{s}$ and $F_{AA} = \boldsymbol{s}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{s}$ (Section 10.3.3), and let the systematic be a calibration error that scales the whole signal, $\delta\boldsymbol{\mu} = \epsilon\boldsymbol{\mu} = \epsilon A\boldsymbol{s}$. Then $b_A = \epsilon A\,\boldsymbol{s}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{s} = \epsilon A\,F_{AA}$ and $\delta A = b_A/F_{AA} = \epsilon A$: a calibration off by 1% moves the amplitude by 1%, as it must. Requirements on calibration, beam knowledge and foreground residuals are set with this formula: the systematic is acceptable if $\delta\boldsymbol{\theta}$ is small against the errors.

7. Validation. For the designs that matter, mock data are generated at the fiducial model and run through the full likelihood with a sampler (for example Cobaya, a Markov chain code of the kind built in Chapter 9), and the posterior is compared with the Fisher ellipses. This is where non-Gaussian posteriors, physical boundaries (a neutrino mass near zero, a tensor ratio near zero) and priors that are not Gaussian are found. *Here:* the chain of Chapter 9 with an emulated spectrum and 1000 maximum-likelihood fits by Fisher scoring, which agreed with the forecast to one or two per cent. For the full problem the chains take cluster hours per design; ours took seconds, because the model was an emulator and the likelihood had no nuisance parameters.

How far a forecast is from reality. Figure 10.20 condenses the comparison with Planck into one picture: the ratio of our forecast errors to the published ones. It is the realistic expectation

stage	research version	laptop version (here)	what the reduction costs
theory	high-accuracy Boltzmann runs to $\ell \sim 5000$	default CAMB, $\ell \leq 3000$, 84 cached runs	$< 0.1\%$ on the errors (step test); more for $\ell > 3000$
noise	per-channel, anisotropic, $1/f$, foreground residuals	white, inverse-variance combined	ω_b, n_s errors $\sim 20\%$ too small vs Planck
nuisances	~ 20 foreground and calibration parameters	calibration only (as a demonstration)	residual optimism in damping-tail parameters
sky	mask with mode coupling, simulated covariance	$(2\ell+1)f_{\rm sky}$ modes	a few per cent at $f_{\rm sky} \gtrsim 0.4$; worse at low ℓ
lensing	lensing reconstruction, lensing-induced covariance	lensed spectra, Gaussian covariance	futuristic τ, A_s gains overstated
low ℓ	non-Gaussian low-ℓ likelihood	Gaussian τ prior 0.0086	no skewness of τ near zero
validation	full-likelihood chains on mock data	chain of Chapter 9 (emulator), 1000 fits	none here; decisive with real complexity

Table 10.4: Each reduction made in the laptop version, what the research analysis does instead, and what the reduction costs.

for a forecast built as above: right in the parameters set by a prior or by the peak positions, optimistic by up to a quarter where unmodelled nuisances live. A research forecast closes most of the remaining gap by adding stages 3 and 4 in full.

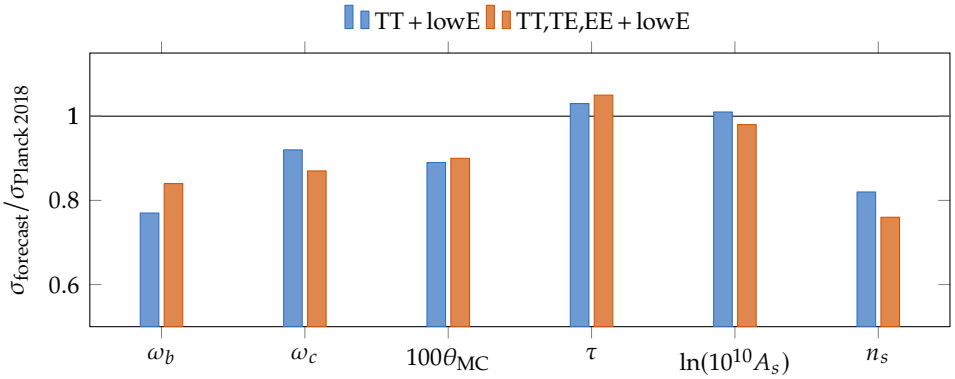


Figure 10.20: Our Fisher forecast errors divided by the published Planck 2018 errors for the same data combinations. A bar at one is a perfect forecast. The amplitude and τ are held by the τ prior and come out right; the densities, the angle and above all ω_b and n_s, which share the damping tail with foregrounds, come out too small, because the idealised experiment has no foreground nuisances to marginalise.

Recap

- The Fisher matrix $F_{ij} = -\,\mathbb{E}[\partial_i \partial_j \ln \mathcal{L}] = \mathbb{E}[S_i S_j]$, at a fiducial model, is the average curvature of the log-likelihood and the covariance of the score; it needs a model of the experiment, not data.
- Cramér–Rao: $\mathrm{Cov}\,\hat{\boldsymbol{\theta}} - \mathbf{F}^{-1}$ is positive semidefinite for every unbiased estimator; maximum likelihood reaches the bound for linear-Gaussian models and for informative data (Cowan's experiment from ~ 500 events on).
- Gaussian data: $F_{ij} = \boldsymbol{\mu}_{,i}^{\mathsf{T}} \mathbf{C}^{-1} \boldsymbol{\mu}_{,j} + \frac{1}{2}\,\mathrm{Tr}[\mathbf{C}^{-1}\mathbf{C}_{,i}\mathbf{C}^{-1}\mathbf{C}_{,j}]$, first derivatives only; the CMB

is the covariance channel, $F_{ij} = \sum_\ell \frac{(2\ell+1)f_{\rm sky}}{2} \partial_i C_\ell \partial_j C_\ell / (C_\ell + N_\ell)^2$.

- Marginalised error $\sqrt{(\mathbf{F}^{-1})_{ii}}$ (invert, then read) versus conditional $1/\sqrt{F_{ii}}$; the difference is a Schur complement. Two-parameter ellipses at $\Delta\chi^2 = 2.30$, 6.18 hold 68.3%, 95.4%.
- Priors and independent experiments add as matrices; a change of parameters is $\boldsymbol{J}^{\mathsf{T}}\mathbf{F}\boldsymbol{J}$; derived quantities use $\nabla g^{\mathsf{T}}\mathbf{F}^{-1}\nabla g$; nuisances are marginalised by inverting.
- Finite-difference steps of a fraction of the expected error, checked by halving and doubling; noisy derivatives make forecasts too optimistic.
- Reproduced: Cowan's two-parameter errors; Knox 1995's tilt and pivot amplitude errors and pivot $\ell_* = 210$; Planck 2018 errors within $\sim 25\%$ (better with a calibration nuisance); $\Omega_m h^3$ and $A_s e^{-2\tau}$ as the degeneracy directions; the chapter-9 chain and 1000 maximum-likelihood fits within one or two per cent.

10.8 Noise alone, or a signal as well?

LISA will deliver, for years, a few numbers per second: the measured changes in the lengths of its three arms, converted into a strain $d(t)$. A strain is a fractional change of length: a passing wave of amplitude h changes an arm by $\delta L/L = h/2$ (Sections T2.1.3 and T2.3), and the detector adds its own noise (Section T2.7). Somewhere in that stream there may be a black hole of a thousand solar masses swallowing a neutron star, its wave completing millions of cycles. Or there may be nothing but the detector's own jitter. Before any parameter is estimated, and long before a dark-matter environment is discussed, there is one question that has to be answered first:

Is this data stream consistent with noise alone, or does it contain a gravitational-wave signal in addition to noise?

The answer shapes everything that follows. It fixes the likelihood we will use for the rest of the chapter, it tells us what the signal-to-noise ratio is, and its curvature is the Fisher matrix that the forecasts of Sections 10.9 and 10.11 are built on.

Why this is not the CMB problem again. In Chapters 7 and 8 the signal was a temperature map. We could look at it, see hot and cold spots, and the noise was characterised by simulating it many times with the instrument model. A gravitational wave gives us nothing like that. Its strain is far smaller than the noise at each instant (in the toy experiment below the noise is 5.314 times larger than the peak of the signal), and there is no foreground map, no picture in which the signal stands out. LISA does have something called a foreground, the confusion signal of the millions of unresolved white-dwarf binaries in our Galaxy (Section T2.7), but it is not a map either: nobody can subtract it, and it enters only statistically, as an extra term in the noise spectrum (Section 10.11.1). Worse, we never know the particular noise realisation that is in the data, only its statistics: its power spectral density, which is itself estimated. So we cannot subtract "the" noise. We have to consider *all* the noise realisations the detector could have produced, weighted by how probable each is, and ask which of them, together with a possible signal, could have produced the stream we recorded.

10.8.1 The event, and what could have produced it

We reason exactly as in the Monty Hall problem of Chapter 5.

The event. What we observed is the whole data stream, or equivalently its Fourier transform $\tilde{d}(f) = \int d(t) e^{-2\pi i f t} \mathrm{d}t$ (the book's Fourier convention, Eq. (T2.48); for a real stream the negative frequencies only repeat the positive ones, Eq. (T2.49)) at the frequencies $f_k = k/T$ of a stretch of length T. Call it $\boldsymbol{d}$.

What could have produced it? Before any number, we list the situations. The data are always the detector noise n plus whatever the sky sent:

- H_0: the sky sent nothing, so $d = n$;
- $H_1(\boldsymbol{\theta})$: a binary with parameters $\boldsymbol{\theta}$ sent a wave $h(\boldsymbol{\theta})$, so $d = n + h(\boldsymbol{\theta})$.

The thing we do not know, under either hypothesis, is the noise realisation n. This is the analogue of the unknown position of the car.

How probable is the event under each hypothesis? Take $H_1(\boldsymbol{\theta})$ first. Of all the noise realisations the detector could have produced, exactly one, $n = d - h(\boldsymbol{\theta})$, turns the prediction $h(\boldsymbol{\theta})$ into the data we hold. Every other realisation would have produced different data. So the probability of the data is the probability that the detector produced that one required realisation. In symbols, summing over all realisations with their probabilities $p_n(n)$ and keeping only the one that produces d,

$$p(\boldsymbol{d} \mid \boldsymbol{\theta}) \overset{(1)}{=} \int p_n(n)\, p(\boldsymbol{d} \mid n, \boldsymbol{\theta})\, \mathrm{d}n \overset{(2)}{=} \int p_n(n)\, \delta\big(d - h(\boldsymbol{\theta}) - n\big)\, \mathrm{d}n \overset{(3)}{=} p_n\big(d - h(\boldsymbol{\theta})\big). \tag{10.38}$$

(1) Marginalisation over the unknown noise realisation (Chapter 1): the total probability of the data is the sum, over every possible realisation, of its probability times the probability of the data given it.
(2) Given the noise and the signal, the data are fixed: $d = h(\boldsymbol{\theta}) + n$ with certainty, a delta function.
(3) The delta function picks out the single realisation $n = d - h(\boldsymbol{\theta})$.

Under H_0 the same argument with $h = 0$ gives $p(\boldsymbol{d} \mid H_0) = p_n(d)$: the data themselves must be the noise. This is the likelihood of Chapter 3, "how plausible is the fluctuation needed to turn the prediction into the data", now for a detector whose noise is a whole random function of time. Notice how little the marginalisation over the noise realisation cost: it is exact, and almost trivial, because once the data and the signal are given the noise is fixed, $n = d - h$. What we genuinely do not know is the statistics of the noise, its spectrum, and integrating over that is the marginalisation that does real work (Section 10.8.6). What remains is to know p_n.

10.8.2 The noise in the frequency domain, from its inputs

Inputs. The physical inputs are three. The noise is *stationary*: its statistical properties do not change during the stretch, so the correlation $C(\tau) = \mathbb{E}[n(t)n(t+\tau)]$ depends only on the lag τ. It has zero mean. And it is *Gaussian*: any finite set of samples has a multivariate Gaussian distribution, which is what we expect when many small independent disturbances add (Chapter 2). The mathematical inputs are two: the correlation dies off over a time τ_c much shorter than T, and the variance is finite. Everything else is derived.

Why go to Fourier space. In the time domain neighbouring samples are correlated, so p_n involves the full $N \times N$ covariance matrix $C_{jm} = C\big((j-m)\Delta t\big)$. A matrix whose entries depend only on $j - m$ is (up to the edges) a circulant matrix, and circulant matrices are diagonalised by

the discrete Fourier transform. So in Fourier space the noise modes should be uncorrelated. Let us check this and find their variance. With N samples spaced by Δt, $T = N\Delta t$, the discrete transform is $\tilde{n}_k = \Delta t \sum_j n_j e^{-2\pi i jk/N}$, an approximation to the continuous one at $f_k = k/T$. Then

$$\begin{aligned}\mathbb{E}[\tilde{n}_k \tilde{n}_l^*] &\overset{(1)}{=} \Delta t^2 \sum_j \sum_m C\big((j-m)\Delta t\big)\, e^{-2\pi i(jk-ml)/N} \\ &\overset{(2)}{=} \Delta t^2 \sum_s C(s\Delta t)\, e^{-2\pi i s l/N} \sum_j e^{-2\pi i j(k-l)/N} \\ &\overset{(3)}{=} \Delta t^2 N\, \delta_{kl} \sum_s C(s\Delta t)\, e^{-2\pi i f_l s\Delta t} \overset{(4)}{=} T\, \delta_{kl} \int C(\tau) e^{-2\pi i f_l \tau} \mathrm{d}\tau \overset{(5)}{=} \frac{T}{2} S_n(f_k)\, \delta_{kl}. \end{aligned} \tag{10.39}$$

(1) Insert the definition of $\tilde{n}_k$ twice and use $\mathbb{E}[n_j n_m] = C((j-m)\Delta t)$, stationarity.
(2) Write $m = j - s$, so $jk - ml = j(k-l) + sl$. Treating the sum over s as independent of j is exact for a periodic stretch and ignores edge terms of relative size τ_c/T, our second mathematical input.
(3) $\sum_{j=0}^{N-1} e^{-2\pi i j(k-l)/N} = N\delta_{kl}$, the orthogonality of the discrete Fourier modes; and $sl/N = f_l\, s\Delta t$ because $f_l = l/(N\Delta t)$.
(4) $N\Delta t = T$, and $\Delta t \sum_s$ becomes $\int \mathrm{d}\tau$ when $\Delta t \ll \tau_c$.
(5) The **one-sided power spectral density** is defined by $S_n(f) = 2\int C(\tau) e^{-2\pi i f\tau} \mathrm{d}\tau$ for $f > 0$ (the Wiener–Khinchin relation); the factor 2 folds the negative frequencies onto the positive ones.

The same steps with $\tilde{n}_l$ instead of $\tilde{n}_l^*$ give $\mathbb{E}[\tilde{n}_k \tilde{n}_l] = 0$ for $0 < k, l < N/2$, so the real and imaginary parts of each mode are uncorrelated and of equal variance. This is the statement $\langle \tilde{n}(f)\tilde{n}^*(f')\rangle = \frac{1}{2}\delta(f-f')S_n(f)$ of [84, eq. (35)] and [87], written for a finite stretch: $\delta(0)$ becomes T.

The density of a noise realisation. Each mode is a complex Gaussian with $\mathbb{E}\,|\tilde{n}_k|^2 = \sigma_k^2 \equiv \frac{T}{2} S_n(f_k)$; its real and imaginary parts are independent with variance $\sigma_k^2/2$ each, so its density is $\frac{1}{\pi\sigma_k^2} e^{-|\tilde{n}_k|^2/\sigma_k^2}$. Different modes are independent, so the density of the whole realisation is the product:

$$\ln p_n(n) \overset{(1)}{=} -\sum_k \frac{|\tilde{n}_k|^2}{\sigma_k^2} + \text{const} \overset{(2)}{=} -2\sum_k \Delta f\, \frac{|\tilde{n}_k|^2}{S_n(f_k)} + \text{const} \overset{(3)}{=} -\frac{1}{2}(n \mid n) + \text{const}, \tag{10.40}$$

(1) Product of independent complex Gaussians; the normalisations $1/(\pi\sigma_k^2)$ do not depend on n.
(2) $1/\sigma_k^2 = 2/(TS_n) = 2\Delta f/S_n$, since the mode spacing is $\Delta f = 1/T$.
(3) For long T the sum is an integral, $-2\int |\tilde{n}|^2/S_n\, \mathrm{d}f$, which is $-\frac{1}{2}(n \mid n)$ with the inner product defined below.

Working definition 10.4 (noise-weighted inner product and signal-to-noise ratio)

For two signals a, b and a detector with one-sided noise spectrum $S_n(f)$,

$$(a \mid b) = 4\,\mathrm{Re}\int_0^\infty \frac{\tilde{a}(f)\,\tilde{b}^*(f)}{S_n(f)}\,\mathrm{d}f$$

[a dot product in which each frequency counts in proportion to how quiet the detector is there]. The noise density is $p_n(n) \propto e^{-(n|n)/2}$, and the **optimal signal-to-noise ratio** of a signal is $\rho = \sqrt{(h \mid h)}$. *Operationally:* evaluate the integral on the frequencies where S_n is finite; for a sky-averaged detector use the sky-averaged waveform with the matching

S_n. The averages over sky position, polarisation and orbital inclination are worked out in Section T2.3, and LISA's S_n, built from its physical noise sources, in Section T2.7.

Putting (10.40) into (10.38), the likelihood of the data under $H_1(\boldsymbol{\theta})$ and under H_0 is

$$p(\boldsymbol{d}\,|\,\boldsymbol{\theta}) \propto \exp\Big[-\tfrac{1}{2}\big(d-h(\boldsymbol{\theta})\,\big|\,d-h(\boldsymbol{\theta})\big)\Big], \qquad p(\boldsymbol{d}\,|\,H_0) \propto \exp\Big[-\tfrac{1}{2}(d\,|\,d)\Big], \tag{10.41}$$

with the same constant in both [87, eq. (7)], [199]. It is the χ^2 of Chapter 3 for correlated Gaussian noise, $\boldsymbol{r}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{r}$, with the covariance diagonalised by the Fourier transform (Example 0.14 sketched it).

Key idea: The gravitational-wave likelihood

We never know the noise realisation in the data, so we weigh all of them by their probability. Under a hypothesis that predicts the signal $h(\boldsymbol{\theta})$, only the realisation $n = d - h(\boldsymbol{\theta})$ reproduces the data, and the likelihood is the probability of that realisation: $p(\boldsymbol{d}\,|\,\boldsymbol{\theta}) \propto e^{-(d-h|d-h)/2}$. The noise statistics enter only through $S_n(f)$, which weights each frequency by the inverse of its noise power.

10.8.3 The best test: the matched filter

The question in words. If the template h were known exactly, which single number computed from the data best separates "noise alone" from "noise plus h"? The Neyman–Pearson lemma of Chapter 4 answers: for two simple hypotheses the most powerful test at any false-alarm rate thresholds the likelihood ratio. From (10.41),

$$\ln\Lambda = \ln\frac{p(\boldsymbol{d}\,|\,H_1)}{p(\boldsymbol{d}\,|\,H_0)} \overset{(1)}{=} -\tfrac{1}{2}\big[(d|d) - 2(d|h) + (h|h)\big] + \tfrac{1}{2}(d|d) \overset{(2)}{=} (d\,|\,h) - \tfrac{1}{2}(h\,|\,h). \tag{10.42}$$

(1) Expand $(d-h|d-h)$ using linearity and the symmetry $(a|b) = (b|a)$ of the real inner product.
(2) The $(d|d)$ terms cancel: the data enter only through their overlap with the template.

Since $(h|h)$ is a fixed number, thresholding Λ is thresholding the **matched filter** $(d\,|\,h)$: the data correlated with the template, each frequency weighted by $1/S_n$. Its sampling distribution under each hypothesis follows from one property of the noise.

The overlap of noise with any signal. $(n|a)$ is a sum of Gaussian modes, so it is Gaussian with mean zero. Its covariance with $(n|b)$ is

$$\mathbb{E}\big[(n|a)(n|b)\big] \overset{(1)}{=} 4\Delta f^2 \sum_{k,l} \frac{\mathbb{E}\big[(\tilde n_k \tilde a_k^* + \tilde n_k^* \tilde a_k)(\tilde n_l \tilde b_l^* + \tilde n_l^* \tilde b_l)\big]}{S_n(f_k)S_n(f_l)} \overset{(2)}{=} 4\Delta f^2 \sum_k \frac{\sigma_k^2\,(\tilde a_k^* \tilde b_k + \tilde a_k \tilde b_k^*)}{S_n(f_k)^2} \overset{(3)}{=} (a\,|\,b). \tag{10.43}$$

(1) $(n|a) = 2\Delta f \sum_k (\tilde n_k \tilde a_k^* + \tilde n_k^* \tilde a_k)/S_n(f_k)$, since $2\,\mathrm{Re}\,z = z + z^*$.
(2) Only $\mathbb{E}[\tilde n_k \tilde n_l^*] = \sigma_k^2 \delta_{kl}$ survives; $\mathbb{E}[\tilde n_k \tilde n_l] = 0$.
(3) $\sigma_k^2 = \frac{T}{2} S_n$ and $T\Delta f = 1$, so $4\Delta f^2 \sigma_k^2 / S_n^2 = 2\Delta f / S_n$, and $2\Delta f \sum_k 2\,\mathrm{Re}(\tilde a_k \tilde b_k^*)/S_n$ is $(a|b)$.

So the noise-weighted inner product is exactly the covariance of the noise projections [87, eq. (8)]. With $\rho^2 = (h|h)$, the normalised statistic $z = (d|h)/\rho$ is

$$z\,|\,H_0 \sim \mathcal{N}(0,1), \qquad z\,|\,H_1 = \rho + \frac{(n|h)}{\rho} \sim \mathcal{N}(\rho, 1). \tag{10.44}$$

The two hypotheses produce two unit Gaussians a distance ρ apart. That is the meaning of the signal-to-noise ratio: the separation, in units of the noise spread, between what noise alone and noise plus the signal would give. The power of the test at false-alarm probability α follows at once: $P_D = \mathrm{P}(z > z_\alpha \mid H_1) = \Phi(\rho - z_\alpha)$, with z_α the upper-α point of $\mathcal{N}(0,1)$ and Φ the standard Gaussian cumulative distribution.

Fitting the amplitude too. So far the template had a known size. Suppose the wave is $A\,h$ with h the template of signal-to-noise ratio ρ and A an unknown strength ($A = 0$ is noise alone, $A = 1$ the template as normalised). The log-likelihood ratio of a signal of strength A against noise alone is (10.42) with $h \to Ah$, and it can be maximised over A by setting its derivative to zero:

$$\ln\Lambda(A) \overset{(1)}{=} A\,(d|h) - \tfrac{1}{2}A^2\rho^2, \qquad \frac{\partial \ln\Lambda}{\partial A} \overset{(2)}{=} (d|h) - A\rho^2 = 0 \;\Rightarrow\; \hat{A} = \frac{(d|h)}{\rho^2} = \frac{z}{\rho},$$
$$\ln\Lambda(\hat{A}) \overset{(3)}{=} \frac{(d|h)^2}{\rho^2} - \frac{(d|h)^2}{2\rho^2} \overset{(4)}{=} \frac{z^2}{2}. \tag{10.45}$$

(1) (10.42) has h in its first term and h twice in its second, and the inner product is bilinear.
(2) A parabola in A with a maximum, at $\hat{A}$.
(3) Insert $\hat{A} = (d|h)/\rho^2$ and $z = (d|h)/\rho$.
(4) Combine the two terms.

The largest log-likelihood ratio over every possible strength is $z^2/2$: the larger the filter output in either sign, the more the data prefer a signal over noise alone. (The same z^2 is the amount by which the fitted template lowers the residual, $(d|d) - (d - \hat{A}h|d - \hat{A}h) = z^2$.) By (10.44) the fitted strength has mean A and standard deviation $1/\rho$: the error on the amplitude of a known template is the inverse of the signal-to-noise ratio, the simplest example of the errors of Section 10.9. In the script, the maximum of $\ln\Lambda(A)$ over a fine grid in A equals $z^2/2$ for each of 2000 noise realisations to 1.421×10^{-14}, and $\hat{A}$ has mean 1 and scatter $0.9885/\rho$ for a signal of unit strength (`16_whiten_chi2.py`).

How far the claim of optimality reaches. The Neyman–Pearson lemma compares two *simple* hypotheses: noise alone against noise plus one exactly specified wave. For that pair the matched filter is the best possible test. It does not say that one filter is best when the wave has unknown parameters. The alternative is then a whole family $H_1(\boldsymbol{\theta})$, and no single test is the most powerful against all its members. The practical answer is the **generalised likelihood ratio**: maximise Λ over the unknown parameters, as just done for A and as Section 10.8.4 does for ϕ_c and t_0 (the unknown phase and arrival time of the wave, the coalescence constants of Section T2.4.3), and take the largest value as the statistic. Its distribution under noise alone is not the one-dimensional Gaussian of (10.44): the maximum over many possible templates is larger than any single filter (the look-elsewhere effect), and with many parameters the distribution is usually found empirically (Section 10.14). One more caveat belongs to LISA specifically. The hypothesis "noise alone" presumes a stretch of data with no signal in it, and LISA will not have one: tens of thousands of Galactic binaries and other sources overlap throughout the mission. "Noise alone" here is the idealised null of a single-source problem; the real analysis fits all the sources and the noise together.

Example: a chirp buried in coloured noise. To see all of this on data we need a stretch short enough to simulate thousands of times (script `08_detection.py`). LISA's sources last years, so for this check we use a ground-based detector: the analytic noise shape of Cutler and Flanagan [87, eq. (4)] and a $10 + 10\,M_\odot$ binary, whose whole chirp from 20 Hz to the innermost stable

circular orbit (ISCO, where the slow inspiral ends, Eq. (T2.46)) at 219.9 Hz lasts 5.959 s and fits in $T = 16$ s sampled at 1024 Hz. Every step is identical for LISA with S_n of Eq. (T2.77) and years of data. The noise is generated directly from (10.39):

Listing 10.5: Coloured Gaussian noise and the inner product (`code/ch10/08_detection.py`)

```
fs, T = 1024.0, 16.0
N = int(fs * T)
dt, df = 1 / fs, 1 / T
fk = np.fft.rfftfreq(N, dt)
band = (fk >= 15.0) & (fk <= 500.0)
Sn = np.full_like(fk, np.inf)
Sn[band] = L.Sn_cf94(fk[band])
sig2 = 0.5 * T * Sn                                   # E |n_k|^2 = (T/2) S_n(f_k)


def noise_fd(n):
    """n realisations of the noise Fourier modes (complex Gaussian, independent across k)."""
    s = np.sqrt(np.where(band, sig2, 0.0) / 2)
    return s * (rng.standard_normal((n, fk.size)) + 1j * rng.standard_normal((n, fk.size)))


def inner(a, b):
    """(a|b) = 4 Re sum a b* / S_n df  over the band."""
    return 4 * df * np.real(np.sum((a * np.conj(b))[..., band] / Sn[band], axis=-1))
```

Line 37 is (10.39): the variance of each Fourier mode. Lines 42–43 draw the real and imaginary parts independently, each with half that variance. Line 48 is the inner product of the working definition, as a sum over the band with $\mathrm{d}f \to 1/T$. The template is the stationary-phase chirp of Eq. (T2.56), the Fourier transform of the inspiral derived in Section T2.5, scaled so that $\rho = 5$.

Figure 10.21 shows one realisation. In the raw strain the chirp is invisible. Dividing each Fourier mode by $\sqrt{S_n}$ (**whitening**) makes the noise equally loud at all frequencies, and even then the chirp is a faint ripple below the noise. The Welch estimate of the spectrum, computed from this one stretch, follows the true S_n: what we can know about the noise is its spectrum, not its realisation.

Over 2000 realisations of each hypothesis (independent noise for the two), the statistic z has mean -7.002×10^{-4} and standard deviation 0.9939 under H_0, and mean 5.001 and standard deviation 0.9945 under H_1, as (10.44) predicts (Figure 10.22, left). At a false-alarm probability of 1% the threshold is $z_\alpha = 2.326$, and the detection probability is $\Phi(5 - 2.326) = 0.9962$; the simulation gives 0.9965.

10.8.4 When the arrival time and phase are unknown

The test of Section 10.8.3 was built for a template we know exactly. Imagine running it on a real detector. A binary that merged somewhere in the Universe sends us a chirp whose shape we can predict from its masses, but two of its numbers are pure accident: the moment t_0 at which it reaches the detector, and the phase ϕ_c of the wave at that moment, which depends on where the two bodies happened to be on their orbit. Neither tells us anything about the source, and neither can be known in advance. A template with the wrong arrival time lines up its oscillations against noise; a template with the wrong phase lines up its crests against the signal's troughs. In both cases the matched filter $(d|h)$ of (10.42) comes out near zero, even with a loud signal in the data.

Must we run a separate matched filter for every phase and every arrival time we can think of, and if we search over all of them, what does noise alone now produce? The answer to the first part is no. The

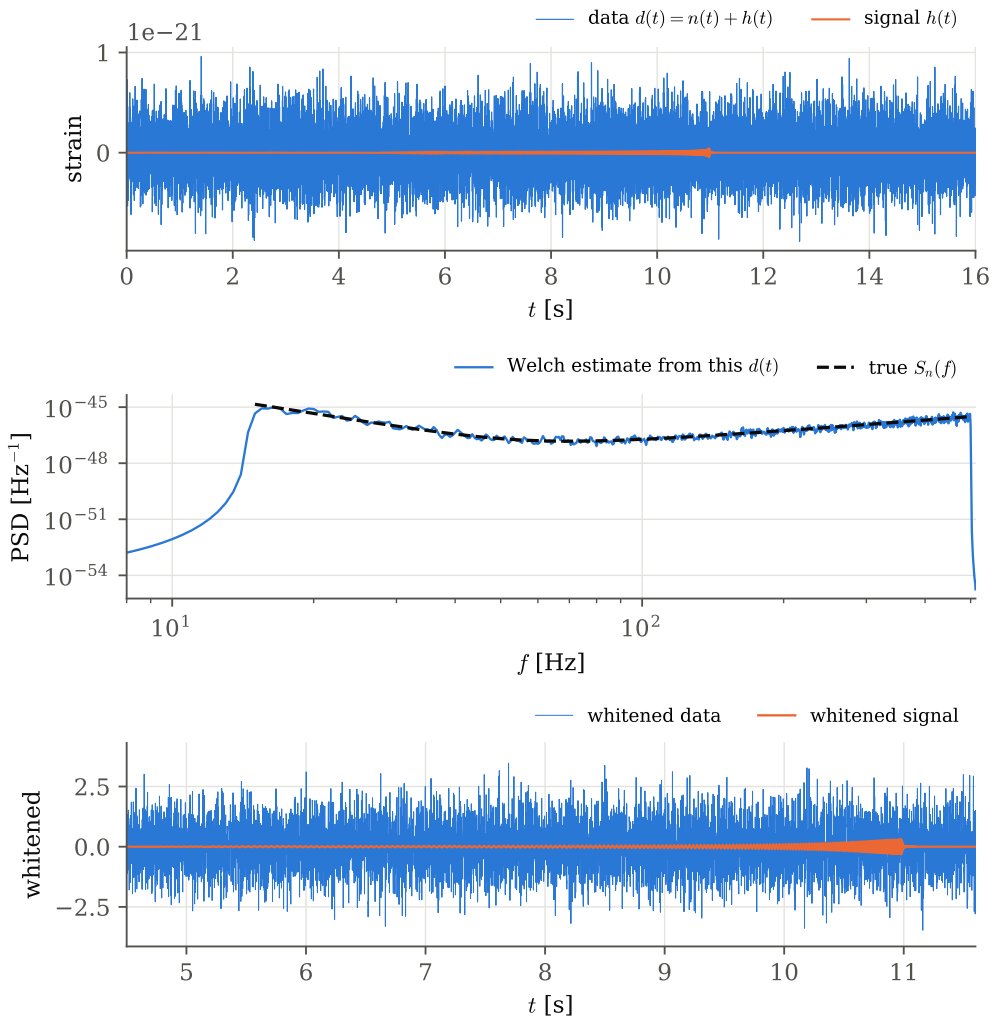


Figure 10.21: One stretch of simulated detector output with a $\rho = 5$ chirp. Top: the data and the signal; the signal is a small fraction of the noise and cannot be seen. Middle: the spectrum estimated from this stretch alone falls on the true $S_n(f)$ (dashed). Bottom: after whitening, the noise is flat and the chirp (orange) is still well below it. Only correlating with the right template brings it out.

phase costs exactly two filters instead of one, and every arrival time comes out of one Fourier transform. The answer to the second part is the price we pay for not knowing.

The phase

The simplest wave first. Take a pure tone of known amplitude A and angular frequency ω, with an unknown phase ϕ_c,

$$h(t;\phi_c) = A\cos(\omega t + \phi_c). \tag{10.46}$$

As a function of ϕ_c this looks hopeless: every value of ϕ_c is a different waveform. But the addition rule $\cos(\alpha+\beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta$, with $\alpha = \omega t$ and $\beta = \phi_c$, separates the part that depends on time from the part that depends on the phase:

$$h(t;\phi_c) \overset{(1)}{=} A\cos\omega t\,\cos\phi_c - A\sin\omega t\,\sin\phi_c \overset{(2)}{=} h_c(t)\cos\phi_c + h_s(t)\sin\phi_c, \tag{10.47}$$

with $h_c(t) \equiv A\cos\omega t$ and $h_s(t) \equiv -A\sin\omega t$.

(1) The addition rule.
(2) Give the two fixed time functions names. All of the dependence on ϕ_c is now in the two numbers $\cos\phi_c$ and $\sin\phi_c$.

So every phase is a mixture of the same two waveforms, h_c and h_s. They differ by a quarter of a cycle ($-\sin\omega t = \cos(\omega t + \pi/2)$), and are called the two **quadratures** of the signal.

The filter of a mixture is the mixture of the filters. The inner product $(a|b)$ of the working definition in Section 10.8.2 is an integral, so it is linear in each slot: $(d \mid \alpha a + \beta b) = \alpha(d|a) + \beta(d|b)$ for real numbers α, β. Applied to (10.47),

$$(d \mid h) = x_c \cos\phi_c + x_s \sin\phi_c, \qquad x_c \equiv (d \mid h_c), \quad x_s \equiv (d \mid h_s). \tag{10.48}$$

The data are correlated with each quadrature once, giving two numbers x_c and x_s; the filter for any phase is then a recombination of these two numbers, with no further reference to the data.

Which phase fits best? Treat (x_c, x_s) as a point in a plane and describe it in polar coordinates: its distance from the origin $r = \sqrt{x_c^2 + x_s^2}$ and its angle ϕ_0, so that $x_c = r\cos\phi_0$ and $x_s = r\sin\phi_0$. Then

$$(d \mid h) \overset{(1)}{=} r\cos\phi_0\cos\phi_c + r\sin\phi_0\sin\phi_c \overset{(2)}{=} r\cos(\phi_c - \phi_0). \tag{10.49}$$

(1) Substitute the polar form of x_c and x_s into (10.48).
(2) The addition rule again, read backwards: $\cos(\phi_c - \phi_0) = \cos\phi_c\cos\phi_0 + \sin\phi_c\sin\phi_0$.

A cosine is at most 1, and reaches it when its argument is zero. So the best phase is $\phi_c = \phi_0$, the angle of the point (x_c, x_s), and the filter there equals the distance r:

$$\max_{\phi_c} (d \mid h) = \sqrt{(d \mid h_c)^2 + (d \mid h_s)^2}. \tag{10.50}$$

A continuum of phases has been searched with two filters and one square root. The best-fitting phase also comes for free, as ϕ_0.

The two filters as one complex number. A point in the plane is a complex number, so define

$$x \equiv x_c + i\,x_s, \qquad |x| = \sqrt{x_c^2 + x_s^2} = \max_{\phi_c}(d \mid h), \qquad \arg x = \phi_0. \tag{10.51}$$

In the frequency domain x takes a compact form. We need the Fourier transforms of the two quadratures, in the convention of Section 10.8.2, $\tilde h(f) = \int h(t)\, e^{-2\pi i f t}\, \mathrm{d}t$. Write the two time functions as complex exponentials,

$$\cos\omega t = \tfrac{1}{2}\big(e^{i\omega t} + e^{-i\omega t}\big), \qquad -\sin\omega t = -\tfrac{1}{2i}\big(e^{i\omega t} - e^{-i\omega t}\big) = \tfrac{i}{2}\big(e^{i\omega t} - e^{-i\omega t}\big). \tag{10.52}$$

The inner product only uses positive frequencies, and there only the $e^{+i\omega t}$ term contributes. Its coefficient is $\frac{1}{2}$ for the cosine and $\frac{i}{2}$ for the sine quadrature. So, at every positive frequency,

$$\tilde h_s(f) = i\,\tilde h_c(f). \tag{10.53}$$

A quarter-cycle delay is multiplication by i. Now compute x_s from the definition of the inner product:

$$x_s = (d \mid h_s) \overset{(1)}{=} 4\,\mathrm{Re}\int_0^\infty \frac{\tilde d\,(i\tilde h_c)^*}{S_n}\,\mathrm{d}f \overset{(2)}{=} 4\,\mathrm{Re}\Big[-i\int_0^\infty \frac{\tilde d\,\tilde h_c^*}{S_n}\,\mathrm{d}f\Big] \overset{(3)}{=} 4\,\mathrm{Im}\int_0^\infty \frac{\tilde d\,\tilde h_c^*}{S_n}\,\mathrm{d}f. \tag{10.54}$$

(1) The inner product of Section 10.8.2, with (10.53).
(2) $(i\tilde h_c)^* = -i\,\tilde h_c^*$, and the constant $-i$ comes out of the integral.

(3) For any complex number $w = a + ib$, $-iw = b - ia$, so $\mathrm{Re}(-iw) = b = \mathrm{Im}\, w$.

And $x_c = 4\,\mathrm{Re}\int \tilde{d}\,\tilde{h}_c^*/S_n\,\mathrm{d}f$ by definition. Putting the two together,

$$x = x_c + i\,x_s = 4\int_0^\infty \frac{\tilde{d}(f)\,\tilde{h}_c^*(f)}{S_n(f)}\,\mathrm{d}f. \tag{10.55}$$

This is the inner product with the "Re" left out. Its real part is the cosine filter, its imaginary part the sine filter, and its modulus is the filter already maximised over the phase.

From the tone to the chirp. Nothing above used the fact that the wave was a pure tone, except to find (10.53). For the stationary-phase chirp of Eq. (T2.56), $\tilde{h}(f) = \mathcal{A} f^{-7/6} e^{-i\Psi(f)}$ with $\Psi(f) = 2\pi f t_c - \phi_c - \pi/4 + \ldots$ (Section T2.4.3), the unknown phase enters Ψ as an additive constant, so it multiplies the whole Fourier transform by one factor $e^{i\phi_c} = \cos\phi_c + i\sin\phi_c$. That is (10.47) again, with $\tilde{h}_c$ the template at $\phi_c = 0$ and $\tilde{h}_s = i\tilde{h}_c$. The complex filter (10.55) therefore searches all phases of the chirp as well. (With $-\phi_c$ inside Ψ and $e^{-i\Psi}$ outside, the factor is $e^{+i\phi_c}$; the opposite convention would only flip the sign of ϕ_0.)

The arrival time

A delay is a phase that grows with frequency. A signal arriving t_0 later is the same waveform shifted, $h(t) \to h(t - t_0)$. Its Fourier transform is

$$\int h(t - t_0)\,e^{-2\pi i f t}\,\mathrm{d}t \overset{(1)}{=} \int h(s)\,e^{-2\pi i f (s+t_0)}\,\mathrm{d}s \overset{(2)}{=} e^{-2\pi i f t_0}\,\tilde{h}(f). \tag{10.56}$$

(1) Substitute $s = t - t_0$, so $t = s + t_0$ and $\mathrm{d}t = \mathrm{d}s$.

(2) The factor $e^{-2\pi i f t_0}$ does not depend on s and comes out; what is left is $\tilde{h}(f)$.

Put the shifted template into the complex filter (10.55). The complex conjugate turns $e^{-2\pi i f t_0}$ into $e^{+2\pi i f t_0}$:

$$x(t_0) = 4\int_0^\infty \frac{\tilde{d}(f)\,\tilde{h}_c^*(f)}{S_n(f)}\,e^{+2\pi i f t_0}\,\mathrm{d}f = \int_0^\infty Q(f)\,e^{+2\pi i f t_0}\,\mathrm{d}f, \qquad Q(f) \equiv \frac{4\,\tilde{d}(f)\,\tilde{h}_c^*(f)}{S_n(f)}. \tag{10.57}$$

One transform for all arrival times. Look at what depends on what. $Q(f)$ is built from the data, the template and the noise spectrum; it does not know about t_0, so it is computed once. The arrival time appears only in $e^{+2\pi i f t_0}$, and an integral of a function of f against $e^{+2\pi i f t_0}$ is, by definition, the inverse Fourier transform of that function, evaluated at the time t_0. So

$$x(t_0) = \mathcal{F}^{-1}[Q](t_0) \quad \text{for every } t_0 \text{ at once.} \tag{10.58}$$

On a stretch of N samples there are N possible arrival times. Filtering at each one separately would take N operations per lag, N^2 in all; the fast Fourier transform delivers all N values in about $N \log_2 N$. For the 16384 lags of our example that is about a thousand times less work per template, and it is how real searches work [200]. In the script, `q` is $Q(f)$ on the frequency grid (zero outside the band), `x` is $x(t_0)$ on every lag, and the statistic maximised over both unknowns, $z_{\max} = \max_{t_0} |x(t_0)|/\rho$, is stored:

Listing 10.6: The filter at every arrival time (`code/ch10/08_detection.py`)

```
        zl.append(inner(d, h0) / norm)
        # complex filter over all lags: x(t0) = 4 sum d h* / S df e^{+2 pi i f t0}; |x| maximises over the phase
        q = np.zeros_like(d)
        q[:, band] = d[:, band] * np.conj(h0[band]) / Sn[band]
        x = 4 * df * np.fft.ifft(np.concatenate([q, np.zeros((q.shape[0], N - q.shape[1]))], axis=1), axis=1) * N
```

What noise alone now gives

Searching costs something. Every time we try another phase or another arrival time, we give the noise another chance to look like a signal. To set a threshold we need the distribution of $z_{\max}$ when the data are noise alone, $d = n$.

One arrival time. Fix t_0. Under H_0 the two filters are $x_c = (n|h_c)$ and $x_s = (n|h_s)$. By (10.43), each is Gaussian with mean zero, and their variances and covariance are inner products of the templates. These follow from (10.53):

$$(h_s \mid h_s) \overset{(1)}{=} 4\int_0^\infty \frac{|i\tilde{h}_c|^2}{S_n}\,\mathrm{d}f \overset{(2)}{=} (h_c \mid h_c) = \rho^2,$$
$$(h_c \mid h_s) \overset{(3)}{=} 4\,\mathrm{Re}\Big[-i\int_0^\infty \frac{|\tilde{h}_c|^2}{S_n}\,\mathrm{d}f\Big] \overset{(4)}{=} 0. \tag{10.59}$$

(1) The inner product of a signal with itself; the real part is not needed because $|\cdot|^2$ is real.
(2) $|i| = 1$; and $(h_c|h_c) = \rho^2$ is the optimal signal-to-noise ratio of the working definition.
(3) The same manipulation as for x_s above, with $\tilde{d}$ replaced by $\tilde{h}_c$.
(4) The integral is a real number, so $-i$ times it is purely imaginary, and its real part vanishes.

So x_c/ρ and x_s/ρ are Gaussians with mean zero, variance one and covariance zero. For jointly Gaussian variables zero covariance means independence, so they are two independent $\mathcal{N}(0,1)$ variables, and

$$\frac{|x|^2}{\rho^2} = \left(\frac{x_c}{\rho}\right)^2 + \left(\frac{x_s}{\rho}\right)^2 \sim \chi^2_2 \tag{10.60}$$

is the sum of the squares of two independent unit Gaussians, a χ^2 with two degrees of freedom (Section 2.13.2). With $p = 2$ the χ^2_p density (2.92) becomes simply $\frac{1}{2}e^{-v/2}$, an exponential, so its tail is $\mathrm{P}(\chi^2_2 > v) = \int_v^\infty \frac{1}{2}e^{-v'/2}\mathrm{d}v' = e^{-v/2}$. With $v = u^2$,

$$\mathrm{P}\Big(\frac{|x|}{\rho} > u \,\Big|\, H_0\Big) = e^{-u^2/2}. \tag{10.61}$$

Compare the known-phase test: there $\mathrm{P}(z > u)$ was a one-sided Gaussian tail, about $e^{-u^2/2}/(u\sqrt{2\pi})$ for large u. Searching over the phase has raised the false-alarm probability by a factor of about $u\sqrt{2\pi}$, roughly 6 at $u = 2.5$: the price of one unknown angle.

All arrival times. Now we take the largest $|x(t_0)|/\rho$ over all $N = 16384$ lags. If the lags were independent, the largest would stay below u only if every one did:

$$\mathrm{P}(z_{\max} > u \mid H_0) = 1 - \big(1 - e^{-u^2/2}\big)^N \approx N\,e^{-u^2/2} \quad \text{while this is small.} \tag{10.62}$$

This is the look-elsewhere effect of Chapter 4: the largest of many noise peaks is larger than a typical one, and a threshold that was safe for one trial is not safe for N. The simulation shows how large the effect is. Under H_0 the median of $z_{\max}$ is 4.226, the 1% threshold rises from $z_\alpha = 2.326$ to 5.123, and for the same $\rho = 5$ the detection probability falls from 0.9965 to 0.54 (Figure 10.22, middle and right).

How many trials, really? The lags are not independent: the filter output is smooth over a time about the inverse of the bandwidth, so neighbouring lags see nearly the same noise. We can measure how many independent trials the search is worth by solving (10.62) for N at the measured threshold, $N_{\mathrm{eff}}e^{-u^2/2} = 0.01$. This gives $N_{\mathrm{eff}} \approx 5000$, fewer than the 16384 lags. It is a rough number: the 1% threshold rests on the twenty largest of the 2000 noise-only values, and N_{eff} grows as $e^{u^2/2}$, so a small error in u becomes a large one in N_{eff}.

The lesson, in one line: not knowing when and in what phase a signal arrives costs nothing in computation (two filters and one Fourier transform) but a great deal in significance, because noise alone now gets thousands of chances. This is why ground-based searches demand $\rho \gtrsim 8$ before claiming a detection, and why LISA forecasts assume a threshold near $\rho \approx 15$ for these long signals [89].

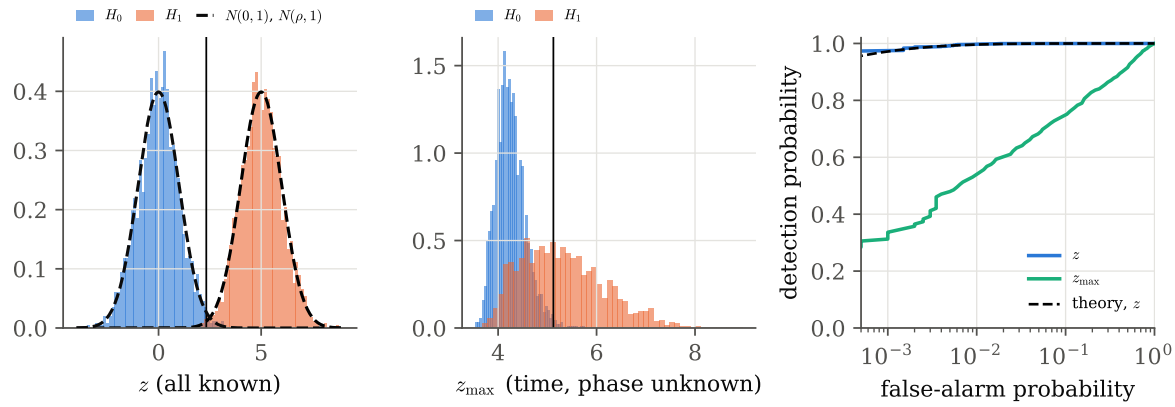


Figure 10.22: Sampling distributions of the detection statistic over 2000 noise realisations, $\rho = 5$. Left: with the template, time and phase known, z is $\mathcal{N}(0,1)$ under H_0 and $\mathcal{N}(\rho,1)$ under H_1 (dashed); the vertical line is the 1% false-alarm threshold. Middle: maximising over arrival time and phase pushes the noise-only distribution up, and the threshold with it. Right: detection probability against false-alarm probability (receiver operating characteristic); the price of not knowing t_0 and ϕ_c is the gap between the two curves.

10.8.5 Whitening, and what one stretch can tell us about the spectrum

Everything above rests on the claim that each Fourier mode of the noise is a complex Gaussian of variance $\sigma_k^2 = \frac{T}{2} S_n(f_k)$ (Eq. (10.39)). Two practical questions follow at once. How do we look at the data so that this claim can be checked? And if the spectrum is not known, what does one stretch of data say about it?

Whitening. Divide each mode by its own standard deviation, $w_k = \tilde{n}_k/\sigma_k$. Then $\mathbb{E}\,|w_k|^2 = 1$ in every bin, however steep the spectrum: the noise has become equally loud at all frequencies, **white**. The same division turns the noise-weighted inner product into an ordinary dot product. For two signals with whitened modes $a_k = \tilde{a}_k/\sigma_k$, $b_k = \tilde{b}_k/\sigma_k$,

$$(a \mid b) = 4\Delta f \sum_k \frac{\mathrm{Re}\,\tilde{a}_k \tilde{b}_k^*}{S_n(f_k)} \overset{(1)}{=} 2 \sum_k \mathrm{Re}\left(a_k b_k^*\right), \tag{10.63}$$

(1) $\sigma_k^2 = T S_n(f_k)/2$ and $\Delta f = 1/T$, so $\mathrm{Re}(\tilde{a}_k\tilde{b}_k^*)/\sigma_k^2 = 2\Delta f\,\mathrm{Re}(\tilde{a}_k\tilde{b}_k^*)/S_n(f_k)$.

so in whitened data the filter $(d|h)$ is a plain correlation between the whitened data and the whitened template, and $(n|n) = 2\sum_k |w_k|^2$ is the sum of squares of independent unit-variance numbers, as in (10.40). This is what the bottom panel of Figure 10.21 shows.

What one stretch says about the spectrum. The natural estimate of σ_k^2 from one stretch is the **periodogram** of that bin, $|\tilde{n}_k|^2$. Its distribution follows from (10.39): the real and imaginary parts of $\tilde{n}_k$ are independent Gaussians of variance $\sigma_k^2/2$, so with a and b two independent $\mathcal{N}(0,1)$ variables,

$$\frac{|\tilde{n}_k|^2}{\sigma_k^2} = \frac{a^2+b^2}{2} = \tfrac{1}{2}\chi_2^2, \qquad \text{density } e^{-v},\ \ v \geq 0, \tag{10.64}$$

the exponential distribution (the χ^2_2 density is $\frac{1}{2}e^{-v/2}$, as used for (10.61)). An exponential has its standard deviation equal to its mean. So a single-bin periodogram is right on average but is wrong by 100% in any one realisation, and a longer stretch does not help: it only supplies more bins, each as noisy as before. The way to a good estimate is to average. Cut the data into M segments, compute the periodogram of each and average the M values in each bin (Welch's method). The sum of M independent exponentials is a gamma variable of mean M and variance M, so the average has mean one and variance $1/M$: the fractional error falls as $1/\sqrt{M}$. This sum is exactly the quantity Q of the next subsection.

Cutting the data into segments has a price. Each segment is the product of the signal with a rectangular window, and in the Fourier domain a product is a convolution: the loud low-frequency part of the spectrum leaks, through the slowly decaying wings of the window's transform, into frequency bins where the true noise is orders of magnitude lower. For detector noise that spans many decades the leakage can dominate the estimate. Segments are therefore multiplied by a smooth taper that falls to zero at the ends (for example a Tukey or Hann window), which suppresses the wings, and overlapping segments recover the information the taper discards. In the script, the periodograms of 400 simulated noise realisations in the band have mean 0.9998 and variance 0.9996 in units of σ_k^2, as (10.64) requires; a Kolmogorov–Smirnov test against the exponential gives $p = 0.08844$; and the real parts of the whitened modes have standard deviation 0.9999 and excess kurtosis -0.002005, as for a Gaussian of variance one half per component after the factor $\sqrt{2}$ (`16_whiten_chi2.py`). On real data the same two checks, an exponential periodogram and Gaussian whitened modes, are what tells us the Gaussian stationary noise model is adequate, or that glitches and drifts are present.

10.8.6 When the noise spectrum is itself estimated

The likelihood (10.41) assumed S_n known. It never is: it is estimated from data near the event, for instance by averaging $|\tilde{n}_k|^2$ over M stretches without a signal. Plugging the estimate $\hat{S}_n$ into the Gaussian likelihood treats a noisy number as exact. What should we do instead?

The question in words. What is the probability of the on-source residual in one frequency bin, regardless of the true noise level in that bin, given what the off-source stretches told us about that level? It is a marginal: the true variance is a nuisance, and we integrate it out, exactly as V was integrated out to obtain Student's t in Chapter 2.

Ingredients. In bin k write $s = \mathbb{E}\,|\tilde{n}_k|^2$ for the unknown true variance, $r = \tilde{d}_k - \tilde{h}_k$ for the on-source residual, and $x_1, \dots, x_M$ for the same bin in M off-source stretches. Each is a complex Gaussian of variance s. With the prior $p(s) \propto 1/s$ (scale-free, Chapter 5), the off-source data give the posterior

$$p(s \mid x) \overset{(1)}{=} \frac{\prod_m \frac{1}{\pi s} e^{-|x_m|^2/s} \cdot \frac{1}{s}}{\int (\cdots) \mathrm{d}s} \overset{(2)}{=} \frac{Q^M}{\Gamma(M)}\, s^{-M-1} e^{-Q/s}, \qquad Q \equiv \sum_{m=1}^{M} |x_m|^2 \equiv M\hat{s}. \tag{10.65}$$

(1) Bayes' theorem with the complex-Gaussian density of each stretch and the prior $1/s$.

(2) Collect $s^{-M} \cdot s^{-1}$ and $e^{-\sum |x_m|^2/s}$; the normalisation is the gamma integral $\int_0^\infty s^{-a-1} e^{-Q/s} \mathrm{d}s = \Gamma(a)/Q^a$ (substitute $y = Q/s$), with $a = M$. $\hat{s}$ is the averaged estimate of the variance.

Marginalise. The density of the on-source residual, regardless of s, is

$$\begin{aligned} p(r \mid x) &\overset{(1)}{=} \int_0^\infty \frac{1}{\pi s} e^{-|r|^2/s}\, p(s \mid x)\, \mathrm{d}s \overset{(2)}{=} \frac{Q^M}{\pi \Gamma(M)} \int_0^\infty s^{-M-2} e^{-(Q+|r|^2)/s} \mathrm{d}s \\ &\overset{(3)}{=} \frac{Q^M}{\pi \Gamma(M)} \frac{\Gamma(M+1)}{(Q+|r|^2)^{M+1}} \overset{(4)}{=} \frac{1}{\pi \hat{s}} \left(1 + \frac{|r|^2}{M\hat{s}}\right)^{-(M+1)}. \end{aligned} \tag{10.66}$$

(1) Marginalisation over the nuisance s.
(2) Insert (10.65) and collect the powers of s: $s^{-1}s^{-M-1} = s^{-M-2}$.
(3) The same gamma integral with $a = M + 1$ and $Q \to Q + |r|^2$.
(4) $\Gamma(M+1)/\Gamma(M) = M$; divide numerator and denominator by Q^{M+1} and use $Q = M\hat{s}$.

This is a two-dimensional Student-t (with $\nu = 2M$ and $t^2 = 2|r|^2/\hat{s}$ it reads $(1 + t^2/\nu)^{-(\nu+2)/2}$) [201]. As $M \to \infty$, $(1 + x/M)^{-(M+1)} \to e^{-x}$ with $x = |r|^2/\hat{s}$: the Gaussian likelihood with the estimate is recovered, because the estimate has become exact. For finite M the tails are heavier. A residual that is large compared with $\hat{s}$ is no longer catastrophic, because it may mean that the noise in that bin was underestimated.

Example: what the plug-in costs. The script estimates the amplitude of the known chirp ($\rho = 5$, true amplitude one) in 2000 experiments, each with its own M off-source stretches, in two ways: the Gaussian likelihood with $\hat{S}_n$ plugged in (maximum-likelihood estimate and its curvature error) and the Student-t likelihood (10.66) summed over bins (posterior on a grid). The amplitude is in units of the true one, so the ideal error is $1/\rho$; we quote widths and scatters times ρ.

	Gaussian, $\hat{S}_n$ plugged in			Student-t, S_n marginalised		
M	claimed width	actual scatter	coverage	claimed width	actual scatter	coverage
2	0.709	2.367	27.2%	1.001	1.226	58.65%
4	0.8662	1.253	49.95%	1.001	1.122	62.25%
8	0.9356	1.087	60.3%	1.001	1.065	64.8%
16	0.9682	1.043	64.4%	1.001	1.036	66.4%
32	0.9843	1.028	66.4%	1.001	1.026	67.35%

Read the plug-in columns first. The claimed width is too small, by a factor close to $\sqrt{(M-1)/M}$ (0.707 for $M = 2$). The reason is Jensen's inequality: $\hat{s}$ is unbiased, but $\mathbb{E}[1/\hat{s}] = \frac{M}{M-1}\frac{1}{s}$ is not, so $(h|h)$ computed with $\hat{S}_n$ is too large on average. Worse, the actual scatter of the estimate is larger than the ideal one, because a bin whose noise was underestimated by chance receives too much weight. Both effects push the coverage of the nominal 68.3% interval down to 27.2% for $M = 2$. The Student-t likelihood gives a more robust estimate (its scatter is half as large at $M = 2$) and an interval closer to honest. What it cannot do is recover information that was never measured: with only two off-source stretches per bin, thousands of separately uncertain noise levels still cost coverage (the Monte Carlo error of each coverage is ±1.043%). Both methods converge for $M \gtrsim 16$ (Figure 10.23). Research analyses therefore estimate the spectrum from long stretches, or describe it by a smooth model with few parameters fitted together with the signal [202, 203].

Watch out: A plugged-in spectrum is an overconfident spectrum

An unbiased estimate of S_n gives a biased estimate of $1/S_n$, the quantity the likelihood actually uses. With few averages the plug-in Gaussian likelihood claims error bars too small by $\sqrt{(M-1)/M}$ and lets badly estimated bins dominate. Either average many stretches, or marginalise the spectrum.

10.8.7 A signal must have the right shape: the χ^2 consistency test

A large filter output does not prove a gravitational wave. A short burst of instrumental noise (a glitch) can also correlate with the template and give a large z. What a glitch lacks is the

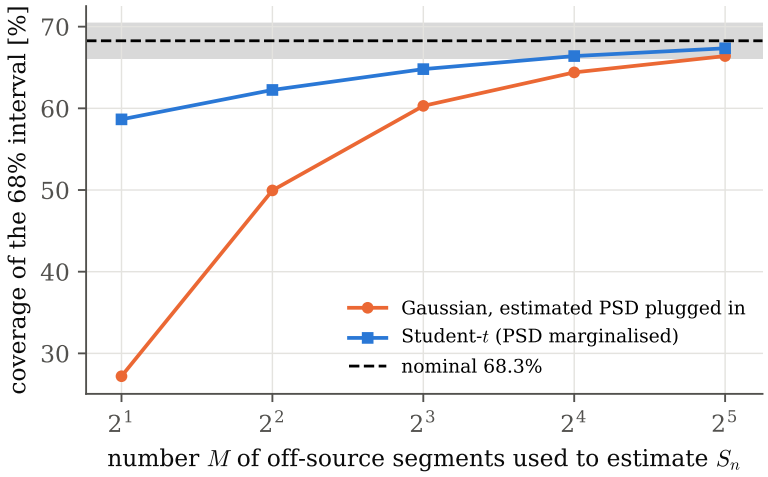


Figure 10.23: Coverage of the nominal 68.3% interval on the signal amplitude when S_n is estimated from M off-source stretches, over 2000 experiments (grey band: ±2 Monte Carlo standard errors). Plugging the estimate into the Gaussian likelihood is overconfident for small M; marginalising the unknown spectrum (Student-t) recovers most of the gap.

shape of the chirp: the template predicts how the signal-to-noise ratio should be shared among frequency bands, and a burst puts its power somewhere else. That prediction is a test with a known distribution.

The question in words. Having found a peak of the filter $x(t_0)$ at some lag, is the way x was built up across frequency what the template predicts, up to noise? Split the band into p pieces with edges chosen so that each carries the same share of the signal-to-noise ratio, $\rho_i^2 = \rho^2/p$ (possible because $4|\tilde{h}|^2/S_n$ is positive). Let x_i be the complex filter (10.55) restricted to piece i, so $\sum_i x_i = x$. For the true signal and noise,

$$\mathbb{E}\, x_i = \frac{\rho^2}{p}\, e^{i\phi_0} = \frac{\mathbb{E}\, x}{p}, \tag{10.67}$$

because the signal's filter output accumulates ρ_i^2 in piece i. So the residuals $r_i = x_i/\rho - x/(p\rho)$ have zero mean when the data contain the template (or only noise). Their noise is Gaussian: by (10.43) the real and imaginary parts of x_i/ρ each have variance $\rho_i^2/\rho^2 = 1/p$. The $2p$ real numbers $\mathrm{Re}\, r_i$, $\mathrm{Im}\, r_i$ therefore have variance $1/p$ each, and two linear constraints, $\sum_i \mathrm{Re}\, r_i = 0$ and $\sum_i \mathrm{Im}\, r_i = 0$ (the sum of the residuals is zero by construction), so

$$\chi^2 \equiv p \sum_{i=1}^{p} |r_i|^2 \sim \chi^2_{2p-2}. \tag{10.68}$$

A sum of squares of $2p$ unit Gaussians would be χ^2_{2p}; each linear constraint removes one degree of freedom (the same count as the $n-1$ of the sample variance in Chapter 2). A glitch correlates with the template mainly in the band where it has power, so one x_i is far from x/p and χ^2 is large.

What the simulation shows. The script (`16_whiten_chi2.py`) divides the chirp band into $p = 8$ pieces ($2p - 2 = 14$ degrees of freedom), finds the lag of the largest $|x|$ and evaluates (10.68) there. For noise alone the mean is 12.89, a little below 14 because the lag was chosen to be the one where the pieces happen to line up; for a true chirp of $\rho = 10$ it is 13.51 and the fraction below the 0.1% threshold 36.12 is 100%. Now take a short burst at 60 Hz, placed at the moment the chirp passes that frequency, and scaled to the same optimal signal-to-noise

ratio as the chirp, $\rho = 10$. The template matches it only in part: the median $|x|/\rho$ is 5.119, well above the 4.222 of noise alone but below the 9.983 of the chirp. Many such bursts would pass a threshold on the filter output alone. Its median χ^2 is 62.65, and only 15.83% of the bursts fall below the threshold (Figure 10.24, left). The test does not make the matched filter better; it rejects the large values of z that come from something other than the template, and detection pipelines weight their statistic by it.

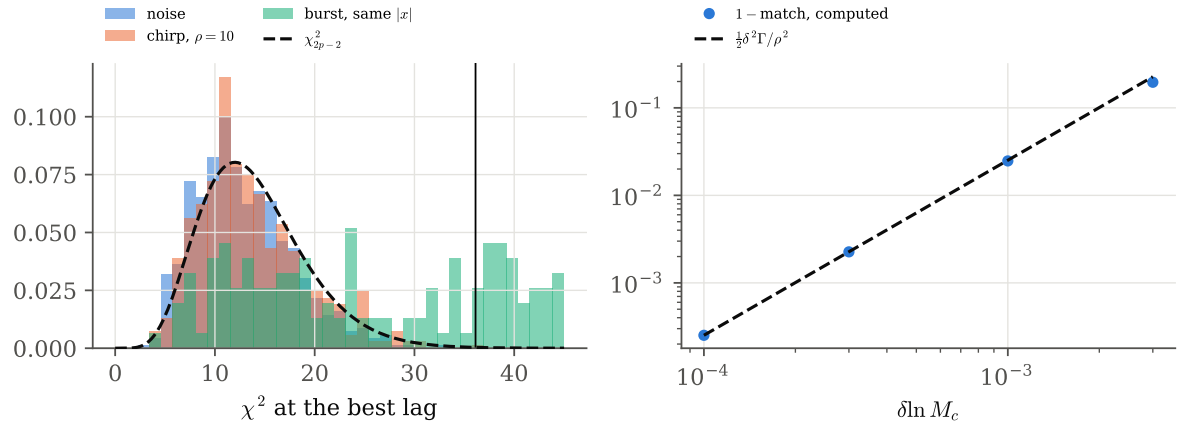


Figure 10.24: Left: the χ^2 statistic of (10.68) at the best lag, for noise alone, for a chirp of $\rho = 10$ (both follow χ^2_{14}, dashed) and for a narrow-band burst with a similar filter output, which does not. The vertical line is the 0.1% threshold. Right: one minus the match between two chirps whose chirp masses (the mass combination of Eq. (T2.42) that sets the sweep) differ by $\delta \ln M_c$, against the Fisher expansion of Section 10.9.5.

What this buys the rest of the chapter. Once a signal is detected, the same likelihood (10.41), now read as a function of $\boldsymbol{\theta}$, is what parameter estimation explores. Its peak is the best fit and its width is the error. The expected curvature of its logarithm at the true parameters is the Fisher matrix of Section 10.1, and for this likelihood it takes a particularly transparent form.

10.9 The Fisher matrix of a chirp

Suppose LISA has detected the inspiral. The likelihood (10.41) of Section 10.8, $p(\boldsymbol{d}\,|\,\boldsymbol{\theta}) \propto e^{-(d-h(\boldsymbol{\theta})|d-h(\boldsymbol{\theta}))/2}$, now becomes a function on parameter space, and its peak sits close to the true masses, arrival time and phase (the chirp mass, the coalescence time t_c and the coalescence phase ϕ_c of the chirp, Sections T2.4 and T2.4.3). How sharp will that peak be? That is what a mission designer, or a theorist asking whether a dark-matter spike could ever be seen, needs before the data exist. In Section 10.1 the answer was the Fisher matrix, the expected curvature of $-\ln\mathcal{L}$ at the true parameters. For the gravitational-wave likelihood it turns into a short, transparent formula, and that formula explains in one picture why some parameters are measured to a part in ten million and others to tens of per cent.

10.9.1 From the likelihood to $\Gamma_{ij} = (\partial_i h|\partial_j h)$

The question in words. If the data were generated by the true parameters $\boldsymbol{\theta}_0$ with a typical noise realisation, how quickly would $\ln\mathcal{L}$ fall as we move away from $\boldsymbol{\theta}_0$? Differentiate $\ln\mathcal{L} = -\frac{1}{2}(d-h|d-h)$ twice, using that the inner product is bilinear and symmetric, and write

$\partial_i \equiv \partial/\partial\theta_i$:

$$-\partial_i\partial_j \ln\mathcal{L} \overset{(1)}{=} -\partial_j\big(d-h\,\big|\,\partial_i h\big) \overset{(2)}{=} \big(\partial_j h\,\big|\,\partial_i h\big) - \big(d-h\,\big|\,\partial_i\partial_j h\big),$$
$$\Gamma_{ij} \equiv \mathbb{E}\left[-\partial_i\partial_j \ln\mathcal{L}\right]_{\theta_0} \overset{(3)}{=} \big(\partial_i h\,\big|\,\partial_j h\big) - \mathbb{E}\,\big(n\,\big|\,\partial_i\partial_j h\big) \overset{(4)}{=} \big(\partial_i h\,\big|\,\partial_j h\big). \tag{10.69}$$

(1) $\partial_i(d-h|d-h) = -2(d-h|\partial_i h)$, because only h depends on $\boldsymbol{\theta}$.
(2) Product rule: the derivative hits either $d-h$ (giving $-\partial_j h$) or $\partial_i h$.
(3) At the true parameters $d - h(\boldsymbol{\theta}_0) = n$, the noise.
(4) $(n|a)$ has mean zero for any fixed signal a (10.43).

This is the **Fisher information matrix** of gravitational-wave astronomy, $\Gamma_{ij} = (\partial h/\partial\theta_i \,|\, \partial h/\partial\theta_j)$ [87, eq. (10)], [84, eq. (41)], written Γ by tradition where Section 10.1 wrote $\mathbf{F}$. Its other face, the covariance of the score, gives the same: the score is $S_i = \partial_i \ln\mathcal{L} = (n|\partial_i h)$ at the truth, and by (10.43) $\mathbb{E}[S_i S_j] = (\partial_i h|\partial_j h)$.

It is also the Gaussian formula (10.20) of Section 10.3 in disguise. Take as data vector the real and imaginary parts of all Fourier modes. Their means are $\operatorname{Re}\tilde{h}_k$ and $\operatorname{Im}\tilde{h}_k$, their covariance is diagonal with entries $\sigma_k^2/2 = \frac{T}{4}S_n(f_k)$ (10.39), with S_n the detector's noise spectrum (for LISA Eq. (T2.77)), and the covariance does not depend on $\boldsymbol{\theta}$. So the trace term vanishes and the mean term is $\sum_k[\operatorname{Re}\partial_i\tilde{h}_k \operatorname{Re}\partial_j\tilde{h}_k + \operatorname{Im}\partial_i\tilde{h}_k \operatorname{Im}\partial_j\tilde{h}_k]/(\sigma_k^2/2) = 4\Delta f \sum_k \operatorname{Re}(\partial_i\tilde{h}_k\,\partial_j\tilde{h}_k^*)/S_n(f_k)$, which is $(\partial_i h|\partial_j h)$. All information sits in the mean channel, because the detector noise does not care what the source is. Then, exactly as before, $\boldsymbol{\Sigma} = \Gamma^{-1}$ is the forecast covariance and $\sqrt{\Sigma_{ii}}$ the marginalised error of θ_i [87, eq. (11)].

10.9.2 The signal model and its derivatives

The waveform is the stationary-phase chirp derived in Section T2.5, $\tilde{h}(f) = \mathcal{A}f^{-7/6}e^{-i\Psi(f)}$: the Fourier transform of the inspiral, in which each frequency is built by the short stretch of the chirp that passes through it, with amplitude $\mathcal{A}$ (T2.56) and phase Ψ. We carry the phase to the order at which the mass ratio first appears. Cutler and Flanagan's restricted post-Newtonian (PN) phase (an expansion in powers of $(v/c)^2$, with v the orbital speed, Section T2.6) [87, eq. (42)] is, in our sign convention (the footnote to Eq. (T2.51) explains why theirs has $e^{+i\Psi}$),

$$\Psi(f) = 2\pi f t_c - \phi_c - \frac{\pi}{4} + \frac{3}{128}\,u^{-5/3}\left[1 + \frac{20}{9}\left(\frac{743}{336}+\frac{11}{4}\eta\right)x - 16\pi x^{3/2}\right],$$
$$u = \frac{\pi G\mathcal{M}f}{c^3}, \qquad x = u^{2/3}\eta^{-2/5}. \tag{10.70}$$

The Newtonian term is Eq. (T2.54): t_c and ϕ_c are the time and phase at which the binary would coalesce (Section T2.4.3), and $\mathcal{M}$ is the chirp mass, the one combination of the two masses that sets the Newtonian sweep (T2.42). The bracket adds the first and the 1.5PN corrections, derived term by term in Section T2.6 as (T2.67), with u the reduced frequency of (T2.58). Here $x = (\pi GMf/c^3)^{2/3}$ is $(v/c)^2$ with the total mass $M = \mathcal{M}\eta^{-3/5}$, and $\eta = m_1m_2/M^2$ is the symmetric mass ratio ([$\eta = 1/4$ for equal masses, $\eta \approx m_2/m_1$ when one is much lighter]). The parameters are $\boldsymbol{\theta} = (\ln\mathcal{M}, \ln\eta, t_c, \phi_c, \ln\mathcal{A})$; logarithms make the errors fractional.

The derivatives are short because every parameter but one sits in the phase. With $\tilde{h} = \mathcal{A}f^{-7/6}e^{-i\Psi}$,

$$\frac{\partial\tilde{h}}{\partial\ln\mathcal{A}} = \tilde{h}, \qquad \frac{\partial\tilde{h}}{\partial\theta} = -i\,\frac{\partial\Psi}{\partial\theta}\,\tilde{h} \quad \text{for } \theta \in \{\ln\mathcal{M}, \ln\eta, t_c, \phi_c\}, \tag{10.71}$$

with $\partial\Psi/\partial t_c = 2\pi f$, $\partial\Psi/\partial\phi_c = -1$, and the mass derivatives obtained term by term from (10.70): a term $u^{-5/3}x^k$ scales as $\mathcal{M}^{-5/3+2k/3}\eta^{-2k/5}$, so its logarithmic derivatives are $(-\frac{5}{3}+\frac{2k}{3})$

and $-\frac{2k}{5}$ times itself, and the explicit η in the 1PN coefficient adds $\frac{20}{9}\frac{11}{4}\eta\,x$. The library does exactly this:

Listing 10.7: Fisher matrix from the logarithmic derivatives of $\tilde{h}$ (`code/ch10/lib_gw10.py`)

```
169 def fisher_from_dlogh(f, amp2, Sn, D):
170     """D: array (P, Nf) of d ln h / d theta_i (complex). Returns the P x P Fisher matrix."""
171     w = 4.0 * amp2 / Sn
172     P = D.shape[0]
173     F = np.empty((P, P))
174     for i in range(P):
175         for j in range(i, P):
176             F[i, j] = F[j, i] = integ(f, w * np.real(D[i] * np.conj(D[j])))
177     return F
178
179
180 def phase_derivs(f, Mc, eta, b=None, order=1.5):
181     """d ln h/d theta for theta = (ln Mc, ln eta, tc, phic, ln A [, beta]) at beta = 0:
182     d ln h = -i dPsi for phase parameters and 1 for ln A; beta enters Psi as beta u^b."""
183     dM, de = dpsi_pn(f, Mc, eta, order)
184     rows = [-1j * dM, -1j * de, -1j * 2 * np.pi * f, -1j * (-1.0) + 0 * f, 1.0 + 0j * f]
185     if b is not None:
186         rows.append(-1j * u_of(f, Mc) ** b)
187     return np.array(rows)
```

Line 176 is (10.69) written with $D_i = \partial_i \ln\tilde{h}$: $(\partial_i h|\partial_j h) = 4\,\mathrm{Re}\int |\tilde{h}|^2 D_i D_j^*/S_n\,\mathrm{d}f$. Lines 183–187 are (10.71); the optional last row is the environmental term of Section 10.10. The integrals use a logarithmic frequency grid, $\int g\,\mathrm{d}f = \int gf\,\mathrm{d}\ln f$, because the integrand varies as a power of f.

10.9.3 Amplitude and phase separate; the phase block is a weighted average

Insert (10.71) into (10.69). Three things happen.

$$\Gamma_{\ln\mathcal{A}\,\ln\mathcal{A}} \overset{(1)}{=} (h|h) = \rho^2,$$
$$\Gamma_{\ln\mathcal{A}\,\theta} \overset{(2)}{=} 4\,\mathrm{Re}\int \frac{\tilde{h}\,\big(-i\partial_\theta\Psi\,\tilde{h}\big)^*}{S_n}\mathrm{d}f = 4\,\mathrm{Re}\Big[\,i\int \frac{|\tilde{h}|^2\partial_\theta\Psi}{S_n}\mathrm{d}f\Big] \overset{(3)}{=} 0,$$
$$\Gamma_{\theta\varphi} \overset{(4)}{=} 4\int \frac{|\tilde{h}|^2}{S_n}\,\partial_\theta\Psi\,\partial_\varphi\Psi\,\mathrm{d}f \overset{(5)}{=} \rho^2\,\big\langle \partial_\theta\Psi\,\partial_\varphi\Psi\big\rangle_w, \qquad w(f) \equiv \frac{4|\tilde{h}(f)|^2}{\rho^2 S_n(f)}. \tag{10.72}$$

(1) Definition of the signal-to-noise ratio.
(2) $\partial\tilde{h}/\partial\ln\mathcal{A} = \tilde{h}$, and $(-i)^* = i$.
(3) The integral is real, so i times it is purely imaginary, and its real part is zero.
(4) $(-i)(-i)^* = 1$: the phases cancel and only $|\tilde{h}|^2$ remains.
(5) $w(f)$ is positive and $\int w\,\mathrm{d}f = 1$, so $\langle g\rangle_w = \int wg\,\mathrm{d}f$ is an average over frequency, weighted by where the signal-to-noise ratio accumulates.

So the amplitude decouples from everything in the phase: $\Delta\mathcal{A}/\mathcal{A} = 1/\rho$ whatever else is unknown, the result Cutler and Flanagan state below their eq. (43). The distance is measured with an error of order $1/\rho$, a few per cent at best. The phase parameters are a different matter: their Fisher matrix is ρ^2 times the matrix of weighted averages of products of the phase derivatives. This is the formula to keep in mind.

What the marginalised error is. Single out one phase parameter β with $g(f) = \partial\Psi/\partial\beta$, and collect the derivatives with respect to the others into a vector $\boldsymbol{e}(f)$. By the Schur complement of

Section 10.4.1, the marginalised variance of β is

$$\frac{1}{\Sigma_{\beta\beta}} \overset{(1)}{=} \Gamma_{\beta\beta} - \Gamma_{\beta e}\Gamma_{ee}^{-1}\Gamma_{e\beta} \overset{(2)}{=} \rho^2\Big[\langle g^2\rangle_w - \langle g\,\boldsymbol{e}\rangle_w^{\mathsf{T}}\langle \boldsymbol{e}\boldsymbol{e}^{\mathsf{T}}\rangle_w^{-1}\langle \boldsymbol{e}\,g\rangle_w\Big] \overset{(3)}{=} \rho^2 \min_{\boldsymbol{c}} \big\langle (g - \boldsymbol{c}\cdot\boldsymbol{e})^2\big\rangle_w. \tag{10.73}$$

(1) The inverse of the $\beta\beta$ entry of Γ^{-1} is the Schur complement of the other parameters (Eq. (10.23)).
(2) Each block is ρ^2 times a weighted average, by (10.72).
(3) Minimise $Q(\boldsymbol{c}) = \langle (g - \boldsymbol{c}\cdot\boldsymbol{e})^2\rangle_w$: setting $\partial Q/\partial \boldsymbol{c} = 0$ gives the normal equations $\langle \boldsymbol{e}\boldsymbol{e}^{\mathsf{T}}\rangle \boldsymbol{c} = \langle \boldsymbol{e}\,g\rangle$ of weighted least squares (Chapter 3), and putting that $\boldsymbol{c}$ back gives $\langle g^2\rangle - \langle g\,\boldsymbol{e}\rangle^{\mathsf{T}}\langle \boldsymbol{e}\boldsymbol{e}^{\mathsf{T}}\rangle^{-1}\langle \boldsymbol{e}\,g\rangle$.

In words: fit the phase signature $g(f)$ of β by the best weighted combination of the other parameters' signatures. Whatever cannot be fitted, the residual $g_\perp = g - \boldsymbol{c}\cdot\boldsymbol{e}$, is what measures β, and

$$\sigma_\beta = \frac{1}{\rho\sqrt{\langle g_\perp^2\rangle_w}}. \tag{10.74}$$

A shift of β by σ_β therefore changes the phase, after the other parameters have absorbed all they can, by a weighted root-mean-square of exactly $1/\rho$ radians. Figure 10.25 draws this.

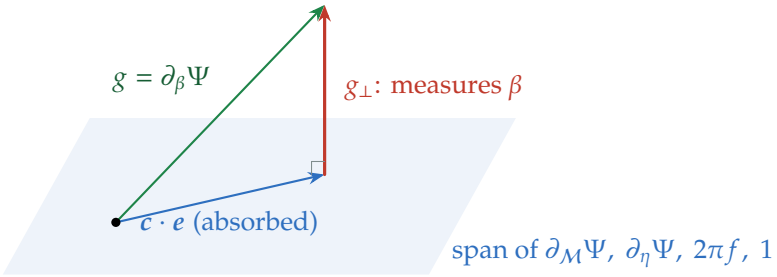


Figure 10.25: The marginalised error as a projection. Functions of frequency form a vector space with the weighted inner product $\langle a\,b\rangle_w$. The phase signature g of a parameter is split into the part that the other parameters can imitate (its projection on their span) and the residual $g_\perp$. Only $g_\perp$ carries information about β: $\sigma_\beta = 1/(\rho\,\|g_\perp\|_w)$. A signature nearly inside the plane is almost unmeasurable, however large it is.

Key idea: What makes a phase parameter measurable

For a chirp in Gaussian noise, $\Gamma_{\theta\varphi} = \rho^2\langle \partial_\theta\Psi\,\partial_\varphi\Psi\rangle_w$ with the weight $w = 4|\tilde h|^2/(\rho^2 S_n)$, and the amplitude decouples. A parameter is measured to $\sigma = 1/(\rho\|g_\perp\|_w)$, where $g_\perp$ is the part of its phase signature that no combination of the other parameters can imitate over the band where the signal-to-noise ratio accumulates.

10.9.4 Examples: timing, bandwidth, and a classic table

Example: the arrival time alone. If only t_c were unknown, $g = 2\pi f$ and nothing absorbs it, so $\sigma_{t_c} = 1/(2\pi\rho\sqrt{\langle f^2\rangle_w})$: the timing error is the period at the root-mean-square frequency of the signal, divided by $2\pi\rho$. For a $1.4 + 1.4\,M_\odot$ binary in the noise of [87, eq. (4)] between 10 Hz and the ISCO (the end of the inspiral, Eq. (T2.46)), the script `09_cf94.py` finds the weight of Figure 10.26, peaked near 55.35 Hz, with $\langle f\rangle_w = 64.26$ Hz and spread $\sigma_f = 44.14$ Hz. At $\rho = 10$, $\sigma_{t_c} = 0.2041$ ms.

Example: arrival time and phase. Now let ϕ_c be unknown too. Its signature is the constant -1, and the best constant fit to $2\pi f$ is $2\pi\langle f\rangle_w$, so $g_\perp = 2\pi(f - \langle f\rangle_w)$ and

$$\sigma_{t_c} = \frac{1}{2\pi\rho\,\sigma_f}, \qquad \sigma_f^2 = \langle f^2\rangle_w - \langle f\rangle_w^2. \tag{10.75}$$

A constant phase offset can imitate the mean frequency of a time shift, but not its spread, so what is left to measure the arrival time is the spread of frequencies in the signal, the **effective bandwidth** σ_f. This is the familiar radar result that timing needs bandwidth. The numbers give 0.3605 ms, and the 2×2 Fisher block gives 0.3605 ms, the same. With the two masses free as well, part of $2\pi f$ is absorbed by the mass terms, and the full matrix gives 0.713 ms (below). Each new nuisance parameter can only remove more of the signature, never add to it.

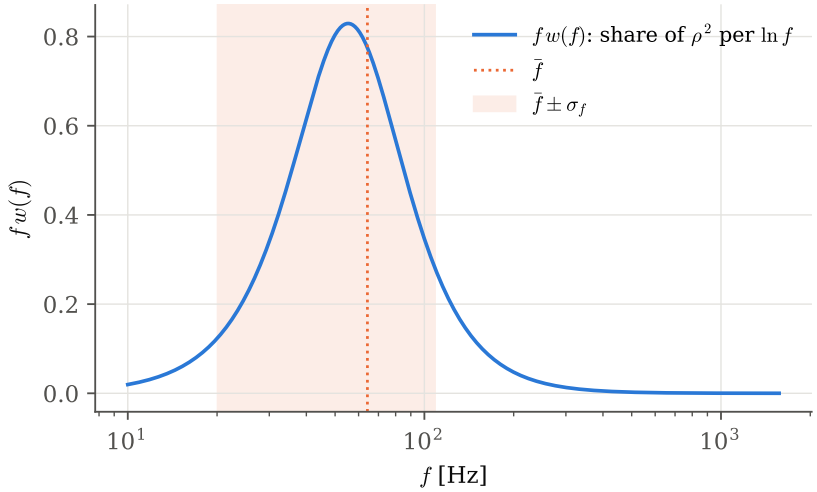


Figure 10.26: Where a $1.4 + 1.4\,M_\odot$ inspiral accumulates its signal-to-noise ratio in the advanced-detector noise of Cutler and Flanagan: $f\,w(f)$ is the share of ρ^2 per unit $\ln f$. The weight is a few tens of hertz wide; its mean (dotted) and spread (band) set the timing errors with and without an unknown phase.

Example: Cutler and Flanagan's Table I. The full five-parameter matrix is a test of the whole machinery, because Cutler and Flanagan published its result for five mass pairs [87, Table I], normalised to $\rho = 10$. They quote the chirp mass and the reduced mass $\mu = \eta M = \mathcal{M}\eta^{2/5}$, so we convert our $(\ln\mathcal{M}, \ln\eta)$ covariance with the Jacobian of $\ln\mu = \ln\mathcal{M} + \frac{2}{5}\ln\eta$ (Section 10.4.5). The comparison:

$m_1 + m_2$ [$M_\odot$]	$\Delta\phi_c$	Δt_c [ms]	$\Delta\mathcal{M}/\mathcal{M}$ [%]	$\Delta\mu/\mu$ [%]	$c_{\mathcal{M}\mu}$
2.0 + 1.0 paper / ours	1.31 / 1.306	0.721 / 0.7212	0.0038 / 0.003859	0.39 / 0.3951	0.899 / 0.899
1.4 + 1.4	1.28 / 1.282	0.713 / 0.7128	0.0040 / 0.004	0.41 / 0.4083	0.906 / 0.906
10 + 1.4	1.63 / 1.635	1.01 / 1.013	0.020 / 0.02036	0.54 / 0.5434	0.927 / 0.927
15 + 5	2.02 / 2.019	1.44 / 1.441	0.113 / 0.1137	1.5 / 1.519	0.954 / 0.9545
10 + 10	1.98 / 1.986	1.43 / 1.43	0.16 / 0.1566	1.9 / 1.944	0.958 / 0.9583

Every entry agrees to within 2.33%, which is the rounding of the published table and of their crude ISCO cut-off. Three features are the projection picture at work. The chirp mass is measured a hundred times better than the reduced mass, because $u^{-5/3}$ dominates the phase at low frequency while the mass ratio enters only through the smaller 1PN and 1.5PN terms. The two mass errors are strongly correlated ($c_{\mathcal{M}\mu} > 0.9$), because the 1PN signature of η is partly

imitated by the Newtonian signature of $\mathcal{M}$. And all errors grow with the total mass, because a heavier binary leaves the band at a lower ISCO frequency and spends fewer cycles in it.

Cutler and Flanagan state the noise density $p(n) \propto e^{-(n|n)/2}$, the Fisher matrix and its inverse in three equations [87, Sec. II B]. Between them lie the variance of a Fourier mode (Eq. (10.39)), the density of one realisation (Eq. (10.40)), the expected curvature of the log-likelihood (Eq. (10.69)) and the weighted-average form (Eq. (10.72)), which turns their table into a statement about projections of phase signatures.

Watch out: Normalised tables hide the distance

Forecast tables are often quoted at a fixed ρ (here 10; Eda et al. use 10, Barack and Cutler 30 [84]). Every Fisher error scales as $1/\rho$, so a table at $\rho = 10$ says nothing about whether a given source reaches $\rho = 10$. Always compute ρ for the source at its distance first, then rescale.

10.9.5 The Fisher matrix as a distance between waveforms

There is a second way to read Γ_{ij}, and it explains how searches are built. If the true wave has parameters $\boldsymbol{\theta}$ and the template has $\boldsymbol{\theta} + \Delta\boldsymbol{\theta}$, how much of the signal-to-noise ratio does the template recover? The **match** M is the normalised overlap

$$\mathsf{M}(h, h') = \frac{(h|h')}{\sqrt{(h|h)(h'|h')}} \le 1, \tag{10.76}$$

the cosine of the angle between the two waveforms in the space with inner product $(\cdot|\cdot)$; the bound is the Cauchy–Schwarz inequality, with equality when $h' \propto h$. A filter built with h' and run on a signal h has mean $\mathsf{M}\rho$ instead of ρ (by (10.44)), so $1 - \mathsf{M}$ is the fractional loss of signal-to-noise ratio. For a small change write the unit-norm waveform $\hat{h} = h/\rho$, so that $\mathsf{M} = (\hat{h}|\hat{h}')$. For two unit vectors $(\hat{h} - \hat{h}'|\hat{h} - \hat{h}') = 2 - 2\mathsf{M}$, and for phase parameters (which leave ρ unchanged, Eq. (10.72)) $\hat{h}' - \hat{h} \simeq \Delta\theta^i\, \partial_i h/\rho$. Therefore

$$1 - \mathsf{M} \overset{(1)}{=} \tfrac{1}{2}\big(\hat{h}' - \hat{h}\,\big|\,\hat{h}' - \hat{h}\big) \overset{(2)}{=} \frac{1}{2\rho^2}\,\Delta\theta^i\Delta\theta^j\,(\partial_i h|\partial_j h) \overset{(3)}{=} \frac{\Delta\boldsymbol{\theta}^{\mathsf{T}}\Gamma\,\Delta\boldsymbol{\theta}}{2\rho^2}. \tag{10.77}$$

(1) Expand the square with $(\hat{h}|\hat{h}) = (\hat{h}'|\hat{h}') = 1$.
(2) Insert $\hat{h}' - \hat{h} \simeq \Delta\theta^i\partial_i h/\rho$ and use bilinearity.
(3) The definition (10.69).

So Γ_{ij}/ρ^2 is the *metric* that measures how different two waveforms are, and the Fisher 1σ ellipsoid $\Delta\boldsymbol{\theta}^{\mathsf{T}}\Gamma\Delta\boldsymbol{\theta} = 1$ is the set of waveforms with $\rho^2(1 - \mathsf{M}) = \frac{1}{2}$: the template loses exactly the amount of log-likelihood that defines one standard deviation. A search bank is laid out by the same metric: the templates are spaced so that the worst match to any possible signal stays above a chosen value (typically 0.97), which is why high-ρ errors and bank densities are both set by Γ. The script (`16_whiten_chi2.py`) checks the expansion for a change of the chirp mass alone: for $\delta\ln\mathcal{M} = 10^{-4}, 3\times10^{-4}, 10^{-3}, 3\times10^{-3}$ the computed $1 - \mathsf{M}$ is 2.518×10^{-4}, 0.002262, 0.02475, 0.1961, against 2.519×10^{-4}, 0.002267, 0.02519 and 0.2267 from (10.77) with $\Gamma_{\ln\mathcal{M}\ln\mathcal{M}} = 1.259\times10^{6}$ (right panel of Figure 10.24). The two agree to a part in 10^3 for the smallest changes and separate once the phase has moved by a radian (where the surface of waveforms has curved away from its tangent plane, the content of Section 10.13.1).

10.10 Environments as power laws in the phase

A binary in vacuum loses energy only to gravitational waves. A binary inside a dark-matter spike, an accretion disc or a gas cloud loses some more: to the wake that drags behind the companion, to the matter it swallows, to the torque of the disc. Each of these makes the binary sweep through frequency a little faster, and by Eq. (T2.53) anything that changes the sweep rate changes the Fourier phase. The projection picture of Figure 10.25 tells us that whether LISA can see such a change depends on one thing only: the shape of the phase change as a function of frequency, and how well the vacuum parameters can imitate it. So the question is:

what shape does an environmental energy loss give to the phase, and how does the shape differ from one environment to the next?

The answer, derived in Section T2.10 for dynamical friction and in general form in Section T2.6, is that at leading order every smooth environment adds a single power of frequency, with an exponent fixed by its physics. That is what turns many separate dark-matter and astrophysics problems into one statistical problem with a single parameter, the exponent.

10.10.1 From an extra power to a power law in the phase

What the physics supplies. The derivation belongs to the theory and is done in Section T2.10; here we recall its result in the form the forecast needs. The orbit is quasi-circular and shrinks slowly (Section T2.4); the environment removes a power $P_{\rm env}$ on top of the gravitational-wave power $P_{\rm GW}$, with a small ratio that varies as a power of the radius, $\epsilon \equiv P_{\rm env}/P_{\rm GW} \propto r^p \ll 1$. Energy balance with the vacuum binding energy $E(f)$ of Eq. (T2.41) then speeds up the sweep by the factor $1+\epsilon$ (Eq. (T2.91)), and Kepler's law $r \propto f^{-2/3}$ turns the power of r into a power of f:

$$\dot f = \dot f_{\rm V}\,(1+\epsilon), \qquad \epsilon = c_f\, f^{-n}, \quad n = \frac{2p}{3}, \tag{10.78}$$

where $\dot f_{\rm V} = K f^{11/3}$, $K = \frac{96}{5}\pi^{8/3}(G\mathcal{M}/c^3)^{5/3}$, is the vacuum chirp rate (T2.43) and $\mathcal{M}$ the chirp mass. Because the curvature of the Fourier phase is $2\pi/\dot f$ (Eq. (T2.53): the slope of Ψ is the arrival time of each frequency, so a faster sweep bends it), integrating the change of curvature twice gives the environmental part of the phase (Eq. (T2.94)):

$$\delta\Psi(f) = -\frac{2\pi c_f}{K}\,\frac{f^{-n-5/3}}{(n+\frac{8}{3})(n+\frac{5}{3})} \overset{(1)}{=} -\frac{2\pi f^2}{\dot f_{\rm V}(f)}\,\frac{\epsilon(f)}{(n+\frac{8}{3})(n+\frac{5}{3})} \overset{(2)}{=} \beta\, u^b, \qquad b = -\frac{5}{3} - n. \tag{10.79}$$

(1) Rewrite $c_f f^{-n-5/3}/K = f^2\epsilon(f)/\dot f_{\rm V}(f)$.

(2) With the reduced frequency $u = \pi G\mathcal{M} f/c^3$ of (T2.58), every power of f is a power of u; the constants are collected in β. The two integration constants, $a_0 + a_1 f$, are a shift of ϕ_c and t_c and are absorbed by those parameters.

The middle form is the one the analysis uses: it gives the phase shift at any frequency directly from the ratio ϵ there. The factor $2\pi f^2/\dot f_{\rm V}$ is 2π times the number of cycles the binary spends near f (per unit $\ln f$), so in words the phase shift is "the fractional speed-up times the number of radians spent there, divided by a number of order ten". The last form is the **parametrised post-Einsteinian** (ppE) phase correction of Yunes and Pretorius [91, eq. (1)]: β carries the strength of the effect and b its physics (Eq. (T2.69), where the same rule is derived for any correction that scales as a power of the orbital speed). In post-Newtonian language, the expansion in powers of $(v/c)^2$ of Section T2.6, a relative correction $\epsilon \propto f^{-n} = (v/c)^{-3n}$ enters at order $-\frac{3}{2}n$ PN.

10.10.2 The exponents of the common environments

Each exponent follows from how $P_{\rm env}/P_{\rm GW}$ scales with r, and $P_{\rm GW} = \frac{32}{5}G^4M^3\mu^2/(c^5r^5)$ (the quadrupole power of a circular binary, Eq. (T2.27)) falls as r^{-5}, steeper than almost anything an environment does. That is why environments matter at large radius, at the low-frequency start of the band.

Dynamical friction in a spike. Chandrasekhar's drag in a density $\rho \propto r^{-\gamma}$ removes $P_{\rm DF} = 4\pi G^2 m_2^2 \rho \ln\Lambda/v$ with $v = \sqrt{GM/r}$ (Eq. (T2.87)). Divided by $P_{\rm GW} \propto r^{-5}$ this is $\propto \rho\, r^{11/2} \propto r^{11/2-\gamma}$, with the constants worked out in Eq. (T2.92) [82, eq. (16)], [88, eq. (16)]. So

$$p = \tfrac{11}{2} - \gamma, \qquad n = \tfrac{11-2\gamma}{3}, \qquad b = \tfrac{2\gamma-16}{3}.$$

For the adiabatic spike $\gamma = 7/3$ this is $b = -34/9 \approx -3.78$, a -3.2PN effect.

Gas accretion onto the central black hole. Suppose the heavy hole grows at a fraction $f_{\rm Edd}$ of the Eddington rate, $\dot M/M = f_{\rm Edd}/t_{\rm S}$ with the Salpeter time $t_{\rm S} \approx 4.5\times10^7$ yr. The companion feels a slowly growing central mass. Its specific angular momentum $\sqrt{GMr}$ is conserved (the force is still central), so $r \propto 1/M$, and Kepler's law $\omega^2 = GM/r^3$ gives $\omega \propto M^2$, hence $\dot f/f = 2\dot M/M$ in addition to the radiative sweep. The fractional speed-up is

$$\epsilon = \frac{2f\,\dot M/M}{\dot f_{\rm V}} \propto \frac{f}{f^{11/3}} = f^{-8/3}, \qquad n = \tfrac{8}{3}, \quad b = -\tfrac{13}{3},$$

a -4PN effect, as for the mass changes discussed by Barausse et al. [92, Sec. XII].

Migration in a thin disc. A companion embedded in a thin accretion disc exchanges angular momentum with it. For type-I migration far from the inner edge, Barausse et al. find $\dot L_{\rm migr}/\dot L_{\rm GW} = 10^{-5} f_{\rm Edd}^{2/5} (M/10^6 M_\odot)^{7/5} (\alpha/0.1)^{-3/5} \tilde r^{7/2}$, with $\tilde r = rc^2/(GM)$ and α the disc viscosity parameter [92, eq. (108)]. On a circular orbit $\dot E = \omega \dot L$ for both torques, so the energy ratio is the same and $p = 7/2$, $n = 7/3$, $b = -4$.

effect	$\epsilon = P_{\rm env}/P_{\rm GW} \propto$	$b = -\frac{5}{3} - n$	PN order
friction, spike $\rho \propto r^{-\gamma}$ [82, 89]	$r^{11/2-\gamma}$	$\frac{2\gamma-16}{3}$ (-3.78 at $\frac{7}{3}$)	$\gamma - \frac{11}{2}$
friction, uniform medium	$r^{11/2}$	$-\frac{16}{3}$	-5.5
collisionless accretion, spike [92]	$r^{9/2-\gamma}$	$\frac{2\gamma-14}{3}$	$\gamma - \frac{9}{2}$
enclosed dark-matter mass (no loss) [82, 88]	$r^{3-\gamma}$	$\frac{2\gamma-11}{3}$	$\gamma - 3$
Eddington accretion onto m_1	$f^{-8/3}$	$-\frac{13}{3}$	-4
type-I migration, thin disc [92]	$r^{7/2}$	-4	-3.5
dipole radiation (Brans–Dicke) [91]	r	$-\frac{7}{3}$	-1
massive graviton (propagation) [91]	—	-1	$+1$
GR: chirp mass, 1PN, 1.5PN, ϕ_c, t_c	—	$-\frac{5}{3}$, -1, $-\frac{2}{3}$, 0, 1	
vector ultralight dark matter [100]	a step at a resonance: not a power law		

Table 10.5: Leading-order exponents of environmental and beyond-GR phase corrections $\delta\Psi = \beta u^b$. The PN order of a relative correction $\epsilon \propto f^{-n}$ is $-\frac{3}{2}n$. The GR exponents in the last block are the signatures every environmental term must be distinguished from.

Table 10.5 collects these. All the dissipative environments sit at $b \lesssim -3$, well below the chirp-mass exponent $-5/3$; the propagation and dipole corrections of modified gravity sit

between $-7/3$ and -1, in the middle of the GR terms. Figure 10.27 shows what that means on the frequency axis.

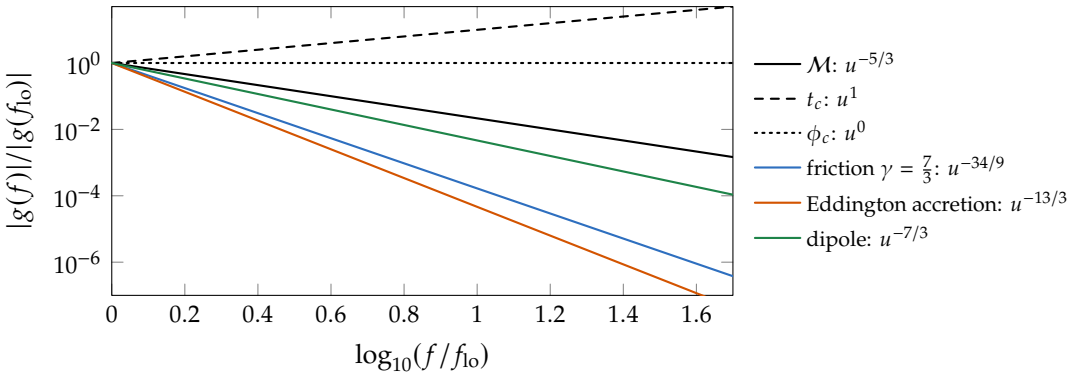


Figure 10.27: Phase signatures over the LISA band of the benchmark binary (a factor 50 in frequency), normalised at the start frequency $f_{\rm lo}$. On these axes each is a straight line of slope b. The environmental signatures fall far more steeply than the chirp-mass one and are concentrated in the first fraction of a decade; a dipole term lies close to the chirp-mass line and is easily imitated by it.

10.10.3 Example: how large is each effect for one source?

Take the benchmark of Section T2.4 that recurs below: a $10^3 + 1.4\,M_\odot$ intermediate mass-ratio inspiral observed for the last five years before its ISCO (the innermost stable circular orbit, where the inspiral ends, Eq. (T2.46)), from $f_{\rm lo} = 19.96\,$mHz up to 1 Hz (its ISCO is at 4.391 Hz, above the LISA band), chirp mass $19.39\,M_\odot$. The script `10_env_phases.py` evaluates $\epsilon(f)$ for each effect with the parameters of the literature (the spike density $\rho_6 = 5.45\times10^{15}M_\odot\,\mathrm{pc}^{-3}$ at $10^{-6}\,$pc of Coogan et al. [89], $\xi = 0.58$, $\ln\Lambda = \ln\sqrt{m_1/m_2}$; $f_{\rm Edd} = 1$ for gas accretion; $f_{\rm Edd} = 0.1$, $\alpha = 0.1$ for the disc) and converts it to $\delta\Psi$ with (10.79). At the start of the band:

effect	b	$\epsilon(f_{\rm lo})$	$\lvert\delta\Psi(f_{\rm lo})\rvert$ [rad]
friction, spike $\gamma = 7/3$	-3.778	0.6913	2.021×10^{6}
friction, spike $\gamma = 3/2$	-4.333	0.01549	35351
collisionless accretion, $\gamma = 7/3$	-3.111	0.006635	27371
enclosed mass, $\gamma = 7/3$ (order of magnitude)	-2.111	4.911×10^{-6}	39.45
Eddington gas accretion onto m_1	-4.333	5.926×10^{-7}	1.353
type-I migration, thin disc	-4	0.03892	1.027×10^{5}

A tiny ϵ can still give a large phase shift, because $2\pi f^2/\dot f_{\rm V}$ is enormous for a binary that spends years near $f_{\rm lo}$: an Eddington-limited growth that changes the sweep rate by six parts in ten million still shifts the phase by more than a radian. Before trusting the closed form, the script compares it with the exact double integral of $2\pi/\dot f - 2\pi/\dot f_{\rm V}$ for a weak spike; after removing the $a_0 + a_1 f$ that t_c and ϕ_c absorb, they differ by 2.285×10^{-6} of the shift.

Figure 10.28 shows the whole band. One number in the table is a warning: for the static $7/3$ spike, $\epsilon(f_{\rm lo}) = 0.6913$, because friction equals radiation at $f_{\rm eq} = 16.75\,$mHz, just below the band (Section T2.10). There the expansion in ϵ is not valid, and the full dephasing Eq. (T2.96) must be used. The power law is a statement about weak environments, which are precisely the ones whose detectability is in question.

The hidden assumption, and a counterexample. "An environment is a power law in the phase" contains an assumption: that the environment changes smoothly and monotonically

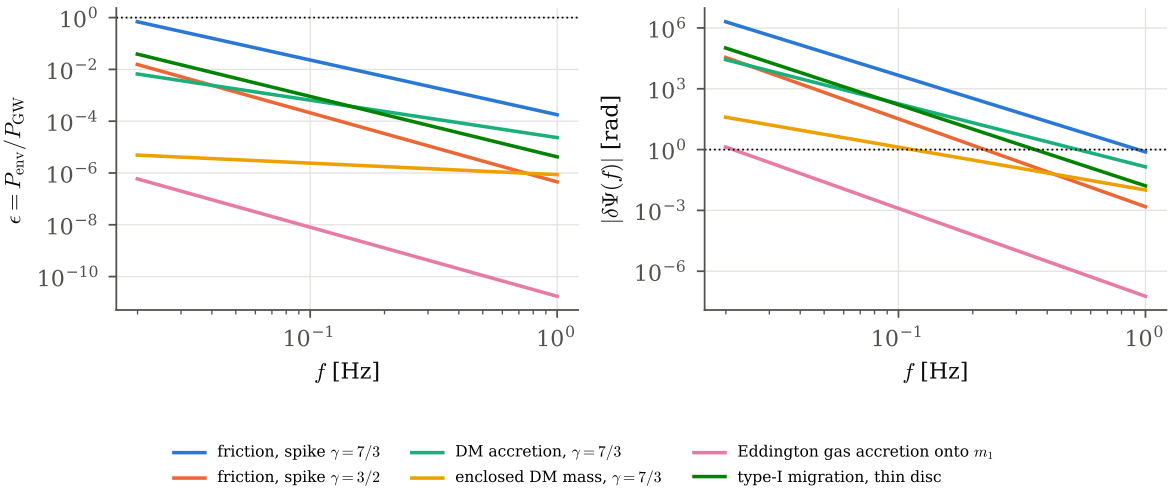


Figure 10.28: The six effects for the $10^3 + 1.4\,M_\odot$ binary over its last five years in the LISA band. Left: the ratio ϵ of environmental to gravitational-wave power; each is a power law falling towards high frequency. Right: the resulting Fourier-phase shift (10.79); the dotted line marks one radian. All shifts are largest at the start of the band.

with radius, so a single power of r describes it across the band. A vector ultralight dark-matter field breaks it. Its oscillating potential does nothing on average until the orbital frequency crosses a resonance with the field, and then shifts the orbit within a narrow band [100]. Chase, López Nacir and Yunes find that the resulting phase correction is a step, and that below the threshold it is partly degenerate with t_c and ϕ_c. In the projection picture this is no surprise: a correction of the form $a_0 + a_1 f$ lies entirely in the span of the ϕ_c and t_c signatures, $g_\perp = 0$, and it is invisible however large it is. Only the sudden switch-on, which no smooth GR term can imitate, can be measured. So the lesson in one line: *a phase correction is measured by its shape, not its size*.

10.11 One forecast for every environment

The two previous sections give us all the pieces. Section 10.9 says that a phase parameter is measured by the part of its signature that the vacuum parameters cannot imitate, and Section 10.10 says that the signature of any smooth environment is a single power, $\delta\Psi = \beta u^b$, with an exponent set by its physics: the ppE form of Eq. (T2.69), with $u = \pi G\mathcal{M}f/c^3$ the reduced frequency, β the strength and b the exponent. So instead of writing one forecast for dark-matter friction, another for accretion and another for a gas disc, we can ask a single question for a given source:

for each exponent b, what is the smallest amplitude β that LISA would detect?

One Fisher matrix per value of b answers it, and the answer is one curve. Every environment then becomes a point, above or below the curve. We build that curve for the benchmark intermediate mass-ratio inspiral and read it.

10.11.1 The noise curve, checked against its source

The forecast weights each frequency by $1/S_n(f)$, the one-sided noise spectrum of Section 10.8.2, so the noise model comes first. We use the sky-averaged LISA sensitivity of Robson, Cornish

and Liu [85, eqs. (10)–(14)]: optical metrology and test-mass acceleration noise, divided by the sky-averaged response of two Michelson channels (the factor $\frac{3}{10}$ of Eq. (T2.37): how strongly a source at a random place on the sky registers), plus the Galactic confusion noise for a four-year mission. Its pieces are derived and explained in Section T2.7; the formula is Eq. (T2.77). The script `11_lisa_robson.py` implements it independently of chapter T2 (the two agree to 4.441×10^{-16}) and checks it against what the paper itself shows and computes.

Their figure. The script renders page 6 of the paper, finds the frame of Figure 4, calibrates its axes and reads the curve of $\sqrt{S_n}$ from the colour of its pixels. Figure 10.29 overlays the digitised points on our curve. Below 30 mHz they agree to within 8.594%, the precision of reading a plotted line; at $f = 1$, 10 and 100 mHz the ratios ours/theirs are $2.880 \times 10^{-19}/2.894 \times 10^{-19}$, $1.201 \times 10^{-20}/1.173 \times 10^{-20}$ and $4.615 \times 10^{-20}/4.019 \times 10^{-20}$. The last is 15% high: above the transfer frequency $f_* = c/(2\pi L) = 19.09$ mHz (their 19.09 mHz), where the wave's phase advances by a radian while light crosses one arm, so the arm starts to average the wave away (Eq. (T2.38)), their plot uses the exact numerical response, which oscillates, while the formula uses the smooth fit $R(f) = \frac{3}{10}/[1 + 0.6(f/f_*)^2]$ of their eq. (9). The wiggles are visible in the digitised points; the fit runs through their average.

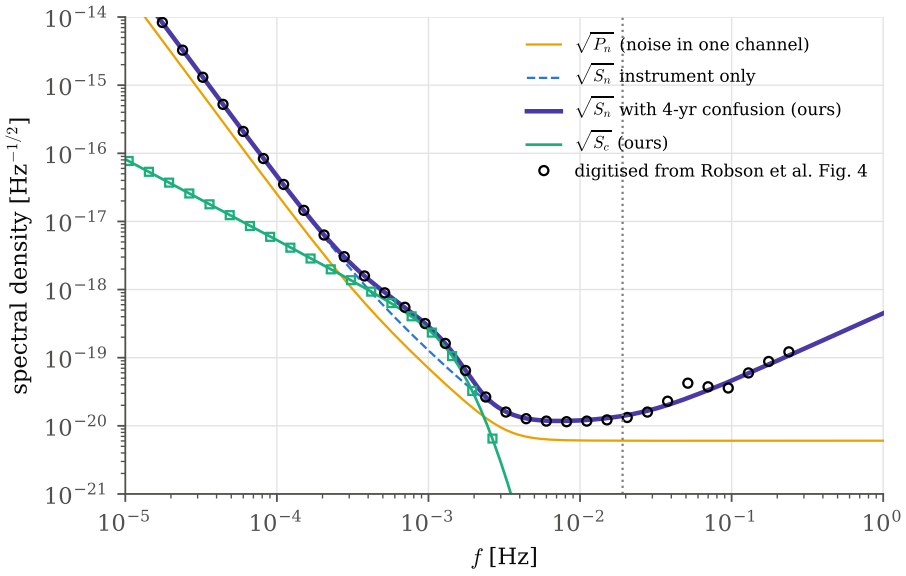


Figure 10.29: The LISA sensitivity used in all forecasts, against Figure 4 of Robson, Cornish and Liu digitised from the published PDF (open circles: $\sqrt{S_n}$; open squares: the Galactic confusion noise $\sqrt{S_c}$). The thick curve includes four years of confusion noise; the thin dashed one is the instrument alone. Above the transfer frequency (dotted) the published curve shows the oscillations of the exact detector response, which the fit smooths.

Their worked numbers. Robson et al. quote three numbers that test both the curve and the signal normalisation.

- *A verification binary.* SDSS J0651+2844 ($0.5 + 0.25\,M_\odot$, $D_L \approx 1$ kpc, $f = 2.6$ mHz) barely changes frequency in four years, so its signal-to-noise integral collapses to one frequency. With the stationary-phase amplitude $|\tilde h| = \frac{1}{2}A_t \dot f^{-1/2}$ (Eq. (T2.51)) and the band $\Delta f = \dot f T$ swept in a time T,

$$\rho^2 \overset{(1)}{=} \frac{16}{5}\int \frac{|\tilde h|^2}{S_n}\mathrm{d}f \overset{(2)}{=} \frac{16}{5}\,\frac{A_t^2 T}{4S_n(f)} \overset{(3)}{=} \frac{h_{\mathrm{GB}}^2}{S_n(f)}, \qquad h_{\mathrm{GB}} = \frac{8\,T^{1/2}(G\mathcal{M}/c^3)^{5/3}\pi^{2/3}f^{2/3}}{5^{1/2}D_L/c},$$

(1) Sky-, polarisation- and inclination-averaged signal-to-noise ratio (Eq. (T2.80)); the $\frac{4}{5}$ in $\frac{16}{5} = 4 \cdot \frac{4}{5}$ is the inclination average (T2.36)), with $|\tilde{h}|$ the face-on stationary-phase amplitude.
(2) $\int |\tilde{h}|^2 \mathrm{d}f = \frac{1}{4} A_t^2 \int \mathrm{d}f/\dot{f} = \frac{1}{4} A_t^2 T$, the amplitude and S_n being constant over the tiny band.
(3) $A_t = 4(G\mathcal{M})^{5/3}(\pi f)^{2/3}/(c^4 D_L)$, so $\frac{16}{5} \cdot \frac{16}{4} = \frac{64}{5}$ and $\sqrt{64/5} = 8/\sqrt{5}$: their eq. (27).

We obtain $\rho = 139.8$; they quote 140. (Without the confusion noise it would be 147.6: at 2.6 mHz the unresolved Galaxy matters.) Our $h_{\mathrm{GB}} = 3.107 \times 10^{-18}\,\mathrm{Hz}^{-1/2}$ is 10% above the 2.8×10^{-18} they print, although the signal-to-noise ratios agree; the printed amplitude is inconsistent with their own $\rho = 140$ and is presumably rounded from slightly different inputs.

- *A GW150914-like binary* five years before merger sweeps from 16 to 29 mHz during four years; we find 15.83 and 28.94 mHz (Section T2.4 derives it by hand).
- *An equal-mass* $10^6 M_\odot$ *binary at* $z = 3$ radiates at 2.93×10^{-5} Hz one year before merger when the detector-frame mass $(1+z)M$ is used; we find 2.927×10^{-5} Hz.

10.11.2 The benchmark source and its vacuum errors

The source is the $10^3 + 1.4\,M_\odot$ binary of Section 10.10.3 at $D_L = 76$ Mpc, the distance of the astrophysical benchmark of Coogan et al. [89]. LISA observes its last five years before the ISCO (Eq. (T2.46)), from $f_{\mathrm{lo}} = 19.96$ mHz to the 1 Hz edge of the band, and the sky-averaged signal-to-noise ratio is $\rho = 13.25$. The parameters are those of Section 10.9.2 (chirp mass, mass ratio, coalescence time t_c and phase ϕ_c, amplitude; their physical meaning is listed in Table T2.1) plus the environmental amplitude,

$$\boldsymbol{\theta} = (\ln \mathcal{M},\ \ln \eta,\ t_c,\ \phi_c,\ \ln \mathcal{A},\ \beta), \qquad \Psi = \Psi_{1.5\mathrm{PN}} + \beta u^b, \tag{10.80}$$

with $\Psi_{1.5\mathrm{PN}}$ the phase (10.70) through 1.5 post-Newtonian order (Section T2.6), evaluated at the vacuum, $\beta = 0$: the question is whether the data could tell an environment apart from no environment. Without the environmental term, the five-parameter errors are $\sigma_{\ln\mathcal{M}} = 2.622{\times}10^{-7}$, $\sigma_{\ln\eta} = 2.700 \times 10^{-4}$, $\sigma_{t_c} = 2.904$ s, $\sigma_{\phi_c} = 6.778$ rad and $\sigma_{\ln\mathcal{A}} = 0.07545 = 1/\rho$. The chirp mass is fixed to a few parts in ten million because the binary completes about a million cycles in the band; this is the reservoir of precision an environment can draw on.

10.11.3 The error on β as a function of the exponent

The script `12_ppe_scan.py` computes the six-parameter Fisher matrix for 815 values of b between -6.5 and 1.6:

Listing 10.8: One Fisher matrix per exponent (`code/ch10/12_ppe_scan.py`)

```
for b in bs:
    F = L.ppe_fisher(f, Mc, eta, DL, Sn, b)
    C = L.inv_scaled(F)
    s = np.sqrt(np.diag(C))
    sig.append(s[5]); sigc.append(1 / np.sqrt(F[5, 5]))
    Dlo.append(3 * s[5] * u_lo**b)
    R = C / np.outer(s, s)
    cMc.append(R[5, 0]); ceta.append(R[5, 1]); ctc.append(R[5, 2]); cphi.append(R[5, 3])
    keep = [0, 2, 3, 4, 5]                          # the same with eta known
    cMc4.append(L.corr(F[np.ix_(keep, keep)])[4, 0])
```

Line 49 is (10.69) with the extra row $\partial \ln \tilde{h}/\partial\beta = -iu^b$. Line 52 records both the marginalised error and the error with all other parameters known, $1/\sqrt{\Gamma_{\beta\beta}}$. Line 53 turns σ_β into a phase: a 3σ shift at the start of the band. The lines that follow redo the calculation in the projection form (10.74), fitting u^b with the vacuum signatures by weighted least squares; the two agree to 2.809×10^{-9}.

Because β multiplies a different power of u for each b, its value is not comparable across b. A physical measure is the phase shift it produces where it is largest, $\sigma_\beta u_{\rm lo}^b$ at $f_{\rm lo}$. Figure 10.30 shows it.

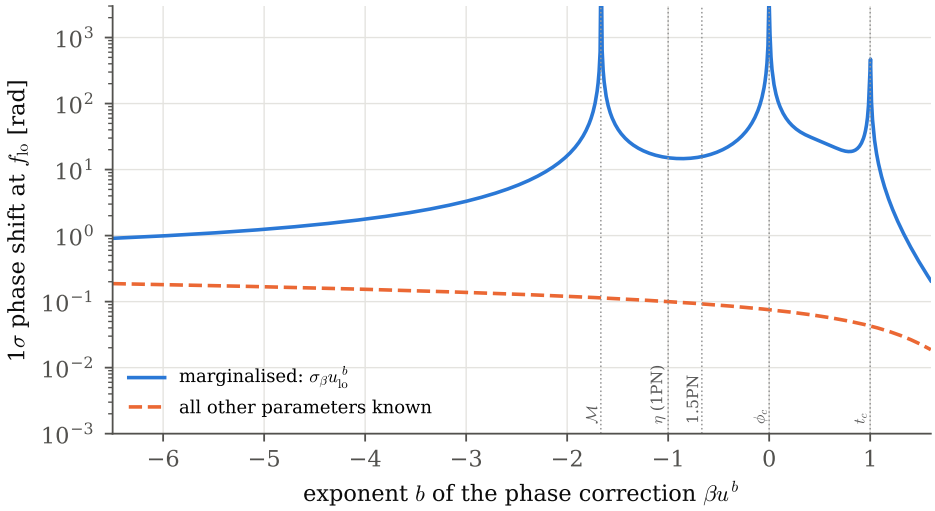


Figure 10.30: The 1σ error of a power-law phase correction, expressed as the phase shift it allows at the start of the band, for the benchmark binary ($\rho = 13.25$). Solid: marginalised over $\ln\mathcal{M}$, $\ln\eta$, t_c, ϕ_c and $\ln\mathcal{A}$. Dashed: with all of them known. The marginalised error diverges where u^b coincides with a vacuum signature: the chirp mass ($b = -5/3$), the phase ($b = 0$) and the arrival time ($b = 1$).

Three things are visible. First, the error diverges at $b = -5/3$, 0 and 1, where u^b is exactly the signature of $\ln\mathcal{M}$, ϕ_c or t_c: there $g_\perp = 0$ and the correction is simply reabsorbed into a wrong chirp mass, phase or arrival time. Second, near the 1PN and 1.5PN exponents (-1, $-2/3$) the error is large but finite: those terms are tied to η with fixed coefficients, so η can imitate only one combination of them. Third, the marginalised error is far larger than the conditional one even away from the singular points: by a factor 11.55 at $b = -4$, 58.43 at the dipole exponent $-7/3$ and 150.8 at the massive-graviton exponent -1. A forecast that holds the masses fixed would be wrong by these factors.

10.11.4 Correlations: why low-frequency effects separate from the chirp mass

The correlation coefficient between two parameters is, in the projection picture, the cosine of the angle between their signatures (after removing the others). For β and $\ln\mathcal{M}$ alone it is $\langle u^b u^{-5/3}\rangle_w / \sqrt{\langle u^{2b}\rangle_w \langle u^{-10/3}\rangle_w}$, and the Cauchy–Schwarz inequality says this is one only if the two functions are proportional over the band. Two decreasing powers look more and more alike as their exponents approach each other, and more and more different as they separate. An effect with $b \ll -5/3$ is concentrated in the first fraction of a decade above $f_{\rm lo}$ (Figure 10.27), while the chirp-mass signature is spread over the whole band. That is why low-frequency effects decorrelate from the chirp mass.

The script shows it directly (Figure 10.31). With η known, the correlation of β with $\ln\mathcal{M}$ is 0.9988 at $b = -2$, 0.9635 at $b = -4$ and 0.8648 at $b = -8$. With η free, the picture changes in an instructive way: η is now the parameter whose signature is closest to the environmental one

(its correlation with β is 0.7417 at $b = -4$), and the correlation with the chirp mass drops to -0.5391. The arrival time, whose signature $2\pi f$ grows with frequency, is the least similar of all, with a correlation near 0.5091 for every steep effect. Across the singular points the correlations jump between ± 1: on one side u^b is slightly steeper than $u^{-5/3}$, on the other slightly flatter, and the sign of the compensating chirp-mass error flips.

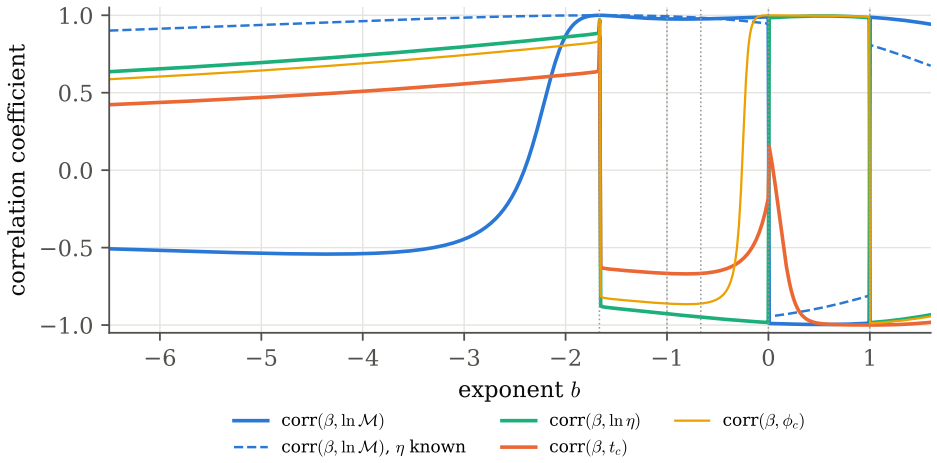


Figure 10.31: Correlation of the environmental amplitude β with the vacuum parameters as a function of the exponent. Dashed: with $\ln \mathcal{M}$ when η is known, which falls steadily as b moves below $-5/3$. Solid: all parameters free; η then takes over part of the chirp-mass role. Dotted verticals mark the GR exponents.

10.11.5 The detectability map

Now the single curve. An environment with exponent b and amplitude β is detected at 3σ if $|\beta| > 3\sigma_\beta(b)$, that is if the phase shift it produces at the start of the band exceeds

$$\delta\Psi_{\mathrm{det}}(b) = 3\,\sigma_\beta(b)\,u_{\mathrm{lo}}^{b}. \tag{10.81}$$

The predicted shift of each physical effect at f_{lo} is the last column of the table in Section 10.10.3. Figure 10.32 puts them together.

Reading the map, in units of the forecast error σ_β for each predicted amplitude:

- friction in a static $\gamma = 7/3$ spike: $1.023 \times 10^6\,\sigma$; in a $\gamma = 3/2$ spike: $22864\,\sigma$;
- collisionless accretion in the 7/3 spike: $9074\,\sigma$;
- type-I migration in a thin disc: $57840\,\sigma$;
- the pull of the enclosed dark matter: $3.36\,\sigma$, marginal;
- Eddington-limited gas accretion onto the heavy hole: $0.875\,\sigma$, not detected.

The dissipative dark-matter and disc effects are not just detectable, they dominate the phase; the problem for them is modelling, not detection (Section 10.12 shows how much the model matters). The two weak effects sit near the curve, where the forecast decides.

The same curve in physical units. For friction, β is proportional to the spike density as long as $\epsilon \ll 1$, so the curve converts into a smallest detectable density. Figure 10.33 shows it as a function of distance (the threshold grows in proportion to D_L, because $\rho \propto 1/D_L$). At 76 Mpc

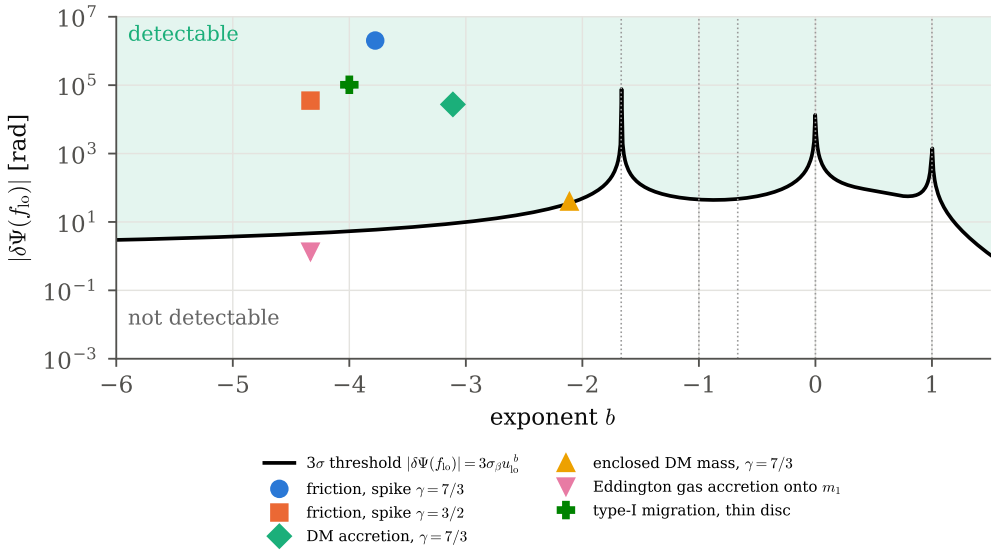


Figure 10.32: Detectability map for the $10^3 + 1.4\,M_\odot$ binary at 76 Mpc ($\rho = 13.25$). The black curve is the phase shift at the start of the band that LISA would detect at 3σ, for every exponent b; the shaded region above it is detectable. Points are the effects of Section 10.10.3 at their exponents. Everything above the curve is seen; the curve moves down in proportion to $1/\rho$ for closer sources.

LISA would see friction from a $\gamma = 7/3$ spike with ρ_6 as low as $1.597 \times 10^{10}\,M_\odot\,\mathrm{pc}^{-3}$ at 10^{-6} pc, five orders of magnitude below the static benchmark; for $\gamma = 2$ and $3/2$ the thresholds are 7.270×10^{10} and 7.148×10^{11}. Above the dotted lines, friction exceeds radiation already at $f_{\rm lo}$ ($\rho_6 > 7.881 \times 10^{15}$ for $\gamma = 7/3$) and the power-law forecast is not applicable; the benchmark density sits just below that line.

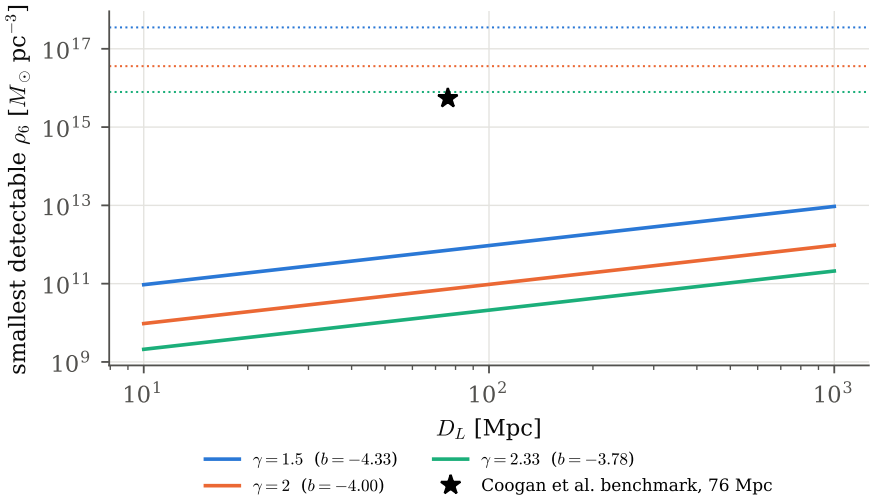


Figure 10.33: Smallest spike density at 10^{-6} pc whose dynamical friction LISA would detect at 3σ in this binary, against luminosity distance, for three spike slopes. Dotted: the density at which friction equals radiation at the start of the band, above which the small-ϵ expansion fails. The star is the astrophysical benchmark of Coogan et al.

Key idea: When is an environment detectable?

Write the environment as $\delta\Psi = \beta u^b$. Fit it with the vacuum parameters by weighted least squares over the band, weight $w \propto |\tilde{h}|^2/S_n$. It is detected at $k\sigma$ when the weighted root-mean-square of the residual phase exceeds k/ρ radians. The rule "a dephasing of more than a radian is detectable" [88] keeps the radian but forgets both the $1/\rho$ and the part that the masses, arrival time and phase can absorb.

How the error scales with the signal strength. In the Fisher approximation every error scales exactly as $1/\rho$: double the distance and the curve rises by two. Whether that scaling holds at the signal-to-noise ratios of real sources is a separate question, answered in Section 10.13 by comparing with a Markov chain.

10.12 Reproducing the published forecasts

A forecast printed in a paper is a claim: "with this source, this detector and this model, the spike slope would be measured to a part in a million". We can trust such a claim only if we can make it again ourselves, with our own code, and land on the same number, or explain why we do not. Two of the claims already passed: the LISA noise curve and its worked examples (Section 10.11.1) and Cutler and Flanagan's table of errors (Section 10.9.4). Here are the two dark-matter forecasts on which most of the later literature builds. Eda, Itoh, Kuroyanagi and Silk [82] put an intermediate-mass black hole inside a static dark-matter spike and asked how well an eLISA-like detector would measure the spike. Kim, Lenoci, Stomberg and Xue [86] repeated the question for wave-like (ultralight) dark matter, with LISA, a halo that loses its inner modes to the black hole, and a full Bayesian posterior.

The procedure is the same for both. First derive the quantity the paper plots, from what is already established. Then rebuild the paper's model, ingredient by ingredient, and compute the same quantity. Then compare, number by number, and when the two disagree, ask what could have produced the disagreement before deciding.

10.12.1 The two phases that papers plot

Both papers summarise the effect of the dark matter by a "phase difference" accumulated between a frequency f and the end of the inspiral. They do not mean the same phase. One is the phase of the Fourier transform, the quantity that enters the likelihood. The other is the phase of the wave in time, which counts cycles. Both follow from the stationary-phase formula of Section T2.5, so we derive both once.

Inputs. The Fourier phase is $\Psi(f) = 2\pi f t(f) - \phi(t(f)) - \pi/4$ (Eq. (T2.51); in words, the Fourier component at f is built by the short stretch of the chirp that passes through f), where $t(f)$ is the time at which the chirp passes through frequency f and $\phi(t)$ is the phase of the wave in time, with $\mathrm{d}\phi/\mathrm{d}t = 2\pi f$. The sweep rate $\dot{f}(f)$ is whatever the physics gives: vacuum (Eq. (T2.43), set by the chirp mass $\mathcal{M}$), or vacuum plus friction (Eq. (T2.91)). Let the signal end at f_I (the ISCO, Eq. (T2.46)) at the time t_c with the phase ϕ_c (the coalescence constants of Section T2.4.3). Then

$$t(f) \overset{(1)}{=} t_c - \int_f^{f_I} \frac{\mathrm{d}f'}{\dot{f}(f')}, \qquad \phi(t(f)) \overset{(2)}{=} \phi_c - 2\pi \int_f^{f_I} \frac{f'\,\mathrm{d}f'}{\dot{f}(f')},$$

$$\Psi(f) \overset{(3)}{=} 2\pi f t_c - \phi_c - \frac{\pi}{4} + 2\pi \int_f^{f_I} \frac{f'-f}{\dot{f}(f')}\,\mathrm{d}f' \equiv 2\pi f t_c - \phi_c - \frac{\pi}{4} + \tilde{\Phi}(f). \tag{10.82}$$

(1) The time it takes to sweep from f to f_I is the sum of $\mathrm{d}t = \mathrm{d}f'/\dot f$ over the frequencies in between.
(2) The phase accumulated over the same interval is $\int 2\pi f'\mathrm{d}t = 2\pi\int f'\mathrm{d}f'/\dot f$.
(3) Insert both into Ψ and collect the two integrals under one sign: $2\pi f\int \mathrm{d}f'/\dot f$ and $2\pi\int f'\mathrm{d}f'/\dot f$ combine into $2\pi\int (f'-f)\mathrm{d}f'/\dot f$.

The integral $\tilde\Phi(f)$ is Eda et al.'s eq. (28c), with their $L(f) = 1+\epsilon(f)$ in $\dot f = \dot f_{\rm V} L$; multiplying out their prefactor $\frac{10}{3}(8\pi G\mathcal{M}/c^3)^{-5/3}$ gives exactly $2\pi/(\dot f_{\rm V} f^{-11/3})$. (With their opposite Fourier sign it enters Ψ with the opposite sign, as the footnote to Eq. (T2.51) explains; the function is the same.) The quantity in their Figure 1 is the difference with and without dark matter, $\Delta\tilde\Phi = \tilde\Phi - \tilde\Phi_{\rm V}$ [82, eq. (29)].

The time-domain phase still to come is the second integral alone, $\Phi(f) = 2\pi\int_f^{f_I} f'\mathrm{d}f'/\dot f$, and its difference with and without dark matter,

$$\Delta\Phi(f) = 2\pi\int_f^{f_I}\left[\frac{f'}{\dot f(f')} - \frac{f'}{\dot f_{\rm V}(f')}\right]\mathrm{d}f', \tag{10.83}$$

is Kim et al.'s eq. (49), plotted in their Figure 3; Eq. (T2.95) evaluated it for a power-law ϵ. The two differ by $\Delta\tilde\Phi = \Delta\Phi - 2\pi f\,\Delta t(f)$, where Δt is the change in the time left before the ISCO: the Fourier phase does not count the cycles that a shift of the arrival time would explain. Both are "accumulated between f and the ISCO", both are dominated by the low-frequency end, and for a power-law ϵ they have the same exponent, $-\frac{5}{3}-n$ (Section T2.10). In the script `13_eda2015.py` the integral (10.82) is two cumulative trapezoid sums run from the ISCO downwards:

Listing 10.9: The Fourier phase accumulated between f and the ISCO (`code/ch10/13_eda2015.py`)

```
45 def phit(f, fdot):
46     """Phi~(f) = 2 pi int_f^{f_ISCO} (f' - f)/fdot df' on the (increasing) grid f, which
       ends at f_ISCO."""
47     A = -cumulative_trapezoid((f / fdot)[::-1], f[::-1], initial=0)[::-1]
48     B = -cumulative_trapezoid((1 / fdot)[::-1], f[::-1], initial=0)[::-1]
49     return 2 * np.pi * (A - f * B)
```

Line 47 is $\int_f^{f_I} f'\mathrm{d}f'/\dot f$ and line 48 is $\int_f^{f_I}\mathrm{d}f'/\dot f$, both accumulated from the top of the grid (the reversed arrays); line 49 combines them as in step (3).

10.12.2 Eda et al. 2015: a static spike and an eLISA forecast

The model. A black hole of $M_{\rm BH} = 10^3 M_\odot$ sits in a spike $\rho = \rho_{\rm sp}(r_{\rm sp}/r)^\alpha$ with $\rho_{\rm sp} = 226\,M_\odot\,\mathrm{pc}^{-3}$ and $r_{\rm sp} = 0.54\,\mathrm{pc}$, the values Eda et al. derive for an NFW minihalo grown adiabatically around the hole [82, Table I]; the slope is $\alpha = 7/3$ for that history and is left free otherwise. A companion of $\mu = 1\,M_\odot$ circles it (they call the companion's mass μ). Friction removes $P_{\rm DF} = 4\pi G^2\mu^2\rho\ln\Lambda/v$ [82, eq. (16)], radiation removes the quadrupole power of a circular binary (Eq. (T2.27)), and Kepler's law uses $M_{\rm BH}$. The ratio is the friction-to-radiation ratio of Eq. (T2.92) with $m_1\to M_{\rm BH}$, $M\to M_{\rm BH}$ and $\xi = 1$:

$$\begin{aligned}\epsilon(f) = \frac{P_{\rm DF}}{P_{\rm GW}} &= \frac{4\pi G^2\mu^2\rho\ln\Lambda\,R^{1/2}(GM_{\rm BH})^{-1/2}}{\frac{32}{5}G^4\mu^2M_{\rm BH}^3/(c^5R^5)}\\ &= \frac{5\pi}{8}\,\frac{c^5\ln\Lambda}{G^{5/2}M_{\rm BH}^{7/2}}\,\rho(R)\,R^{11/2}, \qquad R = \left(\frac{GM_{\rm BH}}{\pi^2 f^2}\right)^{1/3}.\end{aligned} \tag{10.84}$$

The companion's mass cancels, and $\epsilon \propto R^{11/2-\alpha}$. With $\dot{f} = \dot{f}_V(1+\epsilon)$ and $\mathcal{M} = \mu^{3/5}M_{\rm BH}^{2/5} = 15.85\,M_\odot$, (10.82) gives $\Delta\tilde{\Phi}(f)$ for each α, without any expansion in ϵ. The ISCO is at $f_I = 4.397$ Hz.

Example: the dephasing curves. The script renders page 8 of the paper, calibrates the axes of Figure 1 on its tick marks and reads the three curves from the dark pixels; Figure 10.34 overlays them on ours.

The event. With the Coulomb logarithm the text gives, $\ln\Lambda \simeq 3$, our three curves (dashed) run parallel to the published ones and sit below them by the same amount everywhere: the median of $\log_{10}$(ours/theirs) is −0.3672, −0.3674 and −0.355 for $\alpha = 1.5$, 2 and 7/3, across two decades of frequency and fourteen decades of phase.

What could produce a constant factor? We go through the candidates.

- *A misread figure.* Digitisation errors scatter from point to point and do not depend on which curve is read. A common offset of 0.36 in $\log_{10}$ on three separately read curves is not a reading error; the scatter about the offset, 0.02097 to 0.1027 dex, is.
- *Different masses.* A different $M_{\rm BH}$ or μ changes f_I, the vacuum rate and the radius at a given frequency, each with a different power of f. It would tilt the curves, not shift them.
- *A different normalisation of ϵ.* Over the range where $\epsilon \ll 1$, the dephasing is proportional to ϵ (Eq. (T2.95)), and $\epsilon \propto \rho_{\rm sp}\ln\Lambda$. A different value of the product $\rho_{\rm sp}\ln\Lambda$ shifts all curves by the same factor. This fits.

How large is the factor? $10^{0.367} = 2.33$, which is $\ln 10 = 2.303$ to within the scatter. The natural reading is that the authors meant $\log_{10}\Lambda = 3$, a Coulomb logarithm of $\ln 10^3 = 6.908$, and the text abbreviated it as "$\ln\Lambda \simeq 3$". With $\ln\Lambda = 6.908$ (solid curves) the median offsets fall to −0.00495, −0.005213 and −0.009118 dex. The figure cannot separate $\ln\Lambda$ from $\rho_{\rm sp}$, since only their product enters; the Fisher numbers below confirm the reading independently.

The curves also show how large the effect is. At 0.01 Hz the $\alpha = 7/3$ spike has $\epsilon = 10.62$ (friction is ten times stronger than radiation there) and $\Delta\tilde{\Phi} = 3.062\times10^7$ rad; the paper's caption reads this as "a factor of 10^7 more GW cycles". For $\alpha = 1.5$ the same number is 58.96 rad, and at 1 Hz only 1.262×10^{-7} rad: a shallow spike leaves its mark only at the low-frequency end of the band.

The forecast. Eda et al. compute a Fisher matrix for six parameters, $\ln\mathcal{A}$, t_c, Φ_c, $\ln\mathcal{M}$, $\ln\alpha$ and $\ln c_\epsilon$, where c_ϵ is the strength of the friction [82, eqs. (37)–(38)]. Their detector is the eLISA design of the time, with arm length 10^9 m and the noise of their eq. (32a), which `lib_gw10.py` implements as `Sn_elisa`. The band runs from the frequency five years before the ISCO, $f_{\rm ini}(\alpha)$, to f_I [82, eq. (40)]: friction speeds up the early inspiral, so a steeper spike starts the five years at a lower frequency. We find $f_{\rm ini} = 0.02264$ Hz for $\alpha \to 0$ (the vacuum), 0.01481 Hz for $\alpha = 7/3$ and 7.586×10^{-4} Hz for $\alpha = 2.5$: a plateau and then a steep fall above $\alpha \approx 2.3$, the shape of their Figure 2 (Figure 10.35, left). The errors are normalised to $S/N = 10$ by scaling the matrix so that $\Gamma_{\ln\mathcal{A}\ln\mathcal{A}} = \rho^2 = 100$ (Eq. (10.72)). The derivatives with respect to $\ln\alpha$ and $\ln c$ are central differences of the complete waveform, amplitude factor $L^{-1/2}$ included, as in the paper.

The comparison at $\alpha = 7/3$, $S/N = 10$, with their eqs. (41a)–(41f):

	$\Delta\mathcal{A}/\mathcal{A}$	Δt_c [s]	$\Delta\Phi_c$ [rad]	$\Delta\mathcal{M}/\mathcal{M}$	$\Delta\alpha/\alpha$	$\Delta c/c$
Eda et al. (41)	0.1	1	1.3	3.100×10^{-7}	1.200×10^{-6}	5.900×10^{-5}
ours, $\ln\Lambda = 3$	0.1	1.273	1.42	3.280×10^{-7}	3.211×10^{-6}	see text
ours, $\ln\Lambda = 6.9$	0.1	1.213	1.366	3.224×10^{-7}	1.254×10^{-6}	see text

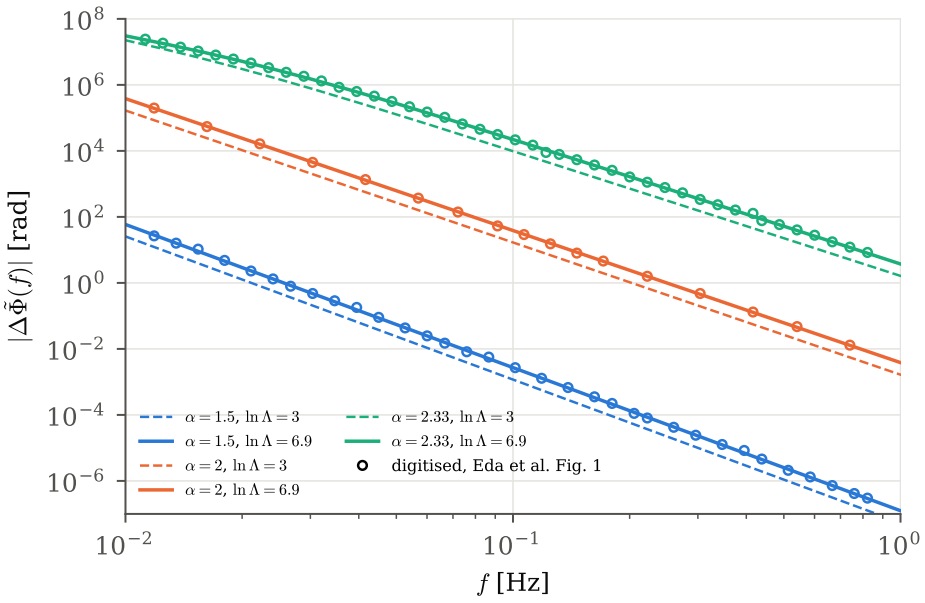


Figure 10.34: The Fourier-phase difference between an inspiral in a dark-matter spike and in vacuum, accumulated between f and the ISCO, for three spike slopes. Open circles: Figure 1 of Eda et al., digitised from the published PDF. Dashed: our calculation with the Coulomb logarithm stated in their text, $\ln\Lambda = 3$; it is low by the same factor 2.3 on every curve. Solid: with $\ln\Lambda = \ln 10^3 = 6.9$, which reproduces the figure.

The amplitude error is $1/\rho$ by construction. The chirp mass agrees to 4%, the phase to 5%, the arrival time to 20%; these three depend on the shape of the noise curve and on where the band starts, which we rebuilt from the paper's description, not from its code. The slope settles the Coulomb logarithm: with $\ln\Lambda = 3$ we get $\Delta\alpha/\alpha = 3.211\times10^{-6}$, almost three times the published value; with $\ln\Lambda = 6.9$ we get 1.254×10^{-6}, within 5% of it. A stronger friction measures the slope better, because the signature it leaves is larger compared with the noise, and the figure and the forecast now agree on the same strength.[4]

Example: why the amplitude's error depends on where it is defined. Our $\Delta c/c$ is 2.200×10^{-6}, twenty-seven times smaller than their 5.900×10^{-5}. The event is a disagreement in one parameter only, the amplitude of the friction, while the slope in the same matrix agrees. What could make the error of an amplitude differ while everything else agrees? The amplitude is not a single number until we say at which frequency it is the amplitude. Our c is ϵ at the pivot $f_p = 0.05\,\mathrm{Hz}$, inside the band: $\epsilon(f) = c\,(f/f_p)^{-n}$, with $n = (11-2\alpha)/3$ from (10.84). The same law, written with the amplitude c' at another pivot f_q, has $c' = c\,(f_q/f_p)^{-n}$. Then

$$\ln c' \overset{(1)}{=} \ln c - n\ln\frac{f_q}{f_p},$$

$$\delta\ln c' \overset{(2)}{=} \delta\ln c + \frac{2}{3}\,\alpha\ln\frac{f_q}{f_p}\,\delta\ln\alpha,$$

$$\sigma^2_{\ln c'} \overset{(3)}{=} \sigma^2_{\ln c} + 2\lambda\,\mathrm{Cov}(\ln c,\ln\alpha) + \lambda^2\sigma^2_{\ln\alpha}, \qquad \lambda \equiv \frac{2}{3}\,\alpha\ln\frac{f_q}{f_p}. \tag{10.85}$$

(1) Take the logarithm of $c' = c\,(f_q/f_p)^{-n}$.

[4] The conclusion of Eda et al. quotes $\pm5\times10^{-6}$ for the 1σ error on α; eq. (41e), which we reproduce, corresponds to $\Delta\alpha = 2.8\times10^{-6}$.

(2) Vary both sides; n depends on α through $\mathrm{d}n = -\frac{2}{3}\mathrm{d}\alpha = -\frac{2}{3}\alpha\,\mathrm{d}\ln\alpha$.
(3) The variance of a linear combination $\delta\ln c + \lambda\,\delta\ln\alpha$ (the propagation rule of Section 10.4.5).

The error of the amplitude grows quadratically with the lever arm λ, the logarithmic distance between the two pivots. Where is Eda et al.'s pivot? Their c_ϵ multiplies $x^{(11-2\alpha)/2}$, where x is the orbital radius in units of R_*, the radius at which the spike, extrapolated inwards, would enclose a mass equal to the black hole's [82, eqs. (10b), (17), (18), (28d)–(28f)]. For their halo $R_* = 0.9829\,\mathrm{pc}$. The orbit in the band has a radius of about $10^{-8}\,\mathrm{pc}$, so their amplitude is defined eight decades in radius, and twelve in frequency ($f_* = 2.195\times10^{-14}\,\mathrm{Hz}$; Eda et al.'s pivot, not LISA's transfer frequency), away from the data. The script evaluates (10.85) at f_*:

Listing 10.10: The amplitude error at another pivot (`code/ch10/13_eda2015.py`)

```
    C = L.inv_scaled(F)
    a73 = 7 / 3
    Rstar = (MBH * L.MSUN * (3 - a73) / (4 * np.pi * RHOSP * L.MSUN / L.PC**3 * RSP**a73))
    ** (1 / (3 - a73))
    fstar = np.sqrt(L.G * MBH * L.MSUN / Rstar**3) / np.pi
    def dlnc_at(fq):
        lev = 2 / 3 * a73 * np.log(fq / 0.05)
        return np.sqrt(C[5, 5] + lev**2 * C[4, 4] + 2 * lev * C[4, 5])
    fq = np.geomspace(1e-16, 1.0, 4000)
    dq = np.array([dlnc_at(x) for x in fq])
    nums.update({f"TenBedaRstar{tag}": Rstar / L.PC, f"TenBedaFstar{tag}": fstar,
                 f"TenBedaCstar{tag}": dlnc_at(fstar), f"TenBedaFbest{tag}":
    fq[np.argmin(dq)],
                 f"TenBedaCbest{tag}": dq.min(), f"TenBedaCorrAC{tag}": C[4, 5] /
    np.sqrt(C[4, 4] * C[5, 5])})
```

Line 181 is R_*, from $4\pi\rho_{\mathrm{sp}}r_{\mathrm{sp}}^{\alpha}R_*^{3-\alpha}/(3-\alpha) = M_{\mathrm{BH}}$; line 182 is Kepler's law at R_*; lines 183–185 are (10.85). At f_* the error is 5.336×10^{-5}, against their 5.900×10^{-5}: the factor of twenty-seven has become ten per cent. In the other direction, the pivot at which c and α decorrelate, $f_q = 0.01658\,\mathrm{Hz}$, gives the smallest possible error, 4.133×10^{-7}. The decorrelation pivot sits where the signal-to-noise ratio accumulates, for the same reason that Knox's tilt and amplitude decorrelate at $\ell_* \approx 210$ in Section 10.6.2: an amplitude measured where the data are is not dragged by the uncertainty of the slope. At our $f_p = 0.05\,\mathrm{Hz}$ the correlation between $\ln c$ and $\ln\alpha$ is already 0.9822; at f_* it is essentially one.

Watch out: An amplitude without a pivot has no error bar

When a power law has a free exponent, the error of its amplitude depends on the frequency (or scale) at which the amplitude is defined, and grows with the logarithmic distance from the data. Quote amplitudes at a pivot inside the band, ideally the decorrelation pivot, or quote the pivot with the error.

Example: how shallow a spike can be measured. Repeating the forecast for every slope gives $\Delta\alpha/\alpha$ as a function of α at $S/N = 10$ (Figure 10.35, right). The error falls by many orders of magnitude between $\alpha = 1.4$ and 2.4, because $\epsilon \propto \rho_{\mathrm{sp}}(r_{\mathrm{sp}}/R)^{\alpha}$ grows rapidly with α at the tiny radii of the band, and turns up near $\alpha \approx 2.5$ where the band start drops and the binary spends its five years with friction in charge. Eda et al. state that the slope is measured to 10% down to $\alpha \approx 1.7$ [82, Sec. IV D]. We find $\alpha = 1.738$ (and 1.783 with $\ln\Lambda = 3$).

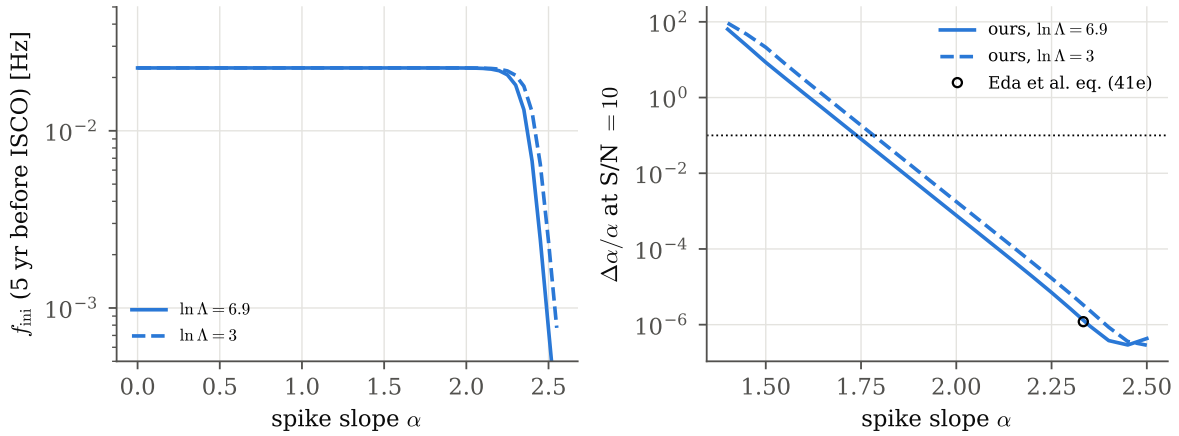


Figure 10.35: Left: the frequency at which the last five years before the ISCO begin, against the spike slope; friction makes steep spikes start lower. Right: the forecast relative error on the slope at $S/N = 10$ with the eLISA noise of Eda et al., for the two Coulomb logarithms; the circle is their eq. (41e) and the dotted line the 10% level, crossed near $\alpha = 1.7$ as they state.

10.12.3 Kim et al. 2023: a wave-like halo, LISA, and a posterior

The model. The benchmark of Kim et al. [86, Table I] is a $m_1 = 10^4 M_\odot$ black hole with a $m_2 = 1\,M_\odot$ companion ($\mathcal{M} = 39.8\,M_\odot$, $q = m_2/m_1 = 10^{-4}$) at $D_L = 203\,\mathrm{Mpc}$, in a spike of slope $\gamma_{\rm sp} = 7/3$ with density $\rho_6 = 25\times10^{15} M_\odot\,\mathrm{pc}^{-3}$ at $r_6 = 10^{-6}\,\mathrm{pc}$. The dark matter is a wave with gravitational Bohr radius $a = 10^{-11}\,\mathrm{pc}$. Each ingredient of their model has a short physical reason, and each is a line of `14_kim2023.py`.

- *Energy balance.* $\mathrm{d}E_o/\mathrm{d}t = -(P_{\rm GW} + \dot E_f)$ with the vacuum orbital energy [86, eq. (36)], so $\dot f = \dot f_{\rm V}(1+\epsilon)$ with $\epsilon = \dot E_f/P_{\rm GW}$, exactly Eq. (T2.91).
- *Friction.* $\dot E_f = 4\pi(Gm_2)^2\rho(<v)\,C/v$, with $\rho(<v)\approx 0.6\rho$ the density of the slower particles [86, eq. (39)]: Chandrasekhar's formula (Eq. (T2.87)) with $\xi = 0.6$ and the Coulomb logarithm replaced by a factor C. For particles, $C_p = \frac{1+\Lambda}{\Lambda}\ln\left[1+\Lambda+\sqrt{\Lambda(2+\Lambda)}\right] - \sqrt{1+2/\Lambda}$ with $\Lambda = v^2 r/(Gm_2)$, which tends to $\ln 2\Lambda - 1$ for $\Lambda \gg 1$ [86, eq. (41)]. For a wave, $C_w(kr) = \mathrm{Cin}(2kr) - 1 + \sin(2kr)/2kr$ [86, eq. (42)], whose limits Section T2.10 derives: friction is suppressed as $(kr)^2/3$ when the de Broglie wavelength exceeds the orbit.
- *The density.* For particles the spike is cut off at twice the Schwarzschild radius, $\rho_6(r_6/r)^\gamma(1-2R_S/r)^\gamma$. For a wave, the modes of low angular momentum fall into the hole: a mode (n,ℓ) decays at a rate $\Gamma_{n\ell}$ that grows steeply as ℓ falls [86, eq. (31)], and the modes with $2\Gamma_{n\ell}t_{\rm halo} > 1$ are gone after the halo's age. The lowest surviving angular momentum ℓ_c sets an inner radius $R_c = a\,\ell_c(\ell_c+1)$ and the profile $\rho_6(r_6/r)^\gamma(1-R_c/2r)^\gamma$ [86, eq. (40)]. We solve $2\max_n \Gamma_{n\ell}t_{\rm halo} = 1$ for continuous ℓ with $t_{\rm halo} = 13.6\,\mathrm{Gyr}$ (a halo formed at $z\approx 20$) and find $\ell_c = 24.03$, 10.22 and 5.16 for $a = 10^{-11}$, 10^{-10} and $10^{-9}\,\mathrm{pc}$, that is particle masses $mc^2 = 9.245\times10^{-14}$, 2.924×10^{-14} and $9.245\times10^{-15}\,\mathrm{eV}$ (their figure labels: 9×10^{-14}, 3×10^{-14}, $9\times10^{-15}\,\mathrm{eV}$).
- *Backreaction.* The friction cannot dissipate more energy than the halo can absorb. Kim et al. cap it by the rate at which the companion sweeps through the binding energy of the shells it crosses: $\dot E_f \to (1/\dot E_f + 2/\dot U)^{-1}$, a smooth minimum of the two [86, eq. (43)].

The cap needs one more step, because $\dot U$ depends on how fast the orbit shrinks, which depends on the capped friction itself. Write x for the capped friction power. A shell of thickness Δr at r

has binding energy $\Delta U = G[m_1 + m_{\rm enc}(r)]\,4\pi r^2\rho\,\Delta r/r$, and the orbit crosses shells at the rate $|\dot r|$:

$$\dot U \overset{(1)}{=} \frac{G(m_1+m_{\rm enc})\,4\pi r^2\rho}{r}\,|\dot r| \overset{(2)}{=} \frac{G(m_1+m_{\rm enc})\,4\pi r\rho\cdot 2r^2}{Gm_1m_2}\,(P_{\rm GW}+x) \equiv \kappa\,(P_{\rm GW}+x),$$
$$\frac{1}{x} \overset{(3)}{=} \frac{1}{\dot E_f} + \frac{2}{\kappa(P_{\rm GW}+x)} \quad\Longrightarrow\quad \kappa x^2 - \big[\kappa(\dot E_f - P_{\rm GW}) - 2\dot E_f\big]x - \kappa \dot E_f P_{\rm GW} \overset{(4)}{=} 0. \tag{10.86}$$

(1) The binding energy per unit radius times the radial speed.
(2) $E_o = -Gm_1m_2/2r$ gives $|\dot r| = |\dot E_o|/(\mathrm{d}E_o/\mathrm{d}r) = (P_{\rm GW}+x)\,2r^2/(Gm_1m_2)$; G cancels.
(3) Kim et al.'s eq. (43) with $\dot U$ from step (2).
(4) Multiply by $x\,\dot E_f\,\kappa(P_{\rm GW}+x)$ and collect powers of x; the positive root is the capped friction.

Kim et al. do not say how they evaluate $\dot U$; taking the radial speed from the total power, as here, is the self-consistent choice, and the quadratic (10.86) is lines 108–110 of the script.

Example: the dephasing curves. Figure 10.36 compares $|\Delta\Phi|$ of (10.83) with Figure 3 (top) of Kim et al., digitised by colour. For the particle halo the median of $\log_{10}$(ours/theirs) is 0.005896 and the largest deviation 0.01903 dex; for the wave halo with $a = 10^{-11}$ pc, −0.005077 and 0.0699; for $a = 10^{-10}$ pc, 0.01456 and 0.1264. For $a = 10^{-9}$ pc the median deviation is 0.2712 dex: there the inner cut-off R_c lies inside the band, the curve plunges where the companion reaches $R_c/2$, and a small change in ℓ_c moves the plunge enough to change the phase there by large factors; in Figure 10.36 our green curve plunges slightly earlier than theirs. The comparison is therefore a test of the decay rates as much as of the friction. At 10 mHz the particle halo has dephased the signal by 4.387×10^5 rad and the wave halo with $a = 10^{-11}$ pc by 2.983×10^5 rad. Everything in the figure depends on ϵ only through (10.83), so this agreement confirms the whole chain from the decay rates to the backreaction cap.

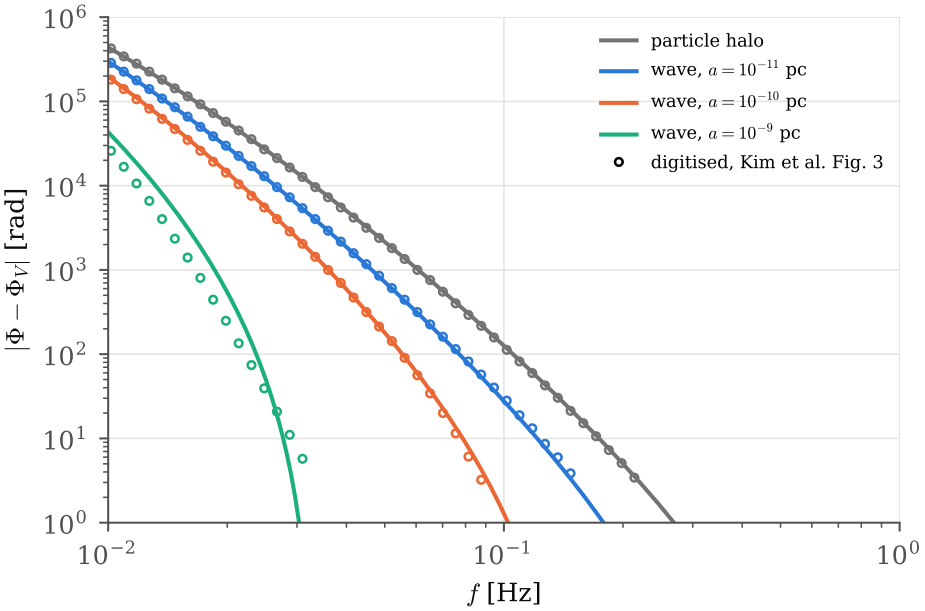


Figure 10.36: The time-domain phase lost between f and the ISCO by the $10^4 + 1\,M_\odot$ inspiral in a particle halo (grey) and in wave halos of three Bohr radii, against the published Figure 3 of Kim et al. (open circles, digitised). The larger the Bohr radius, the more inner modes the black hole has swallowed and the earlier the friction switches off.

Example: the signal-to-noise ratio. With the sky-averaged amplitude $\sqrt{4/5}\,A(f)$ [86, eqs. (44)–(45)] (the $4/5$ is the average over the orbit's inclination, (T2.36); the sky and polarisation average sits in the noise curve, Section T2.3), the Robson et al. noise with four-year Galactic confusion (Sections T2.7 and 10.11.1), and the band from five years before the ISCO, $f_{5\mathrm{yr}} = 0.01269\,\mathrm{Hz}$, to $\min(1\,\mathrm{Hz}, f_I) = 0.4398\,\mathrm{Hz}$, we find $\rho = 15.05$. They state $\rho \simeq 15$.

Example: the forecast against their posterior. Kim et al. do not compute a Fisher matrix. They inject the benchmark signal without noise, $d = h(\boldsymbol{\theta}_{\mathrm{true}})$ (the "median experiment" of Section 10.13), and sample the posterior of $(\mathcal{M}, \log_{10} q, \rho_6, \gamma_{\mathrm{sp}}, \log_{10} a)$ with a nested sampler, maximising the likelihood over the arrival time and the phase and fixing the distance at its best value [86, eqs. (50)–(53), Fig. 4]. Their Figure 4 quotes the 2.5% and 97.5% quantiles of each marginal. We turn each 95% interval into a Gaussian-equivalent width by dividing it by 3.92, and compare with the Fisher matrix of eight parameters, the five plus $(\ln \mathcal{A}, t_c, \phi_c)$, with the derivatives taken by central differences of the complete waveform (amplitude factor $(1+\epsilon)^{-1/2}$ included):

Listing 10.11: The waveform whose derivatives make the Fisher matrix (`code/ch10/14_kim2023.py`)

```
212 Sn = L.Sn_robson(ff, "4yr")
213 nums.update(TenBkimfFive=f5, TenBkimfUp=fup)
214
215
216 def lnh(x, wave=True):
217     """ln h(f) up to the overall amplitude: -(1/2) ln(1+eps) - i Psi, Psi = 2 pi f t(f) -
        Phi(f) (ISCO reference)."""
218     Mc, lq, rho6, gam, la = x
219     eps = eps_kim(ff, Mc, 10**lq, rho6, gam, 10**la if wave else None)
220     fdv = L.fdot_vac(ff, Mc)
221     g = 1 / (fdv * (1 + eps)) - 1 / fdv
```

Line 217 is the integrand of (10.83) divided by f', and line 219 builds the dark-matter part of $\tilde{\Phi}$ from it as in (10.82); line 221 returns $\ln \tilde{h}$ up to a constant, whose real part is the slower-sweep amplitude and whose imaginary part is the phase. Halving and doubling the steps changes no error by more than 3.524×10^{-5}. At the Gaussian level, maximising over $(t_c, \phi_c, \ln \mathcal{A})$ and marginalising over them give the same widths (both are the Schur complement of Section 10.4.1), so the comparison is fair. The result (Figure 10.37):

	$\mathcal{M}\,[M_\odot]$	$\log_{10} q$	$\rho_6\,[10^{15} M_\odot \mathrm{pc}^{-3}]$	γ_{sp}	$\log_{10} a$
posterior (Kim et al.)	4.538×10^{-4}	0.03513	11.53	0.1319	0.1003
Fisher (ours)	8.232×10^{-5}	0.01884	6.978	0.09143	0.06318
ratio	0.1814	0.5364	0.6052	0.6932	0.6302

The order of magnitude is right for every parameter, and the Fisher matrix finds the same structure as the posterior: a correlation of -0.9999 between ρ_6 and γ_{sp} and of 0.9912 between ρ_6 and $\log_{10} q$, the two degeneracies Kim et al. point out in their Figure 4. But every Fisher width is smaller, by a factor 1.4 to 5.5. The matrix is nearly singular (the ratio of its largest to smallest eigenvalue, after scaling to unit diagonal, is 4.160×10^{10}), and along its long axis a 1σ step moves ρ_6 by a quarter of its value and γ_{sp} by 0.09, far beyond the region where the waveform is linear in the parameters. Section 10.13 quantifies this and shows that it, not the code, is the cause. For a particle halo of the same density the four common errors would be $3.211 \times 10^{-5}\,M_\odot$, 0.002314, 0.5984 and 0.007009: the dense inner spike that the wave halo has lost carries most of the information.

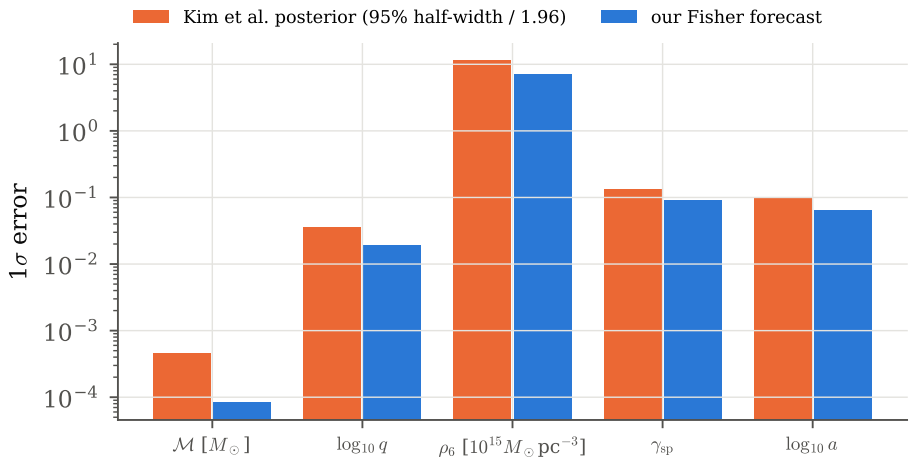


Figure 10.37: The 1σ widths of the five halo and binary parameters for the benchmark of Kim et al. ($\rho \approx 15$): their nested-sampling posterior (Figure 4, 95% half-width divided by 1.96) and our Fisher forecast. The forecast is right to within a factor of a few and too optimistic everywhere, most for the chirp mass, whose posterior is skewed by its correlation with the mass ratio and the density.

10.12.4 Paper against ours

Table 10.6 collects every published gravitational-wave number we have reproduced. Where we agree to better than the precision of reading a plotted line, the paper's model has been rebuilt correctly. Where we disagree, the disagreement has been traced to a cause: a fitted response function (Robson et al. above the transfer frequency f_*, Eq. (T2.38)), a rounded amplitude (the printed $h_{\rm GB}$), an abbreviated Coulomb logarithm and a pivot far outside the band (Eda et al.), or the breakdown of the Gaussian approximation (Kim et al.).

paper	quantity	published	ours
Robson et al. 2019	$\sqrt{S_n}$, 1–30 mHz (Fig. 4)	curve	within 8.594%
	J0651 signal-to-noise ratio	140	139.8
	GW150914-like band, 5 yr before merger	16–29 mHz	15.83–28.94 mHz
	$10^6 M_\odot$ pair at $z = 3$, 1 yr before	2.93×10^{-5} Hz	2.927×10^{-5} Hz
Cutler & Flanagan 1994	Table I, 25 errors at $\rho = 10$	table	all within 2.33%
Eda et al. 2015	$\Delta\tilde{\Phi}(f)$, three slopes (Fig. 1)	curves	-0.009118 to -0.00495 dex ($\ln\Lambda = 6.9$)
	$\Delta\mathcal{M}/\mathcal{M}$ at $S/N = 10$	3.1×10^{-7}	3.224×10^{-7}
	$\Delta\alpha/\alpha$ at $S/N = 10$	1.2×10^{-6}	1.254×10^{-6}
	$\Delta c_\epsilon/c_\epsilon$ (their pivot)	5.9×10^{-5}	5.336×10^{-5}
	Δt_c, $\Delta\Phi_c$	1.0 s, 1.3	1.213 s, 1.366
	slope measured to 10%	$\alpha \approx 1.7$	1.738
Kim et al. 2023	$\lvert\Delta\Phi(f)\rvert$, particle and $a = 10^{-11}$ pc (Fig. 3)	curves	0.005896, -0.005077 dex
	signal-to-noise ratio	$\simeq 15$	15.05
	widths of $(\mathcal{M}, \log q, \rho_6, \gamma, \log a)$	Fig. 4	Fisher $\times$(0.1814–0.6932)

Table 10.6: Published numbers against our reproductions. Deviations in dex are medians of $\log_{10}$(ours/theirs) along the digitised curves.

10.13 When the Fisher ellipse lies

The benchmark of Kim et al. ended on an uncomfortable note (Section 10.12.3): the Fisher matrix found the right parameters to worry about and the right correlations, but every width came out too small, by factors up to five compared with the posterior that a sampler drew from the same model. A forecast that is optimistic by a factor of five can turn "LISA will measure the dark-matter density" into "LISA will see that there is something there". So before any detectability map is believed, the question in words is:

for this source, at this signal-to-noise ratio, does the Fisher ellipse describe the posterior we would actually obtain, and if not, where does it fail?

Of Vallisneri's three readings of Γ^{-1} (Section 10.2.3) [193], the one a forecast relies on is the third: the covariance of the posterior from one experiment. The posterior is something we can compute exactly, by sampling, so the question has an answer that does not depend on approximations.

10.13.1 The linearised signal

The median experiment. To compare a forecast with a posterior we need data. Papers, Kim et al. and Coogan et al. among them [86, 89], use the data with the noise set to zero, $d = h(\boldsymbol{\theta}_0)$, and we do the same. This is not a claim that there is no noise. A noise realisation moves the peak of the posterior away from $\boldsymbol{\theta}_0$, by a random amount whose covariance (at high signal-to-noise ratio) is Γ^{-1} itself, but it leaves the shape of the posterior almost unchanged. Setting $n = 0$ removes the random shift and keeps the shape, the quantity a forecast is about. The noise-free data stand for the typical, or median, experiment.

When is the likelihood exactly the Fisher Gaussian? With $d = h_0 \equiv h(\boldsymbol{\theta}_0)$ the likelihood (10.41) is $\ln \mathcal{L}(\boldsymbol{\theta}) = -\frac{1}{2}(\Delta h|\Delta h)$, with $\Delta h \equiv h(\boldsymbol{\theta}) - h_0$. Expand the signal around the truth, $\Delta h = \Delta\theta^i \partial_i h + R$, where the first term is the tangent (linear) part, the sum over i is implied, and R is everything of second order and higher. Then

$$\begin{aligned} -2\ln\mathcal{L}(\boldsymbol{\theta}) &\overset{(1)}{=} \left(\Delta\theta^i\partial_i h + R \,\middle|\, \Delta\theta^j\partial_j h + R\right) \\ &\overset{(2)}{=} \Delta\theta^i\Delta\theta^j\,(\partial_i h|\partial_j h) + 2\,\Delta\theta^i(\partial_i h|R) + (R|R) \\ &\overset{(3)}{=} \Delta\boldsymbol{\theta}^{\mathsf{T}}\Gamma\,\Delta\boldsymbol{\theta} + 2\,\Delta\theta^i(\partial_i h|R) + (R|R). \end{aligned} \tag{10.87}$$

(1) Insert the expansion of Δh into $-2\ln\mathcal{L} = (\Delta h|\Delta h)$.
(2) The inner product is bilinear and symmetric.
(3) $(\partial_i h|\partial_j h) = \Gamma_{ij}$, (10.69).

If the signal were exactly linear in the parameters, $R = 0$, the log-likelihood would be exactly the quadratic $-\frac{1}{2}\Delta\boldsymbol{\theta}^{\mathsf{T}}\Gamma\Delta\boldsymbol{\theta}$, and with a flat prior the posterior would be exactly the Gaussian of covariance Γ^{-1}, at any signal-to-noise ratio. This is the **linearised-signal approximation** (LSA): it is the whole content of a Fisher forecast. Everything that makes the forecast wrong is in R.

Geometrically (Figure 10.38), as $\boldsymbol{\theta}$ varies the waveforms $h(\boldsymbol{\theta})$ trace out a curved surface in the space of all possible data, and the noise is a cloud of radius about one around the true signal, in the units of the inner product (Eq. (10.43)). The Fisher matrix replaces the surface by its tangent plane. That is harmless if the surface is flat across the cloud and fatal if it bends within it.

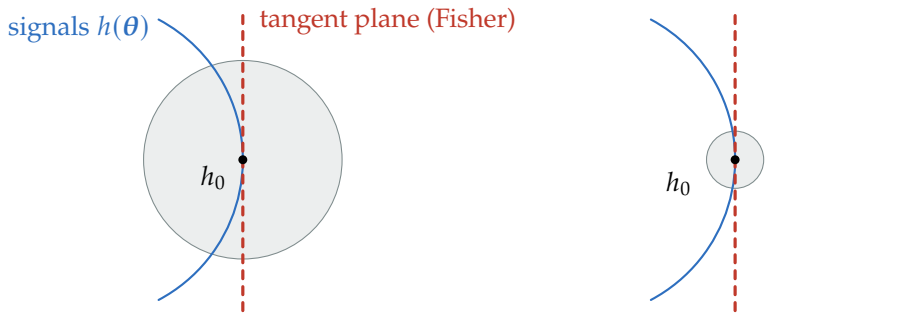


Figure 10.38: The linearised-signal approximation in the space of data. The blue arc is the family of signals $h(\boldsymbol{\theta})$ as one parameter varies, the dashed red line its tangent at the true signal h_0, and the grey disc the region of radius one that the noise fills. When the signal is weak compared with the noise, the true family departs from its tangent inside the disc and the Fisher Gaussian is wrong; when it is strong, the disc is small and the tangent is all that matters.

A test of the approximation. How large may R be? The Fisher 1σ ellipsoid is the surface $\Delta\boldsymbol{\theta}^{\mathsf{T}}\Gamma\Delta\boldsymbol{\theta} = 1$, where the forecast says the log-likelihood has fallen by $\frac{1}{2}$. Vallisneri proposes to walk over that surface and compute there the noise-weighted size of the nonlinear part,

$$|\log r| = \tfrac{1}{2}\,(R \,|\, R), \qquad R = h(\boldsymbol{\theta}) - h_0 - \Delta\theta^i \partial_i h, \tag{10.88}$$

and to accept the forecast if $|\log r| \lesssim 0.1$ everywhere on it [193, eq. (47)]. He shows that this is the logarithm of the ratio of the exact and the linearised likelihood, averaged over the noise realisations that would put the best fit on that surface. On the boundary that the forecast calls "one sigma", the true signals must differ from their tangent-plane stand-ins by much less than the noise. Since the 1σ surface shrinks as $1/\rho$ and R is quadratic in $\Delta\boldsymbol{\theta}$, R shrinks as $\rho \cdot \rho^{-2}$ and $|\log r|$ falls as $1/\rho^2$: every forecast becomes right at high enough signal-to-noise ratio, and the test says how high is high enough.

Key idea: When to trust a Fisher forecast

A Fisher forecast is the posterior of the linearised signal. Check it on its own 1σ surface: the part of the signal change that is not linear in the parameters must have noise-weighted size $\frac{1}{2}(R|R) \lesssim 0.1$ there. Where it fails, sample the posterior.

Example: the benchmark of Kim et al. The script `14_kim2023.py` draws 300 random points on the 1σ surface of its eight-parameter Fisher matrix and evaluates (10.88) with the complete waveform. The median is $|\log r| = 219.5$ (the largest 246.2), and every point fails the test: not by a little, but by three orders of magnitude. The reason is the near-degeneracy of ρ_6 and γ_{sp} (their Fisher correlation is -0.9999): along it the 1σ surface stretches to changes of the density by a quarter and of the slope by a tenth, and there the dephasing, a strongly nonlinear function of both, is nothing like its tangent. Repeating the test with the surface scaled down, as it would be for a louder source, the median falls to 15.07 at a hundred times the signal-to-noise ratio and reaches 0.1 only near $\rho \approx 20000$. For this benchmark the honest forecast is the posterior; the Fisher matrix is a guide to which parameters are degenerate, not to how large the errors are.

10.13.2 A chain against the ellipse

The criterion says whether to trust the forecast; it does not say how wrong it is. For that we sample the posterior with the Metropolis–Hastings chain of Chapter 9 and compare. To keep

the chain small, the problem is reduced to the four parameters that matter for an environment.

The set-up. The source is the $10^3 + 1.4\,M_\odot$ binary of Section 10.11.2, observed by LISA over its last five years from $f_{\rm lo} = 19.96\,$mHz. Its phase is the 1.5PN phase (10.70) (the vacuum template through 1.5 post-Newtonian order, Section T2.6, with chirp mass $\mathcal{M}$ and the coalescence time and phase t_c, ϕ_c of Section T2.4.3) plus a power-law environment (the ppE term of Eq. (T2.69)) written through its size at the start of the band,

$$\Psi(f) = \Psi_{1.5\rm PN}(f;\mathcal{M},\eta,t_c,\phi_c) + A\left(\frac{u}{u_{\rm lo}}\right)^b, \qquad A = \delta\Psi(f_{\rm lo}), \tag{10.89}$$

with $u = \pi G\mathcal{M}f/c^3$ as in Section 10.10 and $u_{\rm lo}$ its value at $f_{\rm lo}$. Writing the amplitude as the phase shift at $f_{\rm lo}$, rather than as β, keeps it of order radians for every b. The free parameters are $(\ln\mathcal{M}, t_c, A, b)$; η is held at its true value, the amplitude $\ln\mathcal{A}$ decouples (Section 10.9.3) and is held too, and ϕ_c is integrated out analytically. The truth is a friction-like effect with $b = -4$ and $A = -19.87\,$rad, three times the Fisher error $\sigma_A = 6.624\,$rad at $\rho = 10$. Then the source is moved closer, $\rho = 10, 20, 40, 80$, with the same A.

Integrating out the phase. The phase ϕ_c is a nuisance with a flat prior on $[0, 2\pi)$, and it can be removed exactly. What probability does the model assign to the data after averaging over ϕ_c? With $\tilde h(\boldsymbol\theta,\phi_c) = \tilde h(\boldsymbol\theta)e^{i\phi_c}$ (the phase enters Ψ as $-\phi_c$), noise-free data $d = h_0$, equal amplitudes $(h|h) = (h_0|h_0) = \rho^2$, and the weight w of (10.72),

$$\begin{aligned}
(d-h\,|\,d-h) &\overset{(1)}{=} (h_0|h_0) + (h|h) - 2(h_0|h) \overset{(2)}{=} 2\rho^2 - 2\rho^2\,\mathrm{Re}\big[Z\,e^{-i\phi_c}\big],\\
Z(\boldsymbol\theta) &\equiv \int w(f)\,e^{i[\Psi(f;\boldsymbol\theta)-\Psi_0(f)]}\,\mathrm{d}f,\\
\mathcal{L}(\boldsymbol\theta) &\overset{(3)}{=} \frac{1}{2\pi}\int_0^{2\pi} e^{-\rho^2+\rho^2|Z|\cos(\phi_c-\arg Z)}\,\mathrm{d}\phi_c \overset{(4)}{=} e^{-\rho^2}\,I_0\big(\rho^2|Z(\boldsymbol\theta)|\big).
\end{aligned} \tag{10.90}$$

(1) Expand the inner product; it is symmetric.
(2) $(h_0|h) = 4\,\mathrm{Re}\int \tilde h_0\tilde h^*/S_n\,\mathrm{d}f$ with $\tilde h_0\tilde h^* = |\tilde h|^2 e^{i(\Psi-\Psi_0)}e^{-i\phi_c}$, and $4|\tilde h|^2/S_n = \rho^2 w$.
(3) Marginalise $e^{-(d-h|d-h)/2}$ over ϕ_c with the flat prior $1/2\pi$; write $Z = |Z|e^{i\arg Z}$.
(4) $\frac{1}{2\pi}\int_0^{2\pi} e^{x\cos\varphi}\mathrm{d}\varphi = I_0(x)$, the modified Bessel function; the shift by $\arg Z$ does not change an integral over a full period.

$|Z| \le 1$ is the overlap of the trial phase with the true one, averaged with the signal-to-noise weight: it is one when the phases agree up to a constant and falls as they drift apart. For large argument $\ln I_0(x) \approx x - \frac12\ln 2\pi x$, so $\ln\mathcal{L} \approx -\rho^2(1-|Z|)$: the likelihood penalises the lost overlap times ρ^2, the same structure as the phase-maximised likelihood of Kim et al. [86, eq. (53)]. The script `15_fisher_vs_mh.py` implements it in three lines:

Listing 10.12: The likelihood with the phase integrated out (`code/ch10/15_fisher_vs_mh.py`)

```
53 def Z(th, psi_true):
54     y = fw * np.exp(1j * (psi(th) - psi_true))
55     return np.sum(0.5 * (y[1:] + y[:-1]) * dlf)
56
57
58 def loglike(th, psi_true, rho):
59     x = rho**2 * np.abs(Z(th, psi_true))
60     return np.log(i0e(x)) + x - rho**2
```

Line 54 is the integrand of Z on the logarithmic grid (`fw` is $f\,w(f)$), line 55 the trapezoid sum, and line 60 is $\ln\mathcal{L} = \ln I_0(x) - \rho^2$, computed as $\ln(e^{-x}I_0(x)) + x$ because $I_0(x)$ itself overflows

for $x = \rho^2 \sim 6400$. The Fisher matrix to compare with is the 5×5 matrix of $(\ln \mathcal{M}, t_c, A, b, \phi_c)$ with ϕ_c removed by its Schur complement, the Gaussian version of the same marginalisation. The chain is the sampler of Chapter 9, with a proposal covariance tuned in three short runs and then 40 000 steps per signal-to-noise ratio; acceptance rates are 0.08865 to 0.3004, and the whole script runs in 156.3 s.

Example: the banana at $\rho = 10$. Figure 10.39 shows the chain against the Fisher ellipses. At $\rho = 10$ the posterior is curved: larger amplitudes $|A|$ pair with flatter exponents b, because a flatter power law must be larger at f_{lo} to produce the same phase drift where the signal-to-noise ratio accumulates. The tail reaches $A \approx -60$ rad and $b \approx -2.5$, far outside the 95% ellipse, and on the other side it runs into the region of steep exponents where the effect is confined to the first few millihertz. Only 28.43% of the samples fall inside the Fisher 68.3% ellipse in the (A, b) plane. The marginal standard deviations from the chain exceed the Fisher errors by factors of 2.163 for A, 1.305 for b, 2.154 for $\ln \mathcal{M}$ and 3.564 for t_c. The chirp mass and the arrival time are dragged along the banana, because each point on it is a different combination of environment and vacuum parameters that fits the same data.

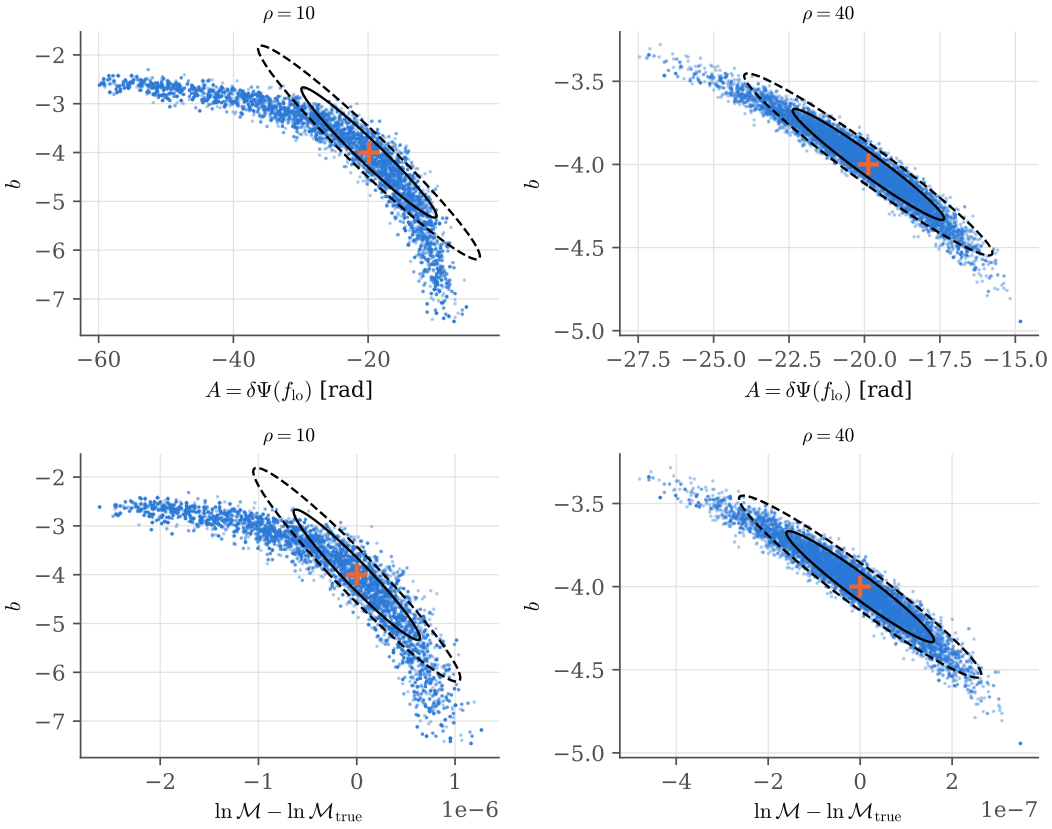


Figure 10.39: Posterior samples from the Metropolis–Hastings chain (dots) against the Fisher 68.3% and 95.4% ellipses (solid, dashed) for a power-law environment with exponent b and phase shift A at the start of the band. Left: at $\rho = 10$ the posterior is a curved banana that the ellipse cannot follow. Right: at $\rho = 40$ it has become an ellipse. Top: environment amplitude against exponent; bottom: chirp mass against exponent. The cross marks the truth.

Example: the approach to the Fisher limit. As ρ grows (Figure 10.40), the ratio of chain to Fisher widths falls towards one: for A it is 2.163, 1.565, 1.172, 1.103 at $\rho = 10, 20, 40, 80$, and for t_c, the worst, 3.564, 1.995, 1.329, 1.117. The fraction of samples inside the 68.3% ellipse climbs

to 50.8, 60.5 and 65.46%. The analogue of Vallisneri's test for this experiment, the difference between the actual fall of $\ln \mathcal{L}$ on the Fisher 1σ surface and the $\frac{1}{2}$ the forecast predicts (median over 200 points), is 0.1471, 0.06673, 0.03061 and 0.01137: it crosses 0.1 between $\rho = 10$ and 20. Here it falls roughly as $1/\rho$, not $1/\rho^2$, because the noise-free drop contains the cross term $\Delta\theta^i(\partial_i h|R)$ of (10.87), which is first order in R.

Notice what the numbers say together. At $\rho = 20$ the median test is already passed, yet the marginal widths are still 20 to 100% larger than the forecast. The median over the 1σ surface hides the directions where the surface bends most, and the marginal widths are set by the tails. This is why Vallisneri asks for the criterion to hold *everywhere* on the surface, not on average, and why he calls it a consistency check rather than a guarantee [193, Sec. VI]. For environmental effects at the signal-to-noise ratios LISA expects for these sources, a signal-to-noise ratio of ten to a few tens, the Fisher error is optimistic by factors of up to two to four, and most so for the vacuum parameters it is correlated with.

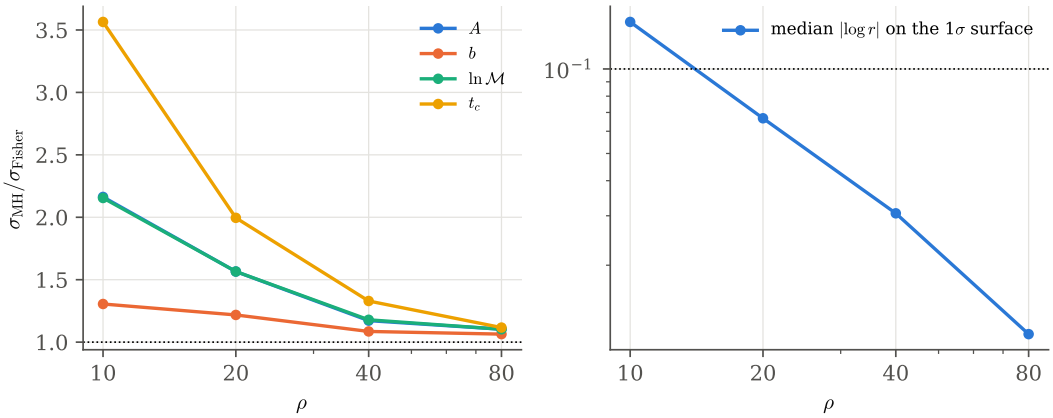


Figure 10.40: Left: marginal standard deviation from the chain divided by the Fisher error, for the four parameters, as the same source is moved closer. All ratios fall towards one, slowest for the arrival time. Right: the median mismatch between the exact and the Fisher log-likelihood on the Fisher 1σ surface; the dotted line is Vallisneri's threshold of 0.1.

Example: a one-sided effect on a vacuum signal. Friction, accretion and migration all take energy from the orbit; none gives it back. Their phase shift therefore has a known sign ($A \le 0$ in our convention), and the right prior is one-sided. Suppose the source is in vacuum, $A = 0$, at the benchmark distance ($\rho = 13.25$), with $b = -4$ known. The Fisher matrix says $\sigma_A = 1.191$ rad, a symmetric Gaussian centred on zero. The posterior cannot be: half of that Gaussian lies at $A > 0$, which the prior forbids. A chain with the prior $A \le 0$ finds a one-sided posterior piled against the boundary, with mean $|A| = 1.168$ rad and a 95% upper limit $|A| < 2.952$ rad, that is $2.479\,\sigma_A$. A half-Gaussian of the Fisher width would give $1.96\,\sigma_A$; most of the excess is the nonlinearity of the previous example at $\rho \approx 13$. The forecast question for a non-detection is "what upper limit would we set?", and the Fisher error answers it only after the prior boundary and the nonlinearity have been put in by hand.

Watch out: A Fisher ellipse knows nothing about boundaries

The Fisher matrix is computed from derivatives at one point and assumes every direction is open. A parameter with a physical boundary (a density that cannot be negative, a dissipative effect of fixed sign, a slope confined to a range by the formation history) has a posterior that the ellipse cannot represent near the boundary, and the published posteriors of Kim et al. show such skewed marginals, with long one-sided tails, for ρ_6 and $\log_{10} a$.

Where the Fisher forecast fails, in one list. The examples above and those of Section 10.12.3 show four distinct mechanisms. Low signal-to-noise ratio makes the signal surface bend within the noise (the banana). Near-degenerate parameters stretch the 1σ ellipsoid into the nonlinear region even at moderate ρ (Kim's ρ_6–$\gamma_{\rm sp}$). Prior boundaries cut the posterior (one-sided effects). And separate parameter regions that fit equally well, which a single expansion point cannot see, give multimodal posteriors; none appeared here, but they are common when the environment can mimic a change of masses. Each is visible to a sampler and invisible to the matrix.

How this is used in real research: Verification here, the instrument in research

Above, the chain verifies the forecast on a four-parameter problem where both can be computed in a minute. In a research analysis the roles reverse: the posterior is computed directly, by nested sampling (Kim et al. use `dynesty` [204]) or by parallel-tempered ensemble samplers built for LISA such as `Eryn` [205], with the full set of 15 or more source parameters, and the Fisher matrix is used before that, to design the proposal distributions, to choose parameters that are nearly uncorrelated, and to locate the degeneracies a sampler will have to cross. The full version of this experiment runs a few million likelihood calls per source on a cluster, with the noise realisation drawn rather than set to zero; ours runs about 3×10^5 calls in under a minute and therefore shows only the median experiment.

10.14 How it is done in research

A collaboration asking "will LISA see the dark-matter spike around this kind of binary?" runs the same chain of reasoning as we did: a likelihood for data that are signal plus Gaussian noise, a detection statistic, a Fisher matrix for a first answer, a sampler for the real one. Every box is larger. The stages, in the order a real analysis passes through them, are drawn in Figure 10.41; at each one the laptop version made a reduction, and each reduction has a cost that can be named.

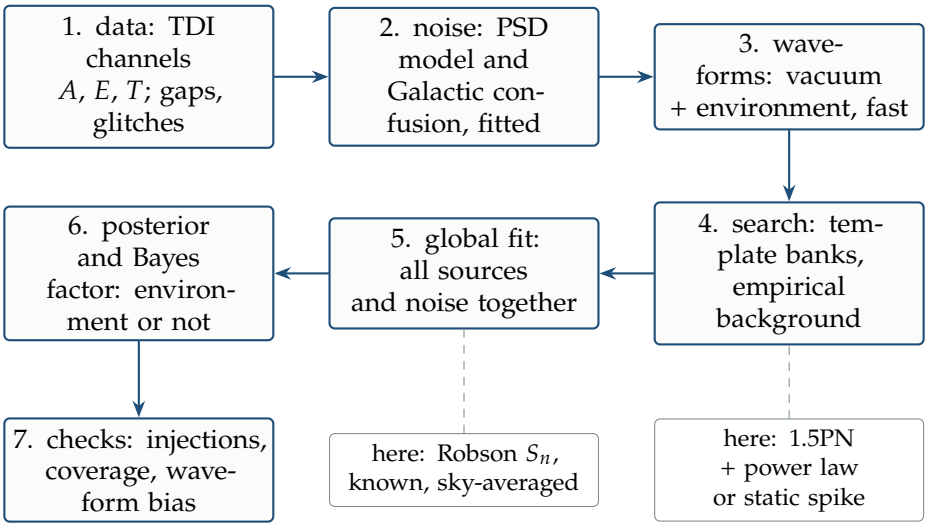


Figure 10.41: The stages of a research analysis of an environmental effect in a LISA source (blue boxes, in order), and the two places where the laptop version cuts most deeply (grey). In a real analysis stages 2 to 5 are not separate: the noise, the resolved sources and the search are iterated inside one global fit.

1. The data. LISA does not record a strain (the fractional change of arm length that a wave causes, Section T2.3). Each spacecraft measures the phase of the laser light arriving from the other two, and the raw measurements are dominated by laser frequency noise many orders of magnitude above any gravitational wave. Time-delay interferometry combines the measurements, delayed by the light travel times along the arms, into channels in which the laser noise cancels [206]; the usual choice is three channels A, E, T whose noises are nearly uncorrelated, with T almost blind to gravitational waves at low frequency. The data come with gaps (antenna repointing, maintenance) and glitches. The community prepares for this with simulated data sets of increasing realism, the LISA Data Challenges [207]. *Here:* a single sky-averaged strain channel in the frequency domain, without gaps. The full analysis keeps the detector response, the antenna patterns F_+, $F_\times$ of (T2.34) that weight the two polarisations h_+, $h_\times$, which carries the sky position and orientation; the reduction loses those parameters and replaces the source-dependent response by its average (Eqs. (T2.36) and (T2.37)), which can be off by tens of per cent for an individual source.

2. The noise. The noise spectrum is not known in advance. It is described by a model with a few parameters for each instrumental component, and fitted together with the signals; with few independent averages the unknown spectrum is integrated out, which turns the Gaussian likelihood into a Student-t [201] as in Section 10.8.6, and glitches are modelled rather than cut [202]. The Galactic confusion noise is the part of the white-dwarf population that cannot be resolved (Section T2.7); it changes during the mission as more binaries are resolved, and it is modulated through the year as the constellation's orientation changes, so the noise is not stationary. *Here:* the Robson et al. spectrum with its four-year confusion fit, known exactly and stationary. A forecast loses nothing by this if the real spectrum matches the fit; an analysis of real data with a fixed spectrum would quote errors that are too small, as Figure 10.23 measured for few averages.

3. The waveforms. The vacuum waveform of an intermediate- or extreme-mass-ratio inspiral is computed from black-hole perturbation theory, not from a 1.5PN expansion (Section T2.6), because the companion completes 10^5 to 10^6 cycles in the band and every small phase error accumulates; fast surrogate codes make this affordable inside a sampler [208]. The environment is then added either as a physical model, such as a dark-matter spike that is heated and thinned by the energy friction deposits in it [89, 209], or agnostically as the ppE power law of Section 10.10 (Eq. (T2.69)). *Here:* the 1.5PN restricted waveform, a static spike, and the power law. The full analysis must also ask whether the model is accurate enough: a waveform error δh shifts the best fit by

$$\delta\theta^i = (\Gamma^{-1})^{ij}\,(\partial_j h \,|\, \delta h), \tag{10.91}$$

the bias formula (10.37) of Section 10.7 with $\mathbf{C} = S_n$ [210]. A model is good enough when this bias is small compared with the statistical error, which requires roughly $(\delta h|\delta h) \lesssim 1$: the waveform error must be below the noise, not merely small. The static spike fails this test by a wide margin for the binaries of Section 10.12, since the heated spike dephases far less [89, 209]; the cost of the reduction is that our environmental amplitudes are those of an idealised model, while the detectability map of Section 10.11.5, which is stated in terms of β and b, is unaffected.

4. The search. The matched filter of Section 10.8.3 is the most powerful test only for a known template, by the Neyman–Pearson lemma. With unknown parameters, a search maximises the filter over a bank of templates: this is a generalised likelihood ratio, and its distribution under noise alone is not a textbook distribution. Real searches therefore measure it empirically; ground-based detectors build the background by shifting the data of one detector in time

against another by more than the light travel time, which destroys any real coincidence and keeps the noise [83, 211]. The phase and the arrival time are handled exactly as in Section 10.8.4, by two quadratures and one fast Fourier transform [200]. *Here:* a single template with known masses, and a look-elsewhere factor measured from 2000 simulations. The full search costs many templates and an empirical background; the reduction overstates the significance of a weak signal found by searching.

5. The global fit. LISA never records a stretch of data without signals: tens of thousands of resolvable Galactic binaries, massive black-hole mergers and extreme-mass-ratio inspirals overlap in time and frequency throughout the mission. "Noise alone" is therefore the idealised null hypothesis of a single-source problem. The research analysis fits all the sources and the noise jointly, with the number of sources itself a parameter, and iterates as the data grow [203]. *Here:* one source in Gaussian noise. The cost is that confusion between sources, and the extra uncertainty it adds to each, is absent from all our errors.

6. The posterior and the decision. Parameters are estimated by sampling the full likelihood, with the detector response and the noise parameters, using nested sampling [204] or tempered ensemble samplers [205] through general frameworks such as `bilby` [212]; the likelihood is made cheap enough by expanding it around a reference waveform (relative binning, [213]). Whether an environment is present is a model comparison: the evidence of the vacuum model against that of the environmental one [89], or of particle against wave dark matter [86]. *Here:* a four-parameter Metropolis–Hastings chain on noise-free data; the posterior width is right for that reduced problem but the model comparison is not done.

7. Checks. Signals are injected into many simulated noise realisations and recovered; the fraction of injections whose true parameters fall inside the nominal $x\%$ credible region must be $x\%$ (a coverage test, as in Section 10.8.6). The bias (10.91) is evaluated for the known approximations of the waveform. And the Fisher matrix is checked against the posterior with Vallisneri's criterion before any forecast is quoted (Section 10.13). *Here:* the comparison of the chain with the ellipse at four signal-to-noise ratios.

Recap

- Detection is inference between $H_0 : d = n$ and $H_1 : d = n + h(\boldsymbol{\theta})$. Only one noise realisation, $n = d - h$, turns a prediction into the data, so $p(\boldsymbol{d} \mid \boldsymbol{\theta}) = p_n(d - h)$; for stationary Gaussian noise $\ln p_n(n) = -\frac{1}{2}(n|n) + \text{const}$ with $(a|b) = 4\,\mathrm{Re}\int \tilde{a}\tilde{b}^*/S_n \,\mathrm{d}f$.
- The likelihood ratio is $(d|h) - \frac{1}{2}(h|h)$: the matched filter. Under the two hypotheses $z = (d|h)/\rho$ is $\mathcal{N}(0,1)$ and $\mathcal{N}(\rho,1)$, $\rho^2 = (h|h)$. An unknown phase costs a second quadrature, an unknown time one FFT, and both raise the threshold (look-elsewhere). An unknown spectrum, integrated out, gives a Student-t.
- $\Gamma_{ij} = (\partial_i h|\partial_j h)$ is the Gaussian Fisher formula with $\mathbf{C} = S_n$. The amplitude decouples ($\sigma_{\ln \mathcal{A}} = 1/\rho$); the phase block is $\rho^2\langle \partial_i \Psi\, \partial_j \Psi\rangle_w$; a parameter is measured by the part $g_\perp$ of its phase signature that the others cannot imitate, $\sigma = 1/(\rho\|g_\perp\|_w)$.
- Any environment with $P_{\mathrm{env}}/P_{\mathrm{GW}} \propto f^{-n}$ adds $\delta\Psi = \beta u^b$, $b = -\frac{5}{3} - n$; friction in a spike $\rho \propto r^{-\gamma}$ has $b = (2\gamma - 16)/3$. One Fisher matrix per b gives a detectability map; the error diverges where u^b is a vacuum signature ($b = -\frac{5}{3}, 0, 1$), and low-frequency effects separate from the chirp mass.
- Reproduced: the Robson et al. LISA curve and its worked numbers, Cutler and Flanagan's

stage	research version	laptop version (here)	what the reduction costs
data	TDI A, E, T time series, response, gaps	one sky-averaged channel, frequency domain	no sky parameters; per-source response off by tens of per cent
noise	fitted PSD, evolving confusion, glitches	Robson S_n, four-year confusion, known	errors too small if the spectrum is not known (Figure 10.23)
waveform	perturbation-theory inspirals, heated spikes	1.5PN, static spike, ppE power law	absolute dephasing overstated; detectability in (b, β) unaffected
search	template banks, empirical background	one template, simulated look-elsewhere	significance of searched signals overstated
sources	global fit of all sources and noise	one source	no source confusion in the errors
inference	full-likelihood sampling, Bayes factors	Fisher matrix; MH on 4 parameters, ~ 40 s	widths right for the reduced problem only; no evidence
checks	injections over noise realisations	noise-free (median) experiment	no scatter of the best fit, no coverage test of the full chain

Table 10.7: Each reduction made in the laptop version of the gravitational-wave forecast, what a research analysis does instead, and what the reduction costs.

Table I (within 2.33%), Eda et al.'s dephasing and errors (with $\ln \Lambda = \ln 10^3$, and their amplitude at its own pivot), Kim et al.'s dephasing, signal-to-noise ratio and, within a factor 1.4 to 5.5, their posterior widths.

- The Fisher matrix is the posterior of the linearised signal. Trust it where $\frac{1}{2}(R|R) \lesssim 0.1$ on its own 1σ surface; otherwise sample. At $\rho = 10$ the chain is up to 3.564 times wider than the ellipse, at $\rho = 80$ within about 10%; degeneracies and one-sided priors break it at any ρ.

10.15 How small a twist can the sky reveal?

Section 10.6 forecast the six numbers of ΛCDM from the temperature and E-mode spectra of a Planck-like experiment, and the forecast came close to what Planck delivered. Every parameter there did the same kind of thing: it changed the size of a spectrum that is already large, C_ℓ^{TT}, C_ℓ^{TE} or C_ℓ^{EE}, and the Fisher matrix weighed that change against the scatter of the measured spectrum. The birefringence part of Chapter T1 then introduced a parameter of a different kind. Light from the last-scattering surface may have its plane of polarization turned by a small angle β on the way to us. Picture the polarization at each point of the sky as a short headless stick. The ordinary sky makes the sticks line up in spoke-and-ring patterns (E) and almost never in swirls (B); a uniform twist turns every stick by β, so a little of every spoke pattern becomes a swirl, and that new swirl remembers the spoke it came from. In the spectra this shows up as a correlation between E and B, and between T and B, which the untwisted sky does not have at all (Section T1.11.2).

That is where the earlier forecast stops being enough. At $\beta = 0$ the twist changes none of the spectra the ΛCDM forecast used, to first order: EE and BB move by an amount proportional to β^2, and TT does not move at all. Its whole first-order effect sits in two spectra, EB and TB, whose expected value in the fiducial model is exactly zero. A spectrum that is zero on average

still scatters from sky to sky, and how much it scatters is not obvious: it is not set by its own size, which is zero, but by the sizes of the other spectra. The forecast must therefore be built from the full covariance of the three fields T, E, B together.

A fence knocked off true. A surveyor wants to know whether a long straight fence has been knocked a tiny angle off its line. Measuring the angle at the near end is hopeless; the angle is far too small for her protractor. Instead she walks to the far end and measures how far the last post sits to the side of the true line. The sideways shift is the length of the fence times the angle, so a longer fence turns the same small angle into a larger shift. What limits her is not the length but the sideways wobble: the posts were never perfectly in line, and her tape shakes in the wind. The angle she can detect is roughly the wobble divided by the length, and every extra post she checks gives her one more independent reading. One thing she should not do is look at how much the posts wobble in total: a small knock adds only (length × angle)^2 to the wobble's square, which is lost among the posts' own irregularities long before the shift itself is.

The polarized sky is that fence. The bright E pattern is its length. The twist moves a fraction 2β of E into the sideways direction, B. The wobble is whatever B was there anyway: the swirls made by gravitational lensing (Section T1.11.4) and the noise of the instrument. Each mode of the sky, each (ℓ, m), is one post. In one sentence: *the information on a small twist is the E power divided by the B wobble, per mode, added up over all the modes we can measure.*

The fence in symbols. Take one mode with a large E of variance C_E, a pre-existing B of variance C_B, and noise of variance N in each. After a small twist β (in radians),

$$E^{\mathrm{o}} = E + n_E, \qquad B^{\mathrm{o}} = 2\beta E + B + n_B, \tag{10.92}$$

from the rotation rule $B' = \sin 2\beta\, E + \cos 2\beta\, B \approx 2\beta E + B$ of Eq. (T1.50). The dials are the lever C_E, the wobble $C_B + N$, and the number n of modes. The naive choice is to look at the B power. Its average is $\mathbb{E}\,|B^{\mathrm{o}}|^2 = 4\beta^2 C_E + C_B + N$, so the twist adds only $4\beta^2 C_E$ to it, and the derivative of this with respect to β vanishes at $\beta = 0$: a measurement of B power contains no information on β to first order. The good choice is the product $E^{\mathrm{o}}B^{\mathrm{o}}$. Its average is $\mathbb{E}[E^{\mathrm{o}}B^{\mathrm{o}}] = 2\beta C_E$, linear in β, because E, B, n_E, n_B are independent with mean zero and only the term $2\beta E \cdot E$ survives. Its variance at $\beta = 0$ is $\mathbb{E}[(E + n_E)^2(B + n_B)^2] = (C_E + N)(C_B + N)$, the product of two independent variances. So one mode measures β with variance $(C_E + N)(C_B + N)/(4C_E^2)$, and n independent modes divide that by n:

$$\sigma^2(\beta) \overset{(1)}{=} \frac{(C_E + N)(C_B + N)}{4\,n\,C_E^2} \overset{(2)}{=} \frac{C_B + N}{4\,n\,C_E} \quad (N \ll C_E). \tag{10.93}$$

(1) The estimate $\hat\beta = \overline{E^{\mathrm{o}}B^{\mathrm{o}}}/(2C_E)$ is the average product over n modes divided by its slope $2C_E$; dividing a quantity by a constant divides its variance by the square of the constant.

(2) When E is measured much better than its noise, $C_E + N \approx C_E$ and one factor of C_E cancels.

The second form is the surveyor's rule: wobble over length, divided by the number of posts. Put numbers in. Let the wobble be one hundredth of the lever, $C_B + N = 0.01\,C_E$, and let there be $n = 1000$ modes. Then $\sigma(\beta) = \sqrt{0.01/4000} = 1.6 \times 10^{-3}$ rad, or $0.09°$, so a twist of $0.35°$ (6.1×10^{-3} rad) stands out at about 3.9σ. The B power of the same modes would grow by $4\beta^2 C_E = 1.5 \times 10^{-4}\,C_E$, while the average of n squared Gaussian values has a relative scatter of $\sqrt{2/n}$ (Chapter 7), so the measured B power itself scatters by $\sqrt{2/1000} \times 0.01\,C_E = 4.5 \times 10^{-4}\,C_E$: the twist would appear in the power at a third of a standard deviation. Lesson: the twist is measured in the cross product, with an error set by the ratio of what is in B anyway to what is in E; its square in the B power is useless at small angles.

A picture: fitting a line through the origin. Section T1.11 drew every mode as a point with its E across and its B up: a thin horizontal cigar, which a twist tilts, and that tilt is why the correlation, not the power, identifies a twist. The same picture also says how well the tilt can be measured. By (10.92) the points scatter around the straight line $B = 2\beta E$ through the origin, so measuring β is fitting the slope of that line, and $\hat{\beta} = \sum E^{\rm o}B^{\rm o}/(2\sum (E^{\rm o})^2)$ is exactly least squares through the origin (Section 3.10.2). Everyone who has fitted a line knows the two dials: the vertical scatter of the points, which is the wobble $C_B + N$, and how far the points spread along the horizontal axis, which is the lever $\sum E^2 \approx nC_E$. A slope is pinned when the points reach far out and sit close to the line.

One idea, many names. Measuring a tiny turn of a bright quantity by looking in the direction at right angles to it is the principle of quadrature readout: a gravitational-wave interferometer runs close to a dark fringe so that a small change of arm length appears as light in an output that is otherwise nearly empty, and the noise that matters is the noise of that empty port, not the brightness of the beam. In statistics it is a slope estimated by regression through the origin, and in the language of the Fisher matrix it is the information about a rotation, which we now compute for the real T, E, B sky. The computation answers three questions: which multipoles carry the information, how the error depends on the noise, the beam, the multipole range and the sky fraction, and how close a Planck-like experiment comes to the errors that were actually published.

10.15.1 The information on an angle, from the covariance of T, E and B

The data. At each multipole ℓ and each m the experiment measures three harmonic coefficients, $T_{\ell m}$, $E_{\ell m}$, $B_{\ell m}$ (Eq. (T1.39)), each the sky plus white noise with the beam divided out. They are Gaussian with mean zero, and different (ℓ, m) are independent, exactly as the temperature $a_{\ell m}$ were in Section 10.6.1. What is new is that at one (ℓ, m) we now have three correlated numbers instead of one or two. Their covariance is the 3×3 matrix of rotated spectra of Section T1.11.2, plus the noise N_ℓ^T in temperature and N_ℓ^P in each of E and B:

$$\mathbf{C}_\ell(\beta) = \begin{pmatrix} C_\ell^{TT} + N_\ell^T & \cos 2\beta\, C_\ell^{TE} & \sin 2\beta\, C_\ell^{TE} \\ \cos 2\beta\, C_\ell^{TE} & c^2 C_\ell^{EE} + s^2 C_\ell^{BB} + N_\ell^P & \tfrac{1}{2}\sin 4\beta\, \Delta_\ell \\ \sin 2\beta\, C_\ell^{TE} & \tfrac{1}{2}\sin 4\beta\, \Delta_\ell & s^2 C_\ell^{EE} + c^2 C_\ell^{BB} + N_\ell^P \end{pmatrix}, \tag{10.94}$$

with $c = \cos 2\beta$, $s = \sin 2\beta$ and $\Delta_\ell \equiv C_\ell^{EE} - C_\ell^{BB}$; the spectra on the right are those of the untwisted sky, which has no EB and no TB. The noise enters E and B with the same N_ℓ^P because white noise in Q and U has no preferred orientation, so it is not twisted by anything.

The formula to use. Equation (10.32) is the Fisher matrix for exactly this situation, a zero-mean Gaussian vector at each (ℓ, m) with an ℓ-dependent covariance; it was written there for the pair (T, E), but nothing in its derivation used the size of the matrix. With one parameter,

$$F_{\beta\beta} = \sum_\ell \frac{(2\ell+1) f_{\rm sky}}{2}\, {\rm Tr}\big[\mathbf{C}_\ell^{-1}\partial_\beta \mathbf{C}_\ell\, \mathbf{C}_\ell^{-1}\partial_\beta \mathbf{C}_\ell\big], \tag{10.95}$$

where $(2\ell+1)f_{\rm sky}$ counts the independent m-modes of a sky fraction $f_{\rm sky}$ (the argument of Section 10.6.1).

The derivative at $\beta = 0$. Differentiate each entry of (10.94) and set $\beta = 0$. The diagonal entries contain $\cos^2 2\beta$ and $\sin^2 2\beta$, whose derivatives $\mp 2\sin 4\beta$ vanish there; the TE entry contains $\cos 2\beta$, whose derivative $-2\sin 2\beta$ also vanishes. Only the two entries that were zero survive: $\partial_\beta \sin 2\beta = 2\cos 2\beta \to 2$ and $\partial_\beta \tfrac{1}{2}\sin 4\beta = 2\cos 4\beta \to 2$. Write $b_\ell \equiv 2C_\ell^{TE}$ and $e_\ell \equiv 2\Delta_\ell$, gather the (T, E) block of the covariance at $\beta = 0$ into a 2×2 matrix $\boldsymbol{A}_\ell$ and the B entry into one number

$\tilde{B}_\ell \equiv C_\ell^{BB} + N_\ell^P$:

$$\mathbf{C}_\ell(0) = \begin{pmatrix} \boldsymbol{A}_\ell & \mathbf{0} \\ \mathbf{0}^\mathsf{T} & \tilde{B}_\ell \end{pmatrix}, \quad \boldsymbol{A}_\ell = \begin{pmatrix} C_\ell^{TT} + N_\ell^T & C_\ell^{TE} \\ C_\ell^{TE} & C_\ell^{EE} + N_\ell^P \end{pmatrix}, \quad \partial_\beta \mathbf{C}_\ell = \begin{pmatrix} \mathbf{0} & \boldsymbol{v}_\ell \\ \boldsymbol{v}_\ell^\mathsf{T} & 0 \end{pmatrix}, \quad \boldsymbol{v}_\ell = \begin{pmatrix} b_\ell \\ e_\ell \end{pmatrix}. \tag{10.96}$$

The trace. The covariance is block diagonal, so its inverse is too, with $\boldsymbol{A}_\ell^{-1}$ and $1/\tilde{B}_\ell$ on the diagonal. Then

$$\begin{aligned} \mathbf{C}_\ell^{-1}\partial_\beta \mathbf{C}_\ell &\overset{(1)}{=} \begin{pmatrix} \boldsymbol{A}_\ell^{-1} & \mathbf{0} \\ \mathbf{0}^\mathsf{T} & \tilde{B}_\ell^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{0} & \boldsymbol{v}_\ell \\ \boldsymbol{v}_\ell^\mathsf{T} & 0 \end{pmatrix} = \begin{pmatrix} \mathbf{0} & \boldsymbol{A}_\ell^{-1}\boldsymbol{v}_\ell \\ \boldsymbol{v}_\ell^\mathsf{T}/\tilde{B}_\ell & 0 \end{pmatrix}, \\ \left(\mathbf{C}_\ell^{-1}\partial_\beta \mathbf{C}_\ell\right)^2 &\overset{(2)}{=} \begin{pmatrix} \boldsymbol{A}_\ell^{-1}\boldsymbol{v}_\ell\boldsymbol{v}_\ell^\mathsf{T}/\tilde{B}_\ell & \mathbf{0} \\ \mathbf{0}^\mathsf{T} & \boldsymbol{v}_\ell^\mathsf{T}\boldsymbol{A}_\ell^{-1}\boldsymbol{v}_\ell/\tilde{B}_\ell \end{pmatrix}, \\ \mathrm{Tr}\left[\left(\mathbf{C}_\ell^{-1}\partial_\beta \mathbf{C}_\ell\right)^2\right] &\overset{(3)}{=} \frac{\mathrm{Tr}(\boldsymbol{A}_\ell^{-1}\boldsymbol{v}_\ell\boldsymbol{v}_\ell^\mathsf{T}) + \boldsymbol{v}_\ell^\mathsf{T}\boldsymbol{A}_\ell^{-1}\boldsymbol{v}_\ell}{\tilde{B}_\ell} \overset{(4)}{=} \frac{2\,\boldsymbol{v}_\ell^\mathsf{T}\boldsymbol{A}_\ell^{-1}\boldsymbol{v}_\ell}{\tilde{B}_\ell}. \end{aligned} \tag{10.97}$$

(1) Multiply the blocks: the top-left block of $\mathbf{C}^{-1}$ meets the top-right block $\boldsymbol{v}_\ell$ of the derivative, the bottom-right number meets the bottom-left row $\boldsymbol{v}_\ell^\mathsf{T}$.
(2) Square the result block by block: (top-right)(bottom-left) gives a 2×2 matrix, (bottom-left)(top-right) a single number; the off-diagonal blocks of the square are products with a zero block.
(3) The trace of a block-diagonal matrix is the sum of the traces of its blocks.
(4) The cyclic property, $\mathrm{Tr}(\boldsymbol{A}^{-1}\boldsymbol{v}\boldsymbol{v}^\mathsf{T}) = \mathrm{Tr}(\boldsymbol{v}^\mathsf{T}\boldsymbol{A}^{-1}\boldsymbol{v}) = \boldsymbol{v}^\mathsf{T}\boldsymbol{A}^{-1}\boldsymbol{v}$, since the last expression is a single number.

Insert this into (10.95): the factor $\frac{1}{2}$ cancels the 2, and

$$\boxed{F_{\beta\beta} = \sum_\ell (2\ell+1) f_{\rm sky}\, \frac{\boldsymbol{v}_\ell^\mathsf{T}\boldsymbol{A}_\ell^{-1}\boldsymbol{v}_\ell}{C_\ell^{BB} + N_\ell^P} = \sum_\ell (2\ell+1) f_{\rm sky}\, \frac{b_\ell^2\tilde{E}_\ell - 2b_\ell e_\ell C_\ell^{TE} + e_\ell^2\tilde{T}_\ell}{(\tilde{T}_\ell\tilde{E}_\ell - (C_\ell^{TE})^2)\,\tilde{B}_\ell}} \tag{10.98}$$

with $\tilde{T}_\ell = C_\ell^{TT} + N_\ell^T$, $\tilde{E}_\ell = C_\ell^{EE} + N_\ell^P$, and the second form written out with the inverse of a 2×2 matrix, $\boldsymbol{A}^{-1} = \begin{pmatrix} \tilde{E} & -C_\ell^{TE} \\ -C_\ell^{TE} & \tilde{T} \end{pmatrix}/\det \boldsymbol{A}$. The error on the angle is $\sigma(\beta) = F_{\beta\beta}^{-1/2}$. The script `code/ch10/17_biref_fisher.py` also evaluates (10.95) numerically, differentiating the full matrix (10.94) by central differences; the two agree to 4.9×10^{-8} in relative terms.

Example: EB alone, and the variance of a cross-spectrum. Drop the temperature: the data at each (ℓ, m) are only (E, B), $\boldsymbol{A}_\ell$ shrinks to the number $\tilde{E}_\ell$ and $\boldsymbol{v}_\ell$ to e_ℓ, and

$$F_{\beta\beta}^{EB} = \sum_\ell (2\ell+1) f_{\rm sky}\, \frac{4\Delta_\ell^2}{\tilde{E}_\ell\tilde{B}_\ell} = \sum_\ell \frac{\left(\partial_\beta C_\ell^{EB}\right)^2}{\mathrm{Var}\,\widehat{C}_\ell^{EB}}, \qquad \mathrm{Var}\,\widehat{C}_\ell^{EB} = \frac{\tilde{E}_\ell\tilde{B}_\ell}{(2\ell+1) f_{\rm sky}}. \tag{10.99}$$

The second form reads the same number as "squared response over variance" of the measured spectrum, the reading of Section 10.6.1, and it identifies the variance of a measured EB when the true EB vanishes. That variance follows directly. On the full sky $\widehat{C}_\ell^{EB}$ is an average over $2\ell+1$ real modes of the product $E\,B$ (Eq. (T1.39)), and for each mode E and B are independent with variances $\tilde{E}_\ell$ and $\tilde{B}_\ell$, so the product has variance $\tilde{E}_\ell\tilde{B}_\ell$; an average of $2\ell+1$ independent copies divides it by $2\ell+1$. In general, for Gaussian fields, the covariance of two measured spectra is $\mathrm{Cov}(\widehat{C}_\ell^{XY}, \widehat{C}_\ell^{ZW}) = (\tilde{C}_\ell^{XZ}\tilde{C}_\ell^{YW} + \tilde{C}_\ell^{XW}\tilde{C}_\ell^{YZ})/(2\ell+1)$, by Isserlis' theorem for the four-point average (Section 7.2.1); this is the formula of [76], Appendix A, and of [80], eq. (21), and with $X = Z = E$, $Y = W = B$ and $\tilde{C}^{EB} = 0$ it gives (10.99). The one-channel forecast of Section T1.13.3 used this variance as an input; here it is a consequence of the Gaussian likelihood.

Example: when does TB help? With only (T, B), the same steps give the information $F_\ell^{TB} = (2\ell+1) f_{\rm sky}\, 4(C_\ell^{TE})^2/(\tilde{T}_\ell \tilde{B}_\ell)$ from one multipole. Divide it by the EB information of the same multipole:

$$\frac{F_\ell^{TB}}{F_\ell^{EB}} \overset{(1)}{=} \frac{(C_\ell^{TE})^2\, \tilde{E}_\ell}{\tilde{T}_\ell\, \Delta_\ell^2} \overset{(2)}{=} \rho_\ell^2\, \frac{\tilde{E}_\ell}{C_\ell^{EE}} \quad (N_\ell^T \ll C_\ell^{TT},\ C_\ell^{BB} \ll C_\ell^{EE}), \qquad \rho_\ell \equiv \frac{C_\ell^{TE}}{\sqrt{C_\ell^{TT} C_\ell^{EE}}}. \tag{10.100}$$

(1) The common factors $(2\ell+1) f_{\rm sky} \cdot 4$ and $\tilde{B}_\ell$ cancel.
(2) Neglect the temperature noise and the B power next to the E power, and multiply and divide by C_ℓ^{EE}.

The correlation coefficient ρ_ℓ between temperature and E is never larger than one in size, and on the acoustic scales its square is well below one. So TB beats EB only at multipoles where the E measurement is swamped by its own noise, $\tilde{E}_\ell/C_\ell^{EE} > 1/\rho_\ell^2$: there the temperature, measured almost without noise, serves as a cleaner copy of the E pattern through their correlation. In the opposite limit, with no noise at all, (10.98) collapses to the EB answer. Set $N = 0$ and $C_\ell^{BB} \ll C_\ell^{EE}$ in the numerator: $b^2\tilde{E} - 2be\, C_\ell^{TE} + e^2\tilde{T} = 4(C_\ell^{TE})^2 C_\ell^{EE} - 8(C_\ell^{TE})^2 C_\ell^{EE} + 4(C_\ell^{EE})^2 C_\ell^{TT} = 4C_\ell^{EE}\big(C_\ell^{TT} C_\ell^{EE} - (C_\ell^{TE})^2\big)$, and the bracket cancels the determinant, leaving $F_\ell = (2\ell+1) f_{\rm sky}\, 4C_\ell^{EE}/\tilde{B}_\ell$, exactly F_ℓ^{EB} in the same limit. A perfectly measured E already contains everything the temperature knows about it.

Example: what EE and BB add. At $\beta = 0$ they add nothing, because their derivatives vanish. At $\beta \neq 0$ they respond, $\partial_\beta C^{EE\prime} = -2 \sin 4\beta\, \Delta$, and one might expect a little extra information. There is none, for a reason worth knowing. Write $\boldsymbol{R}$ for the matrix that turns (E, B) by 2β and leaves T alone. By Eq. (T1.50) the twisted coefficients are $\boldsymbol{R}$ times the untwisted ones, and the noise is unchanged by $\boldsymbol{R}$ in distribution, so $\mathbf{C}_\ell(\beta) = \boldsymbol{R}\, \mathbf{C}_\ell(0)\, \boldsymbol{R}^{\mathsf{T}}$. Then $\partial_\beta \mathbf{C}_\ell(\beta) = \boldsymbol{R}\left[\boldsymbol{J}\mathbf{C}_\ell(0) + \mathbf{C}_\ell(0)\boldsymbol{J}^{\mathsf{T}}\right]\boldsymbol{R}^{\mathsf{T}}$ with $\boldsymbol{J} = \boldsymbol{R}^{\mathsf{T}}\partial_\beta \boldsymbol{R}$ a fixed matrix, and $\boldsymbol{R}^{\mathsf{T}} = \boldsymbol{R}^{-1}$, so in $\mathrm{Tr}[\mathbf{C}^{-1}\partial\mathbf{C}\,\mathbf{C}^{-1}\partial\mathbf{C}]$ every $\boldsymbol{R}$ meets an $\boldsymbol{R}^{-1}$ and cancels. The information is the same at every angle. Turning all the sticks by a known amount changes nothing about how well a further turn can be measured, so as β grows the information moves out of EB and TB into EE and BB, and the total stays fixed. The script finds the total at $\beta = 0.35°$ equal to that at $\beta = 0$ to 4.8×10^{-8}, with the share carried by EE and BB alone at that angle only 9.4×10^{-4}. For any angle of interest, EB and TB are where the information is.

Key idea: The information on a twist

Each (ℓ, m) mode contributes $\boldsymbol{v}^{\mathsf{T}}\boldsymbol{A}^{-1}\boldsymbol{v}/\tilde{B}$ to the Fisher information on β, with $\boldsymbol{v} = (2C^{TE}, 2(C^{EE} - C^{BB}))$ the response of TB and EB, $\boldsymbol{A}$ the (T, E) covariance and $\tilde{B} = C^{BB} + N^P$ the B power already present. For a well-measured E this is $4C^{EE}/\tilde{B}$: the E power is the lever, the B power and noise are the floor, and the B power itself carries no information at small angles.

10.15.2 Which multipoles carry the information

Read (10.98) one multipole at a time, in its well-measured-E form, $F_\ell \approx (2\ell+1) f_{\rm sky}\, 4C_\ell^{EE}/(C_\ell^{BB} + N_\ell^P)$. Three factors compete: the number of modes $2\ell+1$, which grows with ℓ; the lever C_ℓ^{EE}, which falls with ℓ apart from its acoustic peaks; and the floor $C_\ell^{BB} + N_\ell^P$, which is the lensing B power where the noise is low and the noise where it is high. The noise after the beam has been divided out is $N_\ell^P = \Delta_P^2 e^{\ell(\ell+1)\sigma_b^2}$ (Eq. (10.30)), with Δ_P the depth of the polarization map

and $\sigma_b = \mathrm{FWHM}/\sqrt{8\ln 2}$ the width of the beam; it is flat until the beam scale and then grows exponentially.

Figure 10.42 shows the ingredients and the result for two experiments. The Planck-like one combines the three Planck HFI channels used for the CMB, 100, 143 and 217 GHz, with the noise levels and beams of [186], Table 4, by inverse-variance weighting at each ℓ (as in Section 10.6.3); its polarization depth is 52.3 μK-arcmin on large scales. Its noise lies above the lensing BB at every multipole, so the floor is pure noise and the information follows $(2\ell+1)C_\ell^{EE}$ until the beam cuts it: it peaks at $\ell \approx 373$, half of it comes from multipoles below 457, and only 5.1% from $\ell < 200$. The deep experiment, 2 μK-arcmin with a 3′ beam, has noise below the lensing floor out to $\ell \approx 2000$. Its information then traces the acoustic peaks of C_ℓ^{EE} divided by the smooth lensing BB, and it peaks near $\ell \approx 992$.

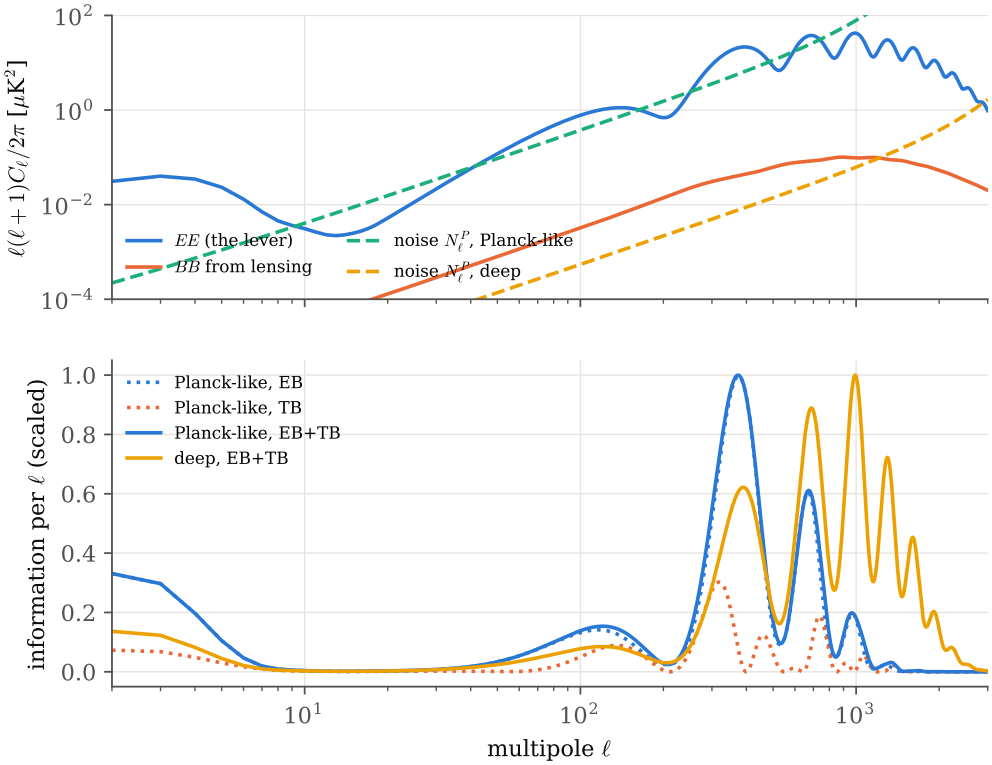


Figure 10.42: Where the information on a twist comes from. Top: the lever and the floor. The E-mode spectrum (blue) is what the twist turns into B; the lensing B modes (orange) and the noise (dashed: Planck-like and deep) are the B that is there anyway. Bottom: the Fisher information on β per multipole, (10.98), scaled to the peak of each experiment's combined EB+TB curve. For the Planck-like experiment the noise is the floor and the information follows the E power out to the beam; TB (dotted orange) adds little because the temperature is a good copy of E only where E is noisy. For the deep experiment the lensing B is the floor and the information follows the acoustic peaks of EE to $\ell \approx 2000$. The large-scale bump at $\ell < 10$ is the reionization peak of EE.

This answers the question of which spectra matter. EE is large and it is the lever: the twist borrows its pattern, so the brighter E is at some multipole, the more B the twist makes there. BB is small and it is the floor: it enters only the denominator, as part of the wobble against which the borrowed B is seen. A brighter EE helps; a brighter BB hurts. Both are needed in the forecast, but only through the covariance, never through a derivative.

Watch out: The reionization bump

The information per multipole at $\ell < 10$ is large, because C_ℓ^{EE} has its reionization peak there while the lensing B is smallest. The published analyses nevertheless start at $\ell = 51$ [62]: on the largest scales the polarization maps are dominated by foreground residuals and instrumental systematics, whose EB is not that of a twist. A forecast that keeps $\ell < 50$ promises information that the real analysis throws away. The forecast numbers for both experiments therefore use $51 \le \ell \le \ell_{\rm max}$.

10.15.3 The dials: depth, beam, multipole range and sky fraction

The sky fraction. Every term of (10.98) carries one factor $f_{\rm sky}$, so $\sigma(\beta) \propto f_{\rm sky}^{-1/2}$ exactly. For the Planck-like experiment the forecast is 0.0415° at $f_{\rm sky} = 0.7$ and 0.0366° at $f_{\rm sky} = 0.9$, in the ratio $\sqrt{0.7/0.9} = 0.88$. With a tapered mask the $f_{\rm sky}$ to use is not the plain fraction of sky kept but $f_{\rm sky} = w_2^2/w_4$, with w_i the average of the i-th power of the mask over the sphere [80], eq. (22). It is the number of independent modes a tapered window leaves, derived in Eq. (8.72): a mode in a region where the mask has weight W has its variance multiplied by W^2, and the modes count fully only where $W = 1$.

The depth. Fix the beam at 5′ and vary Δ_P, with the temperature depth $\Delta_P/\sqrt{2}$ as for detectors that measure both. Equation (10.98) predicts three regimes, which we read off the well-measured-E form $F_\ell \propto C_\ell^{EE}/(C_\ell^{BB} + N_\ell)$ and its noisy-E version $F_\ell \propto (C_\ell^{EE})^2/[(C_\ell^{EE} + N_\ell)(C_\ell^{BB} + N_\ell)]$.

1. *Noise above the E signal* ($N_\ell \gg C_\ell^{EE}$): $F_\ell \propto (C_\ell^{EE})^2/N_\ell^2$, and $N_\ell \propto \Delta_P^2$, so $\sigma(\beta) \propto \Delta_P^2$. Halving the noise improves the error four times.
2. *Noise between the two* ($C_\ell^{BB} \ll N_\ell \ll C_\ell^{EE}$): $F_\ell \propto C_\ell^{EE}/N_\ell$, so $\sigma(\beta) \propto \Delta_P$.
3. *Noise below the lensing B* ($N_\ell \ll C_\ell^{BB}$): $F_\ell \propto C_\ell^{EE}/C_\ell^{BB}$, independent of the noise. The error stops improving: lensing is the floor.

Figure 10.43(a) shows all three. A real experiment mixes regimes across multipoles, so the local slope $\mathrm{d}\ln\sigma/\mathrm{d}\ln\Delta_P$ lies between them: 1.36 at 60 μK-arcmin, where the noise is above the E signal at high ℓ and below it at low ℓ, and 0.54 at 3 μK-arcmin, close to the lensing floor. The floor itself, with $\ell \le 3000$, is $9.8 \times 10^{-4^\circ}$ at 0.1 μK-arcmin; with the lensing B removed it would be $5.2 \times 10^{-5^\circ}$, almost twenty times smaller, which is why a future experiment measuring very small angles would want to remove the lensing B modes first (delensing, Section T1.11.4).

The beam. At fixed depth a wider beam pushes N_ℓ up at lower ℓ and removes the multipoles where most modes live: at 60 μK-arcmin, $\sigma(\beta)$ grows from 0.045° with a 5′ beam to 0.16° with 30′ (Figure 10.43(b)).

The multipole range. Raising $\ell_{\rm max}$ helps only while the new multipoles have $N_\ell \lesssim C_\ell^{EE}$. At 60 μK-arcmin and 5′ the error is 0.065° for $\ell_{\rm max} = 500$, 0.047° for 1000, 0.045° for 1500, and 0.045° for 3000: beyond $\ell \approx 1000$ the noise has won. A deeper map keeps gaining to much higher ℓ (Figure 10.43(c)), because its noise stays below the E signal there.

10.15.4 A Planck-like forecast against the published errors

Their question. The Planck team [214] and later [62] measured the angle by which the CMB polarization appears turned, from the EB (and TB) spectra of the HFI maps over $51 \le \ell \le 1500$. With the detector angles held at their ground calibration, $\alpha_\nu = 0$, the measured angle is the sum $\alpha + \beta$ of Section T1.11.6, and the statistical error is what our forecast predicts. The Planck team

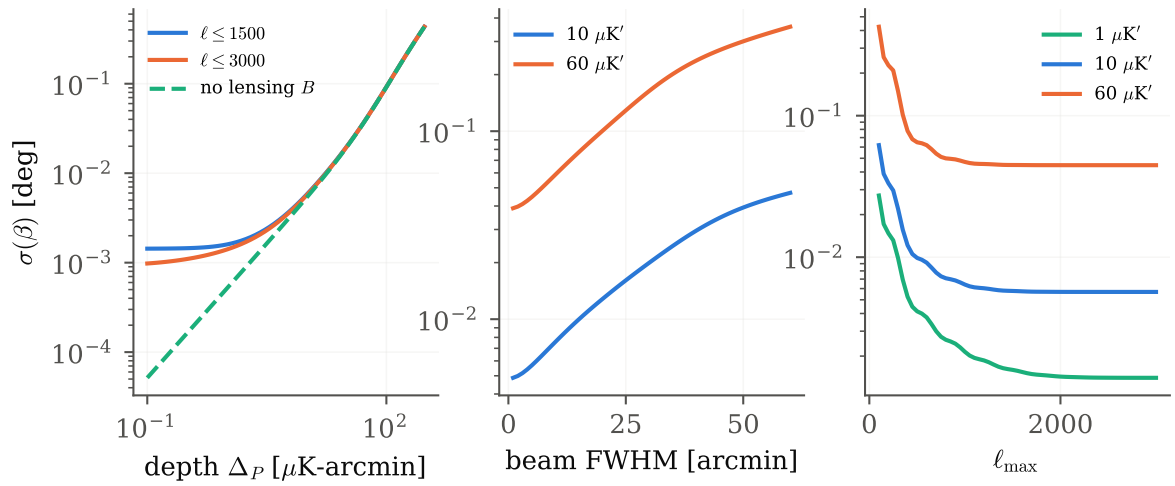


Figure 10.43: The dials of a birefringence forecast, from (10.98) with EB and TB and $f_{\rm sky} = 0.7$. (a) Error against polarization depth with a 5′ beam: steeper than linear for noisy maps, linear in between, and flat below a few μK-arcmin, where the lensing B modes become the floor; without them (dashed) the error would keep falling. (b) Error against beam width at two depths. (c) Error against the largest multipole used: it stops improving where the noise overtakes the E signal, earlier for a noisier map.

found 0.29° ± 0.05° (statistical), Minami and Komatsu 0.289° ± 0.048° with the same data, and Eskilt and Komatsu, from the reprocessed maps (PR4) with more detector splits, 0.288° ± 0.032° [80].

Our forecast. The Planck-like experiment of Section 10.15.2, $51 \le \ell \le 1500$, gives

$$\sigma(\beta) = 0.0428^\circ\ (EB), \qquad 0.0879^\circ\ (TB), \qquad 0.0415^\circ\ (EB + TB) \qquad (f_{\rm sky} = 0.7), \tag{10.101}$$

and 0.0366° for $f_{\rm sky} = 0.9$. TB alone is about twice worse than EB and adds only a few per cent when combined, as (10.100) predicts for an experiment whose E is measured above its noise over most of the range.

Side by side, and why they differ. The published statistical errors, 0.048° to 0.05° for the PR3 maps, are about a third larger than our 0.0366° at a similar sky fraction (their apodized masks keep $f_{\rm sky} = 0.77$–0.97 depending on the map, [62]). Three causes push the real error up. The published analyses use cross-spectra between half-mission maps, each of which has $\sqrt{2}$ times the noise of the full map; this removes the noise bias but loses information wherever the noise matters. Their covariance is built from the observed spectra, which contain the residual Galactic emission, so the floor $\tilde{B}_\ell$ is higher than our CMB-plus-noise model. And the real noise is not white: it rises towards large scales and varies over the sky. The PR4 error is smaller because the reprocessing lowered the noise by 10–30% [79] and because more independent detector splits were used. Lesson: a Fisher forecast with white noise and an ideal sky fraction lands within about a third of the measured error; the remaining gap is a list of known losses, each of which can be put back into the covariance.

10.16 Two angles and one sum

The forecast in Section 10.15 treated the orientation of each detector's polarisation axis as known exactly. In practice that orientation comes from a ground calibration, and if that axis is off by an angle α, every stick the detector sees appears twisted by α (Section T1.11.6). The CMB light then arrives twisted by β on its journey and by α in the instrument, and its spectra depend only

on the sum $\alpha+\beta$. For Planck the ground calibration of α is uncertain by 0.28° [62], almost seven times the statistical error of the forecast (10.101). Section T1.13 described the way out with a door frame next to a crooked painting: the polarized light of our own Galaxy passes through the same detectors but travels too short a distance to be twisted by β, so it carries α alone [76]. A single-channel forecast there showed that the dust in one map can close the degeneracy, at a price. A real experiment has several frequency channels, each with its own detector angle α_ν, and dust that is faint in some and bright in others. How well is β then measured, and what sets the answer?

Four scales, each with its own zero. A post office has four kitchen scales, and each one reads a little off: scale i shows its true load plus an unknown zero error α_i. A parcel of unknown weight β is put on every scale in turn. Each reading is $\alpha_i+\beta$, so the parcel alone can never tell the clerk whether the parcel is heavy or the scales read high: shifting every zero up by a gram and the parcel down by a gram changes no reading at all. The clerk needs a second measurement that involves the zeros but not the parcel. She looks at each scale with its pan empty. That reading is α_i itself, though on some scales the display flickers so much that the empty reading is poor. Now each scale gives its own estimate of the parcel, the loaded reading minus the empty one, and the error of that estimate is set mostly by the worse of the two readings. A scale whose empty reading is useless contributes nothing to the parcel's weight; it only measures its own zero.

The four scales are the four frequency channels of the Planck HFI, 100, 143, 217 and 353 GHz. The parcel is β, which only the CMB carries. The empty pan is the Galactic dust, which sees each channel's α_ν and not β. The flicker is the dust's faintness: at 100 GHz the dust is barely there, at 353 GHz it is bright.

The scales in symbols. Suppose each channel's sum $\alpha_i+\beta$ is measured with error σ_s and its zero α_i with an independent error σ_i. Then $\hat\beta_i=\widehat{(\alpha_i+\beta)}-\hat\alpha_i$ has variance $\sigma_s^2+\sigma_i^2$. The dials are σ_s, the σ_i and the number of channels. The naive choice fixes every α_i at its nominal zero and quotes $\hat\beta$ with error σ_s: it is precise, and wrong by whatever the true α_i are. The good choice fits the zeros too. With $\sigma_s=0.05^\circ$, roughly the Planck-like sum of (10.101), and zeros known to $\sigma_i=0.5^\circ$ from the dust, each channel gives β to $\sqrt{0.05^2+0.5^2}=0.50^\circ$: the sum hardly matters. Four such channels would average to 0.25° if their errors were independent. The CMB part is not independent between channels, because all of them see the same CMB sky, but that part is small here; the dust part, which differs from channel to channel through its noise, is what averages. Lesson: once the detector angles are free, the error on β is set by how well the dust measures the α_ν, not by how well the CMB measures the sum [76].

A picture. In the plane of (α,β) for one channel, the CMB alone allows a band along the line $\alpha+\beta=$ const, infinitely long. The dust adds a band along $\alpha=$ const, a vertical strip. The two cross in a parallelogram, long and thin along the CMB band when the dust strip is wide, and that parallelogram (Figure 10.44) is what the Fisher ellipse in (α,β) looks like. Its long axis is tilted at 45° downwards, so α and β are anti-correlated: a fit that pushes α up pushes β down by the same amount.

One idea, many names. Separating a common offset from individual offsets by adding a measurement that sees only the individual ones is the logic of every calibration network: the zero points of several telescopes from stars that all of them see, the gains of an interferometer's antennas from a calibrator source, the alignment of tracking detectors from cosmic rays that cross them all. In the Fisher matrix it appears as a null direction, $(\delta\alpha_1,\dots,\delta\alpha_N,\delta\beta)\propto(1,\dots,1,-1)$, that the second kind of data lifts. We now build the covariance of several channels, show that without dust it is exactly singular, and compute what the dust buys, channel by channel and for all four channels at once.

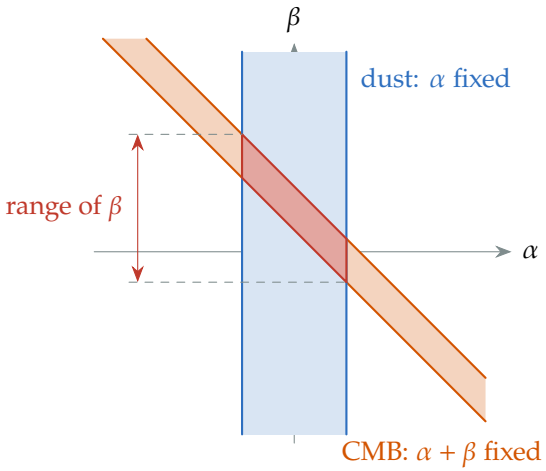


Figure 10.44: Why the dust sets the error on β. The CMB of one channel allows a band of constant $\alpha + \beta$ (orange); the dust allows a strip of α values (blue). They cross in a tilted parallelogram (red) whose height, the error on β, is set by the width of the dust strip once the CMB band is narrow, and whose tilt makes α and β anti-correlated.

10.16.1 The covariance of several channels

The maps. Channel ν sees the CMB, twisted by $\alpha_\nu + \beta$, the dust, twisted by α_ν, and its own noise. By Eq. (T1.74), mode by mode,

$$E_\nu + iB_\nu = e^{2i\alpha_\nu}\, a_\nu \big(E_\mathrm{d} + iB_\mathrm{d}\big) + e^{2i(\alpha_\nu+\beta)}\big(E_\mathrm{c} + iB_\mathrm{c}\big) + n_\nu^E + in_\nu^B, \tag{10.102}$$

where $(E_\mathrm{c}, B_\mathrm{c})$ are the CMB coefficients, $(E_\mathrm{d}, B_\mathrm{d})$ those of one fixed dust pattern, and a_ν the brightness of the dust in channel ν in CMB temperature units. Writing the dust as one pattern times a number per channel assumes that its emission law is the same in every direction of the sky. It is the simplest model, and it makes the dust in different channels perfectly correlated; a dust whose emission law changes across the sky would decorrelate the channels a little and weaken the calibration.

The covariance. The data at one (ℓ, m) are the $2N$ numbers $(E_1, B_1, \ldots, E_N, B_N)$. Write $\boldsymbol{R}(\theta)$ for the 2×2 matrix that turns (E, B) by 2θ, $\boldsymbol{S}_\mathrm{c} = \mathrm{diag}(C_\ell^{EE}, C_\ell^{BB})$ for the CMB and $\boldsymbol{S}_\mathrm{d} = \mathrm{diag}(C_{\ell,\mathrm{d}}^{EE}, C_{\ell,\mathrm{d}}^{BB})$ for the dust pattern. Since the three pieces of (10.102) are independent with mean zero, the 2×2 block between channels μ and ν is the sum of three covariances:

$$\mathbf{C}_{\mu\nu} = \boldsymbol{R}(\alpha_\mu + \beta)\, \boldsymbol{S}_\mathrm{c}\, \boldsymbol{R}(\alpha_\nu + \beta)^\mathsf{T} + a_\mu a_\nu\, \boldsymbol{R}(\alpha_\mu)\, \boldsymbol{S}_\mathrm{d}\, \boldsymbol{R}(\alpha_\nu)^\mathsf{T} + \delta_{\mu\nu} N_{\ell,\nu}^P \mathbb{1}. \tag{10.103}$$

Each term is the rule "if $\boldsymbol{x} = \boldsymbol{R}\boldsymbol{y}$ then $\mathrm{Cov}\,\boldsymbol{x} = \boldsymbol{R}\,\mathrm{Cov}(\boldsymbol{y})\boldsymbol{R}^\mathsf{T}$" applied to one component in two channels; the noise is independent between channels and isotropic in (E, B). The dust model is the one of Section T1.13.3: Planck's power-law fit over 71% of the sky at 353 GHz, $C_\mathrm{d}^{BB} = 0.53\, C_\mathrm{d}^{EE}$, scaled to other frequencies by a modified black body [71]. The Fisher matrix in the $N+1$ parameters $(\alpha_1, \ldots, \alpha_N, \beta)$ is (10.95) with the $2N \times 2N$ trace, and the script computes the derivatives of (10.103) by central differences (Section 10.5).

Without dust the matrix is singular, exactly. Set $a_\nu = 0$. Every entry of (10.103) depends on the angles only through the sums $\alpha_\nu + \beta$. Change every α_ν by $+\delta$ and β by $-\delta$: no sum changes, so the covariance is unchanged, so the likelihood is unchanged, so the curvature of $-\ln\mathcal{L}$ along the direction $\boldsymbol{u} = (1, \ldots, 1, -1)$ is zero, $\boldsymbol{u}^\mathsf{T}\mathbf{F}\boldsymbol{u} = 0$. A Fisher matrix with a zero-curvature direction has a zero eigenvalue and no inverse (Section 10.4.3). For the four Planck channels the script finds the ratio of the smallest to the largest eigenvalue to be 5×10^{-17}, zero to round-off. With the dust, the dust terms depend on α_ν alone, the shift $\boldsymbol{u}$ changes them, and the eigenvalue is lifted.

10.16.2 One channel at a time

Take one Planck channel with its noise and beam, the CMB and the dust, $f_{\rm sky} = 0.7$ and $51 \le \ell \le 1500$, and compute the 2×2 Fisher matrix in (α, β). Table 10.8 lists the marginal errors (the square roots of the diagonal of $\mathbf{F}^{-1}$, Section 10.4.1).

channel	$\sigma(\alpha)$ [deg]	$\sigma(\beta)$ [deg]	$\sigma(\alpha+\beta)$ [deg]	$\rho(\alpha,\beta)$
100 GHz	7	7	0.18	−1.000
143 GHz	1.1	1.2	0.07	−0.998
217 GHz	0.49	0.54	0.12	−0.976
353 GHz	0.3	2.4	2.2	−0.596

Table 10.8: Fisher forecast for the detector angle α and the birefringence angle β from the E and B modes of one Planck HFI channel at a time, with CMB, dust and noise, $f_{\rm sky} = 0.7$, $51 \le \ell \le 1500$. The sum is always measured better than either angle, and the two angles are almost perfectly anti-correlated except at 353 GHz.

Reading the table across frequency. At 100 and 143 GHz the dust is faint. The CMB fixes the sum well (0.07° at 143 GHz, close to the CMB-only value 0.066°), but the dust pins α so weakly that $\beta = (\alpha+\beta) - \alpha$ inherits the large error of α, and the two are anti-correlated with $\rho \approx -1$: the ellipse is the CMB band, cut off far away. At 217 GHz the dust is brighter and $\sigma(\beta)$ drops to 0.54°. At 353 GHz the dust is so bright that it measures α to 0.3°, but it also fills $\tilde{E}_\ell$ and $\tilde{B}_\ell$ with its own power, which is the floor against which the CMB's EB is read (Section 10.15.1). The sum is now known only to 2.2°, and β to 2.4°. Too little dust cannot calibrate, too much blurs the CMB.

Figure 10.45(a) turns this into a curve. For one channel with the same depth at every frequency, the error on β is large at both ends and has a minimum in between, near 280 GHz in our dust model, where the dust and CMB polarization are of comparable size on the scales that carry the information. A lower noise moves the whole curve down and widens the useful range, but hardly moves the minimum. Minami et al. found the same shape for the channels of the planned LiteBIRD satellite, with the best single-channel error, 11′ (0.18°), at 195 GHz [76], Table 2. Their minimum sits at a lower frequency for a reason the curve itself suggests: the minimum is where dust and CMB balance on the scales that carry the information. LiteBIRD's beams are tens of arcminutes wide, so its information comes from ℓ of a few hundred and below, where the E power of the CMB is much smaller and the dust, whose power falls with ℓ, relatively brighter; the balance is then struck at a lower frequency, where the dust is fainter.

10.16.3 All four channels, each with its own angle

Now fit all four channels at once, with five parameters $(\alpha_{100}, \alpha_{143}, \alpha_{217}, \alpha_{353}, \beta)$ and the full 8×8 covariance (10.103). The channels help each other in two ways. The bright dust at 353 GHz is the same dust pattern that is faint at 143 GHz, so it tells the fit what that faint dust looks like; and the CMB, the same in every channel, ties all the sums together. The forecast is

$$\sigma(\beta) = 0.25^\circ \quad (\alpha_\nu \text{ free}), \qquad \sigma(\beta) = 0.044^\circ \quad (\alpha_\nu \text{ known}), \tag{10.104}$$

with each $\sigma(\alpha_\nu)$ between 0.23° and 0.27° and each correlation $\rho(\alpha_\nu, \beta)$ between −0.98 and −0.9. Not knowing the detector angles costs a factor of about six. All four α_ν end up with nearly the same error as β: the scales of the story, whose loaded readings are nearly perfect, so that each zero is known as well as the parcel and no better.

The dials. Shrinking the sky to $f_{\rm sky} = 0.5$ gives 0.3°; enlarging it to 0.9, 0.22°, the $f_{\rm sky}^{-1/2}$ scaling. Lowering every channel's noise by a factor of three gives 0.15°, by ten 0.1°; raising it

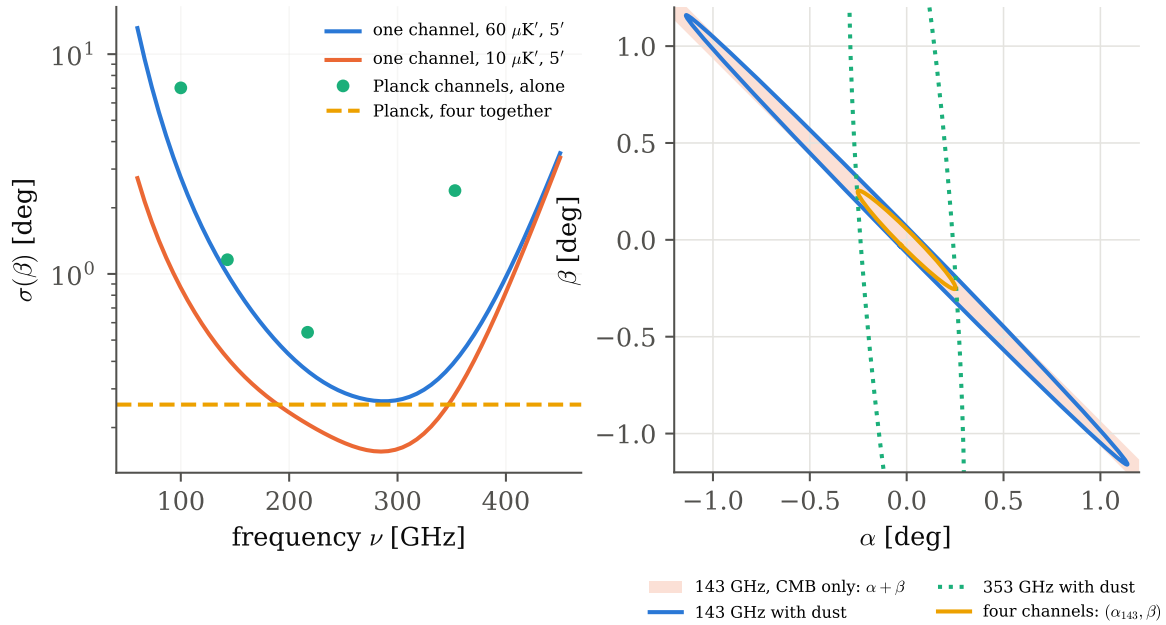


Figure 10.45: Breaking the degeneracy with dust. (a) Error on β from a single channel against its frequency, for a channel of fixed depth (two depths shown) and 5′ beam, with the detector angle free: poor where the dust is too faint to calibrate (low ν) and where it is so bright that it blurs the CMB (high ν). Dots: the four Planck HFI channels, each alone; dashed line: all four together, each with its own α_ν. (b) The 1σ regions in (α, β): the CMB alone at 143 GHz allows a band of constant $\alpha + \beta$ (shaded); the faint 143 GHz dust closes it far away (blue), the bright 353 GHz dust pins α but leaves β loose (dotted green); the four channels together (orange, shown in the plane of α_{143} and β) give the smallest region.

threefold gives 0.42°. Making the dust three times fainter in power gives 0.42°, three times brighter 0.17°. Both the noise and the dust brightness matter, because the dust calibrates α_ν only where it stands above the noise.

Side by side with the measurement. [62] fitted the same five parameters to the Planck 2018 maps and found $\beta = 0.35° \pm 0.14°$, an error about half of our 0.25°. Their setup differs from ours in the direction that helps them. Their masks keep 77–97% of the sky, including regions close to the Galactic plane, where the dust is several times brighter than in the 71% region our dust model describes; our forecast with three times brighter dust gives 0.17°, and with $f_{\rm sky} = 0.9$ on top of that the error would fall further. They did not model the dust spectra at all: the relation Eq. (T1.77) uses the observed $EE - BB$ of each map as the dust template, so the brightest regions contribute fully. And they used every cross-spectrum between half-mission maps of different channels. Lesson: with the detector angles free, the error on β is a statement about the polarized Galaxy, its brightness and the area over which it is used, more than about the CMB.

Watch out: A brighter Galaxy is also a less certain reference

Using more of the bright Galactic plane shrinks $\sigma(\beta)$ in the forecast, but the method assumes that the dust itself has no EB. Any intrinsic dust EB appears as an extra twist γ of the dust and shifts the result to $\beta - \gamma$ (Eq. (T1.78)). The PR4 analysis found β falling as the Galactic mask grows [79], which is what a dust EB concentrated near the plane would do. The Fisher error does not see this bias; only a model of the dust EB, or a calibration that does not rely on the dust, can.

10.17 Simulating the measurement

A forecast is a promise made with a clean model of the data. Twice already the book has checked such a promise by building many skies on the computer and measuring them. In Section 10.6.2 Knox's error bars for a beamed, noisy map were set against a hundred simulated skies per experiment, and they matched. In Section 10.6.6 a Markov chain run on the full likelihood found the Fisher ellipses of ΛCDM where the forecast had put them, except along degeneracies. And in Chapter 8 the simulations showed what a mask does to a temperature spectrum: it couples neighbouring multipoles and changes the scatter, and the MASTER method undoes the first effect on average. The forecast of Section 10.15 made its promise for clean, full-sky T, E, B coefficients with a sky fraction put in by hand. A real polarization map has a hole where the Galaxy is, and that hole does something to polarization that it does not do to temperature.

Think of a field of wheat stalks combed in neat rows by the wind, all leaning east along the rows. Photograph it through a window frame. Inside the frame the rows are perfect. But cut the photograph along the edge of the frame, and ask someone who has only the cut photograph to describe the pattern by its waves. A wave that ends abruptly at the frame is no longer a single wave; it is a bundle of waves pointing in slightly different directions, and for those other directions the same stalks no longer lean along the wave but at an angle to it. Part of a pure "along the rows" pattern is now described as "across", and nothing in the field has changed. The leakage is an artefact of cutting.

On the sky the combed rows are E, the "across" description is B, and the window frame is the Galactic mask. A mask turns some E into B by itself. For a measurement of β this is the dangerous direction: the twist is read as B that remembers E, and the floor against which it is read is the B already present. Leaked B raises that floor. Does it also fake a twist? Does the estimator stay unbiased? And if the covariance used to weight the data and quote the error ignores the leaked B, are the quoted error bars still honest? Those are questions about the whole distribution of the estimate over many skies, and the simulations answer them.

The leak in symbols. On a flat patch, take one pure-E plane wave along the first axis, so that its sticks lie along or across the wavevector: $Q(\boldsymbol{\theta}) = A\cos(k\theta_1)$, $U = 0$ (Eq. (T1.36) with $\varphi_{\boldsymbol{\ell}} = 0$). Multiply it by a window $W(\theta_2)$ that depends only on the other coordinate, a straight edge parallel to the wave's direction of travel. The masked Q is a product, so its Fourier transform is a product of a delta function in ℓ_1 and the transform $\widetilde{W}(\ell_2)$ in ℓ_2; U stays zero. At a wavevector $\boldsymbol{\ell} = (k, \ell_2)$ whose angle is φ, with $\tan\varphi = \ell_2/k$,

$$B(\boldsymbol{\ell}) \overset{(1)}{=} -\sin 2\varphi\; Q(\boldsymbol{\ell}) + \cos 2\varphi\; U(\boldsymbol{\ell}) \overset{(2)}{=} -\sin 2\varphi\; \tfrac{A}{2}\,\widetilde{W}(\ell_2) \overset{(3)}{=} -\frac{2k\ell_2}{k^2+\ell_2^2}\,\tfrac{A}{2}\,\widetilde{W}(\ell_2),$$

$$E(\boldsymbol{\ell}) \overset{(4)}{=} \frac{k^2-\ell_2^2}{k^2+\ell_2^2}\,\tfrac{A}{2}\,\widetilde{W}(\ell_2). \tag{10.105}$$

(1) The definition of B for one wave, Eq. (T1.36).
(2) $U = 0$; the masked Q at (k, ℓ_2) is $\frac{A}{2}\widetilde{W}(\ell_2)$ (the cos splits into two delta functions at $\ell_1 = \pm k$, and we keep the one at $+k$).
(3) $\sin 2\varphi = 2\tan\varphi/(1+\tan^2\varphi)$ with $\tan\varphi = \ell_2/k$.
(4) The same for E, with $\cos 2\varphi = (1-\tan^2\varphi)/(1+\tan^2\varphi)$.

The mask spreads the single wave over a range of ℓ_2 as wide as $\widetilde{W}$, and at every $\ell_2 \neq 0$ a part $\propto \ell_2/k$ of it is read as B. How much in total? Parseval's theorem, $\int |\widetilde{W}(\ell_2)|^2\,\mathrm{d}\ell_2/2\pi = \int W(\theta_2)^2\,\mathrm{d}\theta_2$, turns the question into one about the window on the sky. Let the window keep a strip of width L, many wavelengths wide. The masked wave then carries the power $(A/2)^2\int W^2\mathrm{d}\theta_2 \approx (A/2)^2 L$ in the component at $+k$, and its B part carries $P_B = (A/2)^2\int \sin^2 2\varphi\,|\widetilde{W}|^2\,\mathrm{d}\ell_2/2\pi$. The dials are

the wavenumber k and the shape of the edge. Take the two edges in turn.

$$\begin{aligned}
\text{sharp edges:}\quad & \frac{P_B}{(A/2)^2} \overset{(1)}{=} \int \frac{4k^2\ell_2^2}{(k^2+\ell_2^2)^2}\,\frac{2}{\ell_2^2}\,\frac{\mathrm{d}\ell_2}{2\pi} \overset{(2)}{=} \frac{8k^2}{2\pi}\,\frac{\pi}{2k^3} = \frac{2}{k},\\
\text{tapered edges:}\quad & \frac{P_B}{(A/2)^2} \overset{(3)}{=} \frac{4}{k^2}\int \ell_2^2\,|\widetilde{W}|^2\,\frac{\mathrm{d}\ell_2}{2\pi} \overset{(4)}{=} \frac{4}{k^2}\int W'(\theta_2)^2\,\mathrm{d}\theta_2 \overset{(5)}{=} \frac{\pi^2}{k^2\theta_{\mathrm{apo}}}.
\end{aligned} \tag{10.106}$$

(1) $\sin^2 2\varphi$ from (10.105). The strip $|\theta_2| < L/2$ has $\widetilde{W} = 2\sin(\ell_2 L/2)/\ell_2$, whose square oscillates quickly and averages to $2/\ell_2^2$, since $\sin^2$ averages to $\frac{1}{2}$.

(2) The integral $\int \mathrm{d}x/(k^2+x^2)^2 = \pi/(2k^3)$, found by substituting $x = k\tan u$, which turns it into $k^{-3}\int \cos^2 u\,\mathrm{d}u$ over $(-\pi/2, \pi/2)$.

(3) A taper of width $\theta_{\mathrm{apo}} \gg 1/k$ confines $\widetilde{W}$ to $|\ell_2| \lesssim 1/\theta_{\mathrm{apo}} \ll k$, where $\sin 2\varphi \approx 2\ell_2/k$.

(4) Parseval's theorem again, for the derivative: the transform of W' is $i\ell_2\widetilde{W}$.

(5) A cosine taper rises as $W = \frac{1}{2}[1-\cos(\pi x/\theta_{\mathrm{apo}})]$ over $0 < x < \theta_{\mathrm{apo}}$, so $W' = \frac{\pi}{2\theta_{\mathrm{apo}}}\sin(\pi x/\theta_{\mathrm{apo}})$ and $\int W'^2\mathrm{d}x = \pi^2/(8\theta_{\mathrm{apo}})$; the strip has two edges.

Divide by the power $(A/2)^2 L$ of the wave. A sharp edge leaks the fraction $2/(kL)$: it is as if a strip about one wavelength wide along each cut were read as B, so the leaked fraction falls only like $1/\ell$. A taper leaks $\pi^2/(k^2\theta_{\mathrm{apo}}L)$, falling like ℓ^{-2}, and the ratio of the two,

$$\frac{\text{taper}}{\text{sharp}} = \frac{\pi^2}{2k\theta_{\mathrm{apo}}}, \tag{10.107}$$

does not depend on the size of the kept region. For a 5° taper, $\theta_{\mathrm{apo}} = 0.087$ rad, it is 0.2 at $\ell \approx 290$: the taper cuts the leak fivefold there. At $\ell \approx 60$ the formula gives 0.9, but there $k\theta_{\mathrm{apo}} \approx 5$ and the assumption behind step 3 is only roughly met; the taper helps little on the largest scales either way. The good choice, tapering, costs sky: the tapered part of the mask counts only partly. Lesson: a sharp edge leaks a fraction of E into B that falls like $1/\ell$, a smooth one a fraction that falls like ℓ^{-2}.

Does the leaked B correlate with E and so look like a twist? In (10.105) the product $E(\boldsymbol{\ell})B^*(\boldsymbol{\ell}) \propto -(k^2-\ell_2^2)\,2k\ell_2\,|\widetilde{W}(\ell_2)|^2$ is odd in ℓ_2, while the window of a straight edge has $|\widetilde{W}(\ell_2)| = |\widetilde{W}(-\ell_2)|$. The leaks at $+\ell_2$ and $-\ell_2$ have opposite signs and cancel in a spectrum, which sums over directions: the mask makes B power without a hand. The same holds for any mask on the sphere, because the pseudo-EB spectrum of a masked sky is fed by the true EB only and not by EE or BB [169]. A leak adds wobble to the floor; it does not fake the twist.

One idea, many names. Spreading a pure tone into neighbouring frequencies by cutting it off is spectral leakage, met for the temperature in Chapter 8 and for gravitational-wave data in Section 10.8.5; here the neighbours differ in direction, and direction is what separates E from B. The research tools built on this are the pseudo-C_ℓ estimators of polarization: MASTER generalised to spin-2 fields [168, 169], which undoes the leak in the mean spectra, and "pure" B-mode estimators, which build the B map so that it never receives E in the first place [215]. We use neither: the twist passes through the mask untouched, as Section 10.17.1 shows, so the angle can be measured from the raw masked spectra; the leak then matters only for the error bar.

10.17.1 The simulated skies

The script `code/ch10/18_biref_mc.py` draws 1000 independent skies, each a Gaussian realisation of the lensed fiducial T, E, B spectra with $EB = TB = 0$, on a HEALPix grid of $N_{\mathrm{side}} = 256$

(pixels of 13.7′). The lensing B modes are drawn as a Gaussian field with the lensed BB spectrum; real lensing B is not Gaussian (Section T1.11.4), which matters for the scatter of BB itself but very little for EB. Each sky is smoothed with a 7.22′ Gaussian beam, that of the 143 GHz channel [186], and the pixel window. Two noise levels are added, white in each pixel:

- *Planck-like*: 52.3 μK-arcmin in Q and U and 25.5 μK-arcmin in T, the depths of the combined HFI channels of Section 10.15.2;
- *deep*: 3 μK-arcmin in Q, U and $3/\sqrt{2}$ in T, below the lensing floor.

Two Galactic-like masks remove $|b| < 17.5°$: one with a sharp edge, one with a 5° cosine taper eating into the kept region, with effective sky fractions $w_2^2/w_4 = 0.699$ and 0.656 (Section 10.15.3). The analysis uses $51 \le \ell \le 600$ in 22 bins of 25 multipoles. The full research version (Section 10.18) works at $N_{\rm side} = 2048$ to $\ell = 1500$ with the real noise of each map; the smaller grid keeps a thousand skies under half an hour on four cores. It costs the information above $\ell = 600$: by Section 10.15.3, a 60 μK-arcmin map measures β to 0.065° with $\ell \le 500$ and to 0.045° with $\ell \le 1500$. Every Fisher number set against the simulations is computed for exactly the simulated set-up, so the comparison is fair.

Two choices make the simulation cheap and the leakage visible. The E part of each sky (with its correlated T) and the B part are turned into maps separately, masked, and transformed back. Because the transform is linear, their sum is the masked sky, and the B coefficients of the masked E part are exactly the leaked B. The noise is drawn once per sky at unit depth and scaled to each noise level, again by linearity.

The twist passes through the mask. The twist is applied after masking, to the coefficients of the masked sky. That is legitimate because the two operations commute. Masking multiplies $Q + iU$ at each pixel by the real number W; twisting multiplies it by the constant $e^{2i\beta}$; at each pixel $W\, e^{2i\beta}(Q+iU) = e^{2i\beta}\, W(Q+iU)$, so the masked twisted sky is the twisted masked sky. Its spin-2 coefficients are then the twisted ones of Eq. (T1.49), and the pseudo-E and pseudo-B of the masked sky turn into each other by exactly the rule Eq. (T1.50). On one sky, twisting the Q, U maps before the mask and twisting the masked coefficients after it agree to 2×10^{-16} of the largest coefficient, round-off. The consequence is the relation of Section T1.11.5, now on the masked sky: for each sky, the measured pseudo-spectra obey $\widetilde{C}_\ell^{EB\prime} = \frac{1}{2}\sin 4\beta\,(\widetilde{C}_\ell^{EE} - \widetilde{C}_\ell^{BB}) + \cos 4\beta\,\widetilde{C}_\ell^{EB}$, whatever the mask has done to them. No mode-coupling matrix is needed to measure the angle.

10.17.2 The estimator

The model. Divide every measured pseudo-spectrum by w_2, the average of W^2, so that it estimates the full-sky spectrum on average, and bin it. For a small twist the relation above becomes, bin by bin,

$$e_b = 2\beta\, D_b + n_b, \qquad e_b \equiv \widetilde{C}_b^{EB}, \quad D_b \equiv \widetilde{C}_b^{EE} - \widetilde{C}_b^{BB}, \tag{10.108}$$

where n_b is the chance EB the sky and the noise would have shown without a twist, and $\sin 4\beta = 4\beta$ to a relative accuracy of $(4\beta)^2/6 \approx 10^{-4}$. The template D_b is measured on the same sky; its noise bias cancels in the difference, because white noise puts the same power into EE and BB. For the joint fit we add $t_b = \widetilde{C}_b^{TB} = 2\beta\, G_b + \dots$ with $G_b = \widetilde{C}_b^{TE}$.

The estimate. With the data vector $\boldsymbol{e}$, the template $\boldsymbol{D}$ and a covariance $\boldsymbol{V}$ for $\boldsymbol{n}$, the least-squares fit of the slope 2β minimises $(\boldsymbol{e} - 2\beta\boldsymbol{D})^{\mathsf{T}}\boldsymbol{V}^{-1}(\boldsymbol{e} - 2\beta\boldsymbol{D})$ (Section 3.10.5). Setting the derivative with respect to β to zero,

$$\hat\beta \overset{(1)}{=} \frac{\boldsymbol{D}^{\mathsf{T}}\boldsymbol{V}^{-1}\boldsymbol{e}}{2\,\boldsymbol{D}^{\mathsf{T}}\boldsymbol{V}^{-1}\boldsymbol{D}}, \qquad \hat\sigma(\hat\beta) \overset{(2)}{=} \frac{1}{2\sqrt{\boldsymbol{D}^{\mathsf{T}}\boldsymbol{V}^{-1}\boldsymbol{D}}}. \tag{10.109}$$

(1) ∂_β of the quadratic form is $-4\boldsymbol{D}^{\mathsf{T}}\boldsymbol{V}^{-1}(\boldsymbol{e} - 2\beta\boldsymbol{D})$; set it to zero and solve.
(2) $\hat{\beta}$ is linear in $\boldsymbol{e}$ with coefficient vector $\boldsymbol{c} = \boldsymbol{V}^{-1}\boldsymbol{D}/(2\boldsymbol{D}^{\mathsf{T}}\boldsymbol{V}^{-1}\boldsymbol{D})$, so its variance is $\boldsymbol{c}^{\mathsf{T}}\boldsymbol{V}\boldsymbol{c} = 1/(4\boldsymbol{D}^{\mathsf{T}}\boldsymbol{V}^{-1}\boldsymbol{D})$ when $\boldsymbol{V}$ is the true covariance of $\boldsymbol{e}$.

The quoted error is honest only if $\boldsymbol{V}$ is the true covariance. We try three.

(a) *The model covariance, which ignores the leak.* $V_b = \tilde{E}_b\tilde{B}_b/\nu_b$ from (10.99), with the full-sky signal-plus-noise spectra and $\nu_b = f_{\mathrm{sky}}\sum_{\ell\in b}(2\ell+1)$ modes. This is what the forecast assumed.

(b) *The measured spectra.* The same formula with each sky's own measured $\widetilde{C}_b^{EE}$ and $\widetilde{C}_b^{BB}$ in place of the model, as the published analyses do [62, 80]. The measured BB contains the leaked B, so this covariance knows about the leak without being told.

(c) *The simulations.* The covariance of the bandpowers over the other half of the skies (even skies for odd, odd for even), with the Hartlap correction of Section 8.13.2 for the noise in its inverse. It contains everything the simulations contain, bin-to-bin correlations included.

For the joint fit each V_b becomes the 2×2 covariance of (e_b, t_b) from the general Gaussian formula of Section 10.15.1, $\mathrm{Cov}(EB, TB) = C_\ell^{TE}\tilde{B}_\ell/\nu$. The angle put into every sky is $\beta_{\mathrm{in}} = 0.35°$.

10.17.3 What a thousand skies say

The leak. Figure 10.46(a) shows the B power of the masked E part of the skies, averaged over the 1000 skies: pure leakage, since those skies had no B at all. In the lowest bin, $51 \le \ell < 76$, it is 0.8% of the E power for the tapered mask and 1.3% for the sharp one, which is 1.5 and 2.5 times the lensing B there. In the bin centred on $\ell = 288$ the sharp edge still leaks 0.24% of the E power, 0.59 of the lensing B, while the taper has cut it to 0.048% of E, 0.12 of the lensing B. Both behave as the cut wave predicts. At the sharp edge the leaked fraction falls by about the factor by which ℓ grows, the $1/\ell$ of (10.106). The ratio of taper to sharp edge is 0.20 at $\ell = 288$, where (10.107) gives 0.2, and 0.60 at $\ell = 63$, where the formula's assumption $k\theta_{\mathrm{apo}} \gg 1$ fails. (The slight rise of the tapered curve above $\ell \approx 500$ comes from the pixel grid, which also mixes E and B near its resolution limit.) The Planck-like noise lies well over an order of magnitude above the leak at every multipole, so for it none of this matters. For the deep noise the leak is the largest part of the floor on large scales. Panel (b) shows the other half of the promise: the average measured EB of the twisted skies follows $\frac{1}{2}\sin 4\beta(C^{EE} - C^{BB})$ at every multipole, with no extra EB from the mask.

The bias. With the simulation covariance and both EB and TB, the average of $\hat{\beta} - \beta_{\mathrm{in}}$ over the skies is $0.0034° \pm 0.002°$ for the Planck-like noise and $0.00029° \pm 0.0002°$ for the deep noise, tapered mask, and similar for the sharp one (the error is the Monte Carlo error of an average of 1000 values, the scatter divided by $\sqrt{1000}$). Is an offset this size, or a larger one, something we would reasonably expect from an unbiased estimator? An unbiased average lands more than one Monte Carlo error from the truth a third of the time and more than two errors away one time in twenty. Over all eight set-ups (two noise levels, two masks, EB alone or with TB) the largest offset is 2.0 Monte Carlo errors, which is 0.06 of the scatter of a single measurement. The eight are not eight independent checks, though. They are measured on the same thousand skies, with the noise of each sky scaled from one draw, so a chance lean of the particular skies drawn shows up in all of them together, and the eight offsets move as one rather than averaging out. What the skies establish is an upper limit: a bias as large as three Monte Carlo errors, 0.09 of the scatter of a single measurement, would have been seen, and none was. That is what the commutation argument predicts, since it says the estimator needs no mask correction at all; a smaller bias than a tenth of the scatter could only be tested with many more skies.

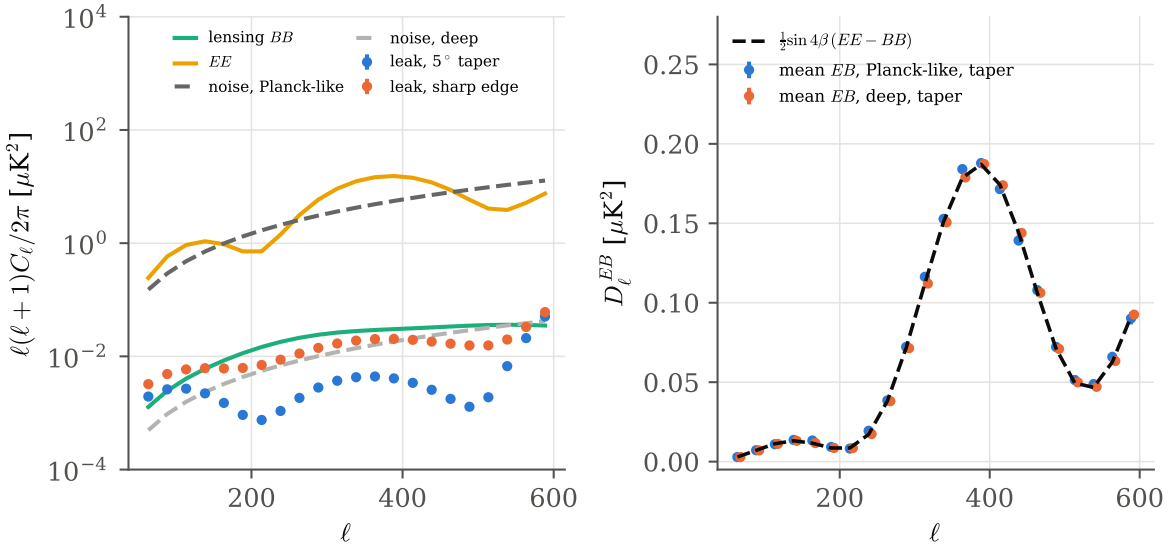


Figure 10.46: (a) E-to-B leakage of the two masks, measured as the B power of skies that contain only E (dots, with the error of the average over 1000 skies), against the lensing B, the E power and the two noise levels, all with the beam. Relative to E, the sharp edge leaks a fraction that falls roughly like $1/\ell$, the tapered mask one that falls faster, so the taper gains most at high ℓ. (b) The average EB of the skies twisted by 0.35°, tapered mask, Planck-like and deep noise (shifted slightly in ℓ for visibility), against the prediction of Eq. (T1.53): the mask adds no EB.

The scatter against the forecast. Table 10.9 sets the standard deviation of $\hat{\beta}$ over the skies against two Fisher errors for exactly the simulated set-up: the forecast of (10.98) with the sky fraction w_2^2/w_4, and the same with the measured leak added to the floor $\tilde{B}_\ell$. For the Planck-like noise the estimator reaches the forecast: the scatter is 1.034 times the Fisher error with the taper and 1.038 with the sharp edge, where the scatter itself is known to 0.001° from the finite number of skies, and adding the leak to the floor changes the forecast by less than one per cent. For the deep noise the scatter is 1.079 and 1.102 times the plain forecast. With the taper, the forecast that includes the leak, 0.00514°, accounts for most of the difference. With the sharp edge, the scatter lies between the two forecasts: counting the leak as extra noise independent of everything else overstates its harm, and the weights built from the simulations, which know how the leaked B is correlated across bins, recover part of the loss.

noise	mask	Fisher	Fisher, leak in floor	scatter of $\hat{\beta}$	ratio	bias
Planck-like	taper	0.0638	0.0639	0.066 ± 0.001	1.034	0.0034 ± 0.002
Planck-like	sharp	0.0618	0.0621	0.0641 ± 0.001	1.038	0.0025 ± 0.002
deep	taper	0.00486	0.00514	0.00524 ± 0.0001	1.079	0.00029 ± 0.0002
deep	sharp	0.0047	0.00563	0.00518 ± 0.0001	1.102	0.00033 ± 0.0002

Table 10.9: The estimator of β on 1000 simulated skies twisted by $\beta_{\rm in} = 0.35°$, EB and TB together, simulation covariance, $51 \le \ell \le 600$; all angles in degrees. "Ratio" is the scatter over the Fisher error. The Planck-like forecast is reached within a few per cent; for the deep noise the leak raises the floor and the scatter.

The error bars. A $\pm1\sigma$ interval should contain the true angle in 68.3% of the skies; with 1000 skies the fraction we measure carries a Monte Carlo error of ±1.5 percentage points (the binomial standard deviation $\sqrt{0.683 \times 0.317/1000}$). Figure 10.47(b) shows the coverage for the three covariances.

- *Planck-like noise:* model, measured spectra and simulations give 68%, 67.9% and 67.5% with

the taper, 69.3%, 69.3% and 67.9% with the sharp edge, all consistent with 68.3% given the Monte Carlo error. When the noise dominates the floor, ignoring the leak costs nothing.

- *Deep noise, model covariance (leak ignored):* 64.8% with the taper and 64.7% with the sharp edge, about 2.4 and 2.4 Monte Carlo errors short of 68.3%. The coverage does not stand alone: the quoted error, 0.00473° for the sharp mask, is smaller than the measured scatter of that estimator, 0.00504°, by about 3 times the Monte Carlo error of the scatter, and the Fisher error with the leak in the floor says why it must be. The error bars are too short, and an analysis that used them would claim more significance than its data hold.
- *Deep noise, measured spectra:* 69% with the taper, correct, but 74.5% with the sharp edge, too high. The measured BB contains the leak, and the formula treats it as independent noise, so the quoted error, 0.00569°, is larger than the scatter, 0.00502°: safe, but wasteful.
- *Deep noise, simulations:* 67.9% and 67%, correct within the Monte Carlo error, as they must be: this covariance contains everything the skies contain.

Figure 10.47(a) shows the shape behind these numbers. The distribution of $(\hat{\beta} - \beta_{\rm in})/\sigma_{\rm Fisher}$ over the skies is a Gaussian of nearly unit width, for both noise levels; a linear estimator of many nearly Gaussian bandpowers is itself nearly Gaussian.

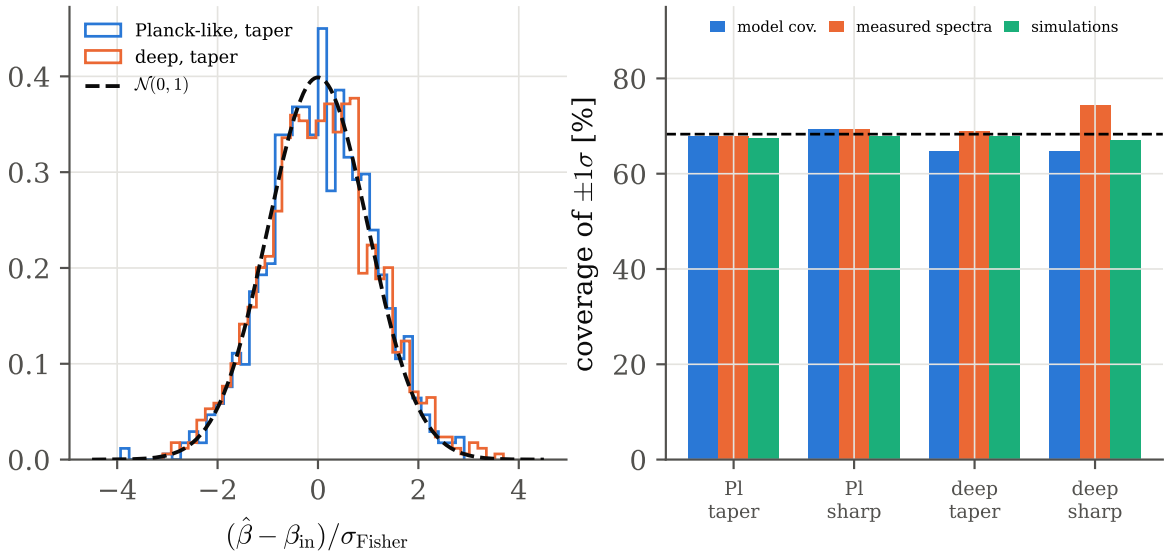


Figure 10.47: (a) The distribution of the estimated angle over 1000 skies, in units of the Fisher error of each set-up (tapered mask, EB and TB, simulation covariance): close to a unit Gaussian, as the forecast promised, slightly wider for the deep noise, where the leak raises the floor (Table 10.9). (b) The fraction of skies whose $\pm1\sigma$ interval contains $\beta_{\rm in}$, for the three covariances (model, which ignores the leak; measured spectra; simulations) and four set-ups; the dashed line is 68.3%, and the Monte Carlo error of each bar is ±1.5 points. Only the deep noise separates the covariances: ignoring the leak gives too short error bars, the measured spectra too long ones at a sharp edge.

The answer. Is the forecast reached on a sky with a hole, noise and a beam? For a Planck-like map it is, within a few per cent, by an estimator on raw masked spectra with no bias the skies can detect, with honest error bars from any of the three covariances. For a map deep enough to see below the lensing floor the mask's leak becomes part of the floor; the estimator still shows no bias, its scatter grows to 1.079–1.102 times the plain forecast, and the error bars are honest only if the covariance knows about the leak. A tapered mask keeps the loss small, and simulations make the error bars right. Every one of these statements is an average over 1000

skies with its Monte Carlo error attached, and a different set of random skies would move each number only within that error.

Watch out: A covariance from the model is only as good as the model

The forecast covariance knows the CMB, the noise and the sky fraction, and nothing else. Whatever the analysis does to the sky that the forecast did not, a mask, a filter, a beam asymmetry, enters the true scatter but not the quoted error. The test is the one done here: simulate the whole analysis on skies with a known answer and count how often the error bars cover it.

10.18 How a birefringence measurement is made in research

The published measurements of Sections 10.15.4 and 10.16.3 follow the same three steps, forecast, simulate, measure, at full scale and with every known imperfection of the instrument put in.

The data. Not one map per channel but two or more independent splits (half-missions, detector sets), so that only cross-spectra between splits are used and the noise bias never appears [62]. For four Planck channels in two splits this gives sixteen maps and many cross-spectra, each with its own detector angle in the model. The masks remove the Galactic plane, carbon-monoxide regions and bright point sources and are tapered by half a degree to a degree [62, 79], close to the masks of the Planck likelihood [75].

The spectra. Pseudo-spectra corrected for the mask with the spin-2 version of MASTER, in the codes NaMaster or PolSpice [169], and corrected for the leakage of temperature into polarization caused by the beams. Section 10.17.1 showed that the angle itself does not need the mask correction; the real analyses apply it anyway because the dust template and the CMB template of Eq. (T1.77) must be compared at the same multipoles.

The likelihood. The relation Eq. (T1.77), written for every pair of maps, is turned into a Gaussian likelihood for (α_ν, β) whose covariance is built from the observed spectra with the formula of (10.99), divided by an effective sky fraction [80], eqs. (21)–(22); it is sampled with an MCMC (Chapter 9). This is choice (b) of Section 10.17.2, which the simulations of Section 10.17.3 found honest with a tapered mask, and safe but wasteful at a sharp edge on a map deep enough for the leak to matter.

The simulations. End-to-end simulations of the instrument, with no twist: ten FFP10 skies for the PR3 analysis gave $\beta = 0.010^\circ \pm 0.030^\circ$ [62], six hundred NPIPE skies for PR4 [79]. Here the simulations are the research instrument, not a check: they contain the scanning, the beams, the noise correlations and the detector systematics, for which no closed form exists, and the published pipeline is trusted only because it returns zero on skies that have no twist. Ten skies whose estimates scatter by 0.030° pin their mean to $0.030^\circ/\sqrt{10} \approx 0.01^\circ$, a third of the scatter of one sky; excluding a bias of a tenth of the scatter, as Section 10.17.3 did, takes hundreds.

How this is used in real research: Simulations that check, simulations that measure

In Section 10.17 the simulations *verify*: the Fisher error and the coverage have analytic predictions, and a thousand Gaussian skies test them. In a real analysis the same machinery *is* the measurement of everything that has no analytic answer: the covariance of spectra from an inhomogeneous noise and a complicated mask, the bias from beam systematics, the distribution of $\hat{\beta}$ under the null hypothesis. Chapter 13 measures the sum $\alpha + \beta$ on the Planck SMICA maps with the estimator of Section 10.17.2, calibrated by simulations of exactly this kind.

	research analysis	here
maps	$N_{\text{side}} = 2048$, four channels, split maps, real noise	$N_{\text{side}} = 256$, one effective channel, white noise
multipoles	$51 \le \ell \le 1500$	$51 \le \ell \le 600$
foregrounds	real Galaxy, dust as the reference for α_ν	none in the simulations; Gaussian dust in the forecast
spectra	spin-2 MASTER, beam leakage corrected	raw masked pseudo-spectra (the angle commutes with the mask)
covariance	from the observed spectra	three options compared over 1000 skies
validation	FFP10 (10 skies), NPIPE (600 skies)	1000 Gaussian skies with a known twist

Table 10.10: The research version of a birefringence measurement and the laptop-scale version used for the forecast and the simulations.

One chain of reasoning leads from the twist to its measurement. A twist moves a little of the bright E pattern into B, so its information is the E power measured against the B already there; the instrument's own angle hides it until a component that sees only the instrument, the Galaxy, is added; and on a real sky the mask adds to the B already there without faking a twist, so the forecast survives, provided the error bars are built from a covariance that has seen the mask.

Recap

- A twist β enters the T, E, B covariance only through EB and TB at first order. The Fisher information per mode is $\boldsymbol{v}^{\mathsf{T}}\boldsymbol{A}^{-1}\boldsymbol{v}/\tilde{B}$, with $\boldsymbol{v} = (2C^{TE}, 2(C^{EE} - C^{BB}))$; for a well-measured E it is $4C^{EE}/(C^{BB} + N^{P})$: E is the lever, the B already present the floor. B power carries no information at small angles; the information on a rotation is the same at every angle.
- TB adds to EB only where E is noise-dominated by more than $1/\rho^2$. A Planck-like experiment measures the sum $\alpha + \beta$ to 0.0415° ($f_{\text{sky}} = 0.7$, $51 \le \ell \le 1500$), within about a third of the published statistical errors; the error scales as $f_{\text{sky}}^{-1/2}$, as Δ_P^2 to Δ_P with depth, and flattens at the lensing floor.
- With one detector angle per channel the CMB-only Fisher matrix is exactly singular. The dust lifts the null direction; four Planck-like channels give $\sigma(\beta) = 0.25^\circ$, about six times the error with known angles, and the error is set by the brightness of the dust and the noise, not by the CMB.
- A mask turns E into B (a fraction falling like $1/\ell$ at a sharp edge, like ℓ^{-2} with a taper) but makes no EB on average. The twist commutes with the mask, so the template estimator on raw pseudo-spectra needs no mask correction.
- Over 1000 simulated skies the estimator shows no bias; a bias of a tenth of its scatter would have been seen. With Planck-like noise it reaches the Fisher error within a few per cent and every covariance gives honest error bars. Below the lensing floor the leak raises the scatter to 1.079–1.102 times the plain forecast; a model covariance that ignores it covers only 64.8% (taper) and 64.7% (sharp edge), the measured spectra over-cover at a sharp edge, and the simulation covariance covers correctly.

Reading the literature

- **The Fisher matrix and the Cramér–Rao bound, textbook level.** Wasserman, *All of Statistics*, Chapter 9 on the Fisher information matrix and the asymptotic normality of maximum likelihood [Wasserman pp. 133–134]; Cowan, *Statistical Data Analysis*, Chapters 6 and 9 on the in-

formation inequality, the variance of maximum-likelihood estimators and confidence regions [Cowan §§6.6–6.8, Ch. 9]; Casella and Berger, *Statistical Inference*, Section 7.3 [CasellaBerger §7.3]; Sivia, *Data Analysis*, Chapter 3 on the Gaussian approximation to a posterior [Sivia Ch. 3]; Hobson et al., *Bayesian Methods in Cosmology*, Chapter 5 on Fisher forecasts [HobsonBMC Ch. 5].

- **Fisher forecasts in cosmology.** Tegmark, Taylor and Heavens on the Gaussian formula and its use [192]; Knox on the CMB [166]; Jungman et al. on cosmological parameters from the CMB [196]; Coe on reading Fisher matrices [194]; Heavens and Verde on statistical methods and forecasting in cosmology [195, 197]; Planck 2018 for the instrument and the parameters [25, 186].
- **Gravitational-wave data analysis.** Finn on detection as hypothesis testing [199]; Cutler and Flanagan on the inner product and the Fisher matrix of an inspiral [87]; Barack and Cutler on LISA forecasts [84]; Vallisneri on the uses and abuses of the Fisher matrix [193]; Cutler and Vallisneri on waveform systematics [210]; Allen et al. on the matched-filter search [200]; Röver, Meyer and Christensen on an unknown noise spectrum [201]; Cornish and Littenberg on noise and glitch modelling [202]; the LIGO–Virgo guide to detector noise [83]; Usman et al. on a search and its background [211]; Zackay, Dai and Venumadhav on relative binning [213]; Speagle, Ashton et al. and Karnesis et al. on samplers [204, 205, 212].
- **LISA.** The mission proposal [216]; Robson, Cornish and Liu on the sensitivity curve [85]; Tinto and Dhurandhar on time-delay interferometry [206]; Katz et al. on fast inspiral waveforms [208]; Littenberg and Cornish on the global fit [203]; the LISA Data Challenges [207].
- **Environments and dark matter.** Yunes and Pretorius on the ppE framework [91]; Barausse, Cardoso and Pani on astrophysical environments [92]; Eda et al. on dark-matter spikes [82]; Kavanagh et al. and Coogan et al. on heated spikes and their detectability [89, 209]; Speeney et al. on relativistic spikes [88]; Kim et al. on wave dark matter [86]; Chase, López Nacir and Yunes on vector dark matter [100].
- **Cosmic birefringence and polarization on a cut sky.** Minami et al. on determining the detector angles and the birefringence angle together [76]; Minami and Komatsu, Diego-Palazuelos et al. and Eskilt and Komatsu on the measurements from Planck and WMAP [62, 79, 80]; the Planck team's measurement with the angles fixed [214]; Hivon et al. and Alonso et al. on pseudo-spectra and their spin-2 version [168, 169]; Smith on B modes that receive no E [215]; the Planck likelihood paper for the masks and multipole ranges of a real analysis [75].

Chapter summary

Before an experiment exists, we can still ask how sharp its likelihood will be. A model of the signal, a model of the noise and a guess for the truth say what the data would look like on average; the average curvature of $-\ln\mathcal{L}$ at the truth is the Fisher matrix, and its inverse is the covariance the experiment will deliver. It equals the covariance of the score, no unbiased estimator beats it, and maximum likelihood reaches it when the data are informative. For Gaussian data it has a closed form with a mean part and a covariance part. The CMB lives in the covariance part and reproduces Knox's and Planck's numbers; a gravitational wave lives in the mean part, where the formula becomes $(\partial_i h|\partial_j h)$ and says that a phase parameter is measured by the part of its signature the others cannot imitate. That one sentence turns many separate dark-matter and astrophysics questions into one: an environment is a power law in the phase, and its detectability is a single curve in the exponent. Every forecast was then reproduced against the papers that published it, and every one was tested against a Markov chain, which showed where the Gaussian picture breaks: at low signal-to-noise ratio, along degeneracies

and at prior boundaries. A twist of the CMB polarization showed the covariance part at work on a parameter whose signal sits in spectra that are zero on average: the information on the angle is the E power over the B already present, the instrument's own angle makes the problem singular until the Galaxy is added as a reference, and a thousand simulated masked skies showed the estimator free of any detectable bias, the forecast reached for a Planck-like map, and the error bars honest once the covariance knows what the mask does to B. In the pipeline

$$\text{physical model} \to \text{field or data} \to \text{summary } S \to \boxed{P(S \mid \theta) \to \text{infer physics}},$$

the Fisher matrix works the last two arrows before the data exist. Chapter 12 uses it, with Monte Carlo derivatives, to forecast $f_{\rm NL}$ from topological summaries, Chapter 11 to forecast distance measurements, and Chapter 14 collects the ways in which a forecast and a posterior part company.

The Fisher matrix

concept	operational meaning	used later in
Fisher matrix	$F_{ij} = -\mathbb{E}[\partial_i \partial_j \ln \mathcal{L}] = \mathbb{E}[S_i S_j]$ at a fiducial $\boldsymbol{\theta}_0$; needs a model, not data	Chapters 11, 12 and 14
Cramér–Rao bound	$\mathrm{Cov}\,\hat{\boldsymbol{\theta}} - \mathbf{F}^{-1} \geq 0$ for unbiased estimators; reached by linear-Gaussian models	Chapter 14
Gaussian Fisher formula	$\boldsymbol{\mu}_{,i}^{\mathsf{T}} \mathbf{C}^{-1} \boldsymbol{\mu}_{,j} + \frac{1}{2}\,\mathrm{Tr}[\mathbf{C}^{-1}\mathbf{C}_{,i}\mathbf{C}^{-1}\mathbf{C}_{,j}]$	Chapters 11 and 12
CMB Fisher matrix	$\sum_\ell \frac{(2\ell+1) f_{\rm sky}}{2} \partial_i C_\ell \partial_j C_\ell / (C_\ell + N_\ell)^2$	Chapter 13
marginal vs conditional	$\sqrt{(\mathbf{F}^{-1})_{ii}}$ vs $1/\sqrt{F_{ii}}$; the gap is a Schur complement	Chapters 11 to 13
ellipses	$\Delta\chi^2 = 2.30, 6.18$ for 68.3%, 95.4% in two parameters	Chapters 11 and 13
priors, experiments, Jacobians	add matrices; $\boldsymbol{J}^{\mathsf{T}}\mathbf{F}\boldsymbol{J}$; $\nabla g^{\mathsf{T}} \mathbf{F}^{-1} \nabla g$	Chapter 11
pivot	the scale where amplitude and slope decorrelate; amplitude errors grow away from it	Chapters 11 and 13
numerical derivatives	steps a fraction of the expected error; halve and double	Chapters 12 and 13
bias of a systematic	$\delta\boldsymbol{\theta} = \mathbf{F}^{-1}[\boldsymbol{\mu}_{,i}^{\mathsf{T}} \mathbf{C}^{-1} \delta\boldsymbol{\mu}]_i$	Chapters 11 and 13

Gravitational waves

concept	operational meaning	used later in
noise-weighted inner product	$(a\|b) = 4\,\mathrm{Re} \int \tilde{a}\tilde{b}^*/S_n \,\mathrm{d}f$; covariance of noise projections	Chapter 14
GW likelihood	$p(\boldsymbol{d} \mid \boldsymbol{\theta}) = p_n(d-h) \propto e^{-(d-h\|d-h)/2}$	Chapter 14
matched filter, SNR	$(d\|h)$; $z \sim \mathcal{N}(0,1)$ vs $\mathcal{N}(\rho,1)$, $\rho^2 = (h\|h)$	Chapter 14
unknown phase, time, PSD	two quadratures; one FFT over lags; Student-t after integrating out S_n	Chapter 14
GW Fisher matrix	$\Gamma_{ij} = (\partial_i h \| \partial_j h)$; $\sigma = 1/(\rho \|g_\perp\|_w)$; $\sigma_{\ln \mathcal{A}} = 1/\rho$	Chapter 14
ppE environment	$\delta\Psi = \beta u^b$, $b = -\frac{5}{3} - n$; friction $b = (2\gamma - 16)/3$	Chapter 14
detectability map	$3\sigma_\beta(b) u_{\rm lo}^b$ against each effect's predicted shift	Chapter 14
linearised signal	Fisher = posterior of $h \approx h_0 + \Delta\theta^i \partial_i h$; check $\frac{1}{2}(R\|R) \lesssim 0.1$ on the 1σ surface	Chapter 14
median experiment	noise-free data $d = h(\boldsymbol{\theta}_0)$: the shape of a typical posterior without its random shift	Chapter 14
waveform bias	$\delta\theta^i = (\Gamma^{-1})^{ij}(\partial_j h \| \delta h)$; model good enough if $(\delta h \| \delta h) \lesssim 1$	Chapter 14

A twist of the CMB polarization

concept	operational meaning	used later in
information on a twist	$F_{\beta\beta} = \sum_\ell (2\ell+1) f_{\rm sky}\, \boldsymbol{v}^{\mathsf{T}} A^{-1} \boldsymbol{v}/\tilde{B}$, $\boldsymbol{v} = (2C^{TE}, 2(C^{EE} - C^{BB}))$; $\approx 4C^{EE}/\tilde{B}$ per mode	Chapter 13
EB versus TB	$F^{TB}/F^{EB} = \rho^2 \tilde{E}/C^{EE}$: TB helps only where E is noisy	Chapter 13
rotation invariance	$F(\beta) = F(0)$; at $\beta \neq 0$ information moves from EB, TB into EE, BB	
depth regimes	$\sigma \propto \Delta^2$, $\propto \Delta$, then the lensing floor	Chapter 13
α–β degeneracy	null direction $(1, \ldots, 1, -1)$ without dust; dust sets $\sigma(\beta)$ when the α_ν are free	Chapter 13
$E \to B$ leakage	a cut wave is read partly as B; fraction $\propto 1/\ell$ at a sharp edge, $\propto \ell^{-2}$ with a taper; no EB on average	Chapter 13
twist through a mask	$We^{2i\beta}(Q+iU) = e^{2i\beta} W(Q+iU)$: pseudo-$E$, B rotate exactly; no MASTER needed for the angle	Chapter 13
honest error bars	weights and errors from a covariance that contains the leak: measured spectra or simulations	Chapter 13

Chapter 11

Large-Scale Structure as Data

Where this chapter is going

A galaxy redshift survey records two angles and a redshift for each galaxy. Using a fiducial cosmology to convert redshift to distance, we can place each galaxy in three dimensions. The resulting catalogue may contain hundreds of thousands or even millions of points in a volume several billion light years across. No theory predicts where any one of those points should be. What a cosmological model predicts is how the points *cluster*: how much more often than chance two galaxies sit 100 h^{-1}Mpc apart, how much the clustering depends on direction relative to our line of sight, and how both change with redshift. Every cosmological number that comes out of a galaxy survey is extracted from that clustering pattern.

Getting from the table to the numbers is a statistics problem from start to finish. The points are a random sample of an underlying continuous density field, so the first question is what sampling does to the statistics: it adds a noise term, the shot noise, whose size we can derive exactly from Poisson's law. The survey sees only part of the sky and is thinner at large distance, so the second question is how to estimate the correlation function $\xi(r)$ and the power spectrum $P(k)$ without letting the survey's edges and its varying depth leak into the answer. The answers are pair-count estimators built from a random catalogue (the Landy–Szalay estimator) and a weighted Fourier estimator (the FKP estimator); in both cases the best combination follows from writing out the first-order fluctuations and choosing the combination in which they cancel, or the weights that minimise the variance. With a good estimate of the clustering in hand, two features carry most of the cosmology. The baryon acoustic peak is a ruler of known length, so its apparent size measures distance. The squashing of the clustering along the line of sight, caused by galaxies' own motions, measures how fast structure grows. Each becomes a parameter fit, and each parameter error can be forecast with the Fisher matrix of Chapter 10 and checked with mock surveys.

The same catalogue holds more than its two-point function. Galaxies live in dark matter haloes and the cosmic web has large empty regions; the halo model and void statistics describe what these structures add. Closest to us, the galaxies' velocities are measured directly, and the local expansion rate, the flows towards nearby superclusters and the ages of passively evolving galaxies give a second, independent route to H_0. Figure 11.1 shows how the pieces connect. In the pipeline that runs through the book (physical model, field, summary, sampling distribution, inference) galaxy clustering lives in the middle three boxes: the summaries of a point process are built and their sampling distributions derived, so that Chapters 9 and 10 can turn them into parameter constraints, and the same catalogues hand Chapter 12 the cosmic web as a field whose shape the two-point function does not see.

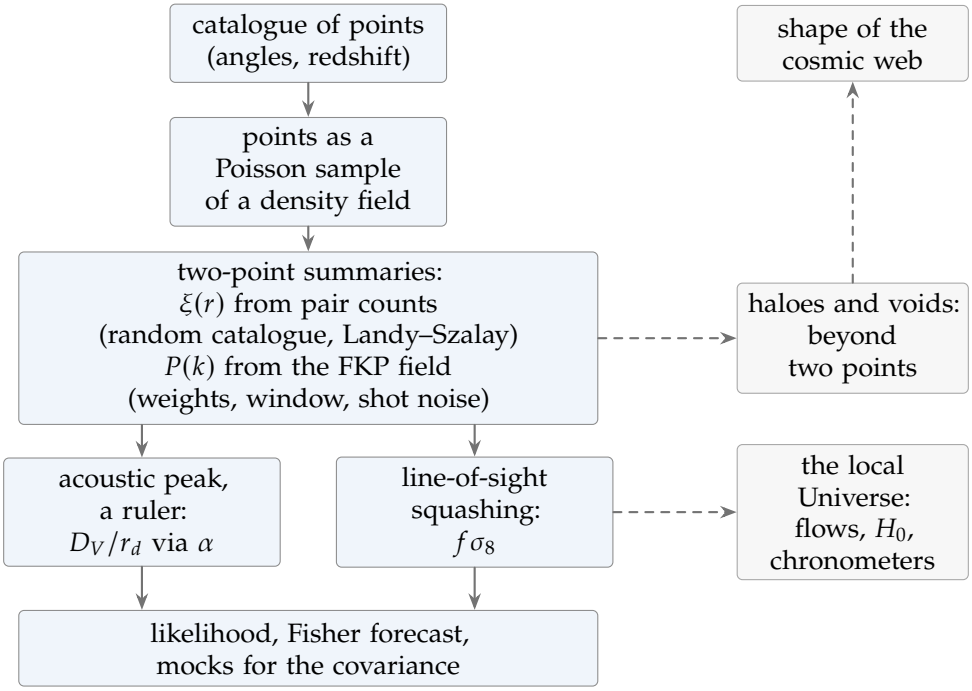


Figure 11.1: The route from a galaxy catalogue to cosmology. The main column is the two-point route: the points are treated as a random sample of a density field, their clustering is summarised by $\xi(r)$ or $P(k)$, and two features of that summary, the acoustic peak and the anisotropy caused by galaxy motions, become parameters. The grey boxes on the right are the other routes through the same data: the structures that two-point statistics miss (haloes, voids, and the shape of the web), and the local Universe, where velocities are measured directly.

11.1 Galaxies as a random sample of a smooth density field

Picture the Sloan luminous red galaxies of 2005: 46,748 points spread through $0.72\,(h^{-1}\mathrm{Gpc})^3$ [42]. That is about one galaxy per cube of $25\,h^{-1}$Mpc on a side, a very sparse sprinkling. The theory of structure formation does not talk about points at all. It talks about a smooth density contrast $\delta(\boldsymbol{x})$, the matter density divided by its mean, minus one, whose two-point function $\xi(r)$ and power spectrum $P(k)$ it predicts (Chapters T1 and 7). Between the two sits one step that every galaxy analysis takes, usually in a single sentence: the galaxies are assumed to be a random sample of the smooth field, more of them where the field is high and fewer where it is low.

That single sentence determines the noise of every clustering measurement in the chapter. If we know exactly what "random sample" means, we can say how the statistics of the points differ from the statistics of the field: how much extra scatter the counts in a cell have, why the power spectrum of the points sits above the power spectrum of the field by a constant, and when that constant swamps the signal. So the question, in words, is:

> *If galaxies are scattered at random with a probability that follows a fluctuating density, how are the pair statistics of the galaxies related to the pair statistics of the density?*

This is a question in the *predictive* direction. We start from a hypothesis, a particular density field, and ask what catalogues it can produce and how often. The measurements of the later sections run the other way. The Universe has produced one catalogue, and we ask which fields, and behind them which cosmological parameters, could have produced it. To answer that we need, for every candidate, the probability of the catalogue we actually observed, and that probability is exactly what the predictive question supplies. So the Poisson sampling worked

out here is the part of the chain that every estimator, error bar and likelihood of the chapter rests on.

11.1.1 The inputs, and the counts they imply

Cut space into small cells of volume ΔV, labelled by i, and let N_i be the number of galaxies in cell i. Write $\bar{n}$ for the mean number of galaxies per unit volume, and $\delta_g(\boldsymbol{x})$ for the density contrast of the smooth field that the galaxies follow. (On large scales δ_g is the matter contrast times a bias factor, $\delta_g = b\,\delta$; see the cosmology notes. None of the sampling results depends on that relation.) The cells are small enough that δ_g is constant across each one; call its value in cell i δ_i. We now state what we assume, separating physics from mathematics, as for the Poisson process of Section 2.5.1.

Physical inputs, all of them statements *given the field*, that is, with the whole map of δ_g held fixed:

(G1) *Rate set by the local density.* The expected number of galaxies per unit volume at $\boldsymbol{x}$ is $\bar{n}\,[1 + \delta_g(\boldsymbol{x})]$. Inside one cell this rate is constant.

(G2) *No memory between places.* The counts in non-overlapping volumes are independent. A galaxy here does not make another galaxy nearby more or less likely, beyond what the field already does.

(G3) *One at a time.* Galaxies are points; two never sit at exactly the same position.

Mathematical inputs: δ_g is smooth on the scale of a cell, and $1 + \delta_g \geq 0$ everywhere, because a rate cannot be negative.

Inside one cell, with the field held fixed, (G1)–(G3) are exactly the three inputs of Section 2.5.1, with volume in place of time: no memory, a constant rate, and events one at a time. The derivation there needed nothing else, so it applies word for word and gives

$$N_i \,|\, \delta_g \;\sim\; \text{Poisson}(\lambda_i), \qquad \lambda_i = \bar{n}\,\Delta V\,(1 + \delta_i), \tag{11.1}$$

with different cells independent of one another given the field (G2). The field itself is random, drawn once by the Universe. A point process built this way, a Poisson process whose rate is itself a random field, is called a **Cox process** or doubly stochastic Poisson process. Cosmologists call the same assumption the **local Poisson model** [Bernardeau2002 §6.3, p. 135].

Working definition 11.1 (Cox process for galaxies)

Galaxies form a **Cox process** with mean density $\bar{n}$ and contrast δ_g if, given the field $\delta_g(\boldsymbol{x})$ [one realisation of the smooth density, held fixed], the count in any small cell is Poisson with mean $\bar{n}\,\Delta V\,(1 + \delta_g)$ and counts in disjoint cells are independent. *Operationally:* draw a field, then draw a Poisson count in every cell with that cell's mean; repeating both steps generates the ensemble of possible catalogues.

11.1.2 From the inputs to counts in cells and pair counts

Two layers of randomness now act on every count: the Universe chose the field, and then the sampling chose the galaxies. To find the mean and variance of N_i we take them one layer at a time. The tool is the **law of total variance** [the spread of N is the average spread at fixed field plus the spread of the field's own prediction]: $\text{Var}(N) = \mathbb{E}[\text{Var}(N \,|\, \delta_g)] + \text{Var}(\mathbb{E}[N \,|\, \delta_g])$, and its two-variable version $\text{Cov}(N_i, N_j) = \mathbb{E}[\text{Cov}(N_i, N_j \,|\, \delta_g)] + \text{Cov}(\mathbb{E}[N_i \,|\, \delta_g], \mathbb{E}[N_j \,|\, \delta_g])$. Both

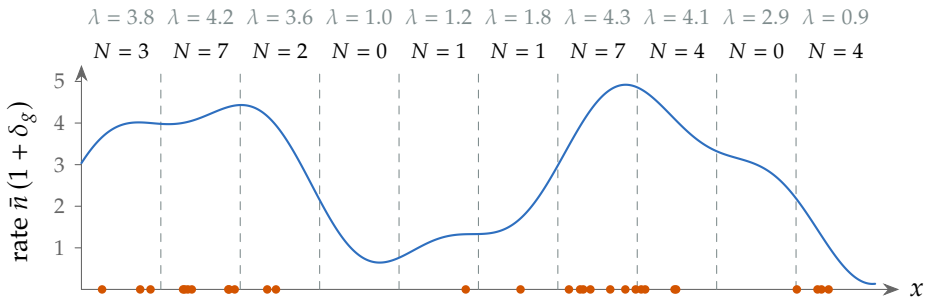


Figure 11.2: A Cox process in one dimension. The blue curve is one realisation of the smooth rate $\bar{n}(1+\delta_g)$; the orange dots are the galaxies drawn from it. Each cell (dashed lines) has an expected count λ (grey, the area under the curve in that cell) and an actual count N (black), which is a Poisson draw with that mean. Cells where the curve is high get many points, but the counts still scatter: the last cell, with $\lambda = 0.9$, happens to hold four galaxies, and the one before it, with $\lambda = 2.9$, holds none.

follow from writing $N - \mathbb{E}\,N = (N - \mathbb{E}[N \mid \delta_g]) + (\mathbb{E}[N \mid \delta_g] - \mathbb{E}\,N)$ and noticing that the cross term averages to zero, because the first bracket has zero mean at every fixed field. We also need the statistics of the field: $\mathbb{E}[\delta_i] = 0$, $\mathbb{E}[\delta_i^2] = \sigma^2$ (the variance of the field in one cell) and $\mathbb{E}[\delta_i \delta_j] = \xi(r_{ij})$, the two-point function at the separation of the two cells (Section 7.1.2).

For one cell:

$$\mathbb{E}[N_i] \overset{(1)}{=} \mathbb{E}\big[\mathbb{E}[N_i \mid \delta_g]\big] \overset{(2)}{=} \mathbb{E}\big[\bar{n}\Delta V(1+\delta_i)\big] \overset{(3)}{=} \bar{n}\Delta V, \tag{11.2}$$

$$\begin{aligned}\mathrm{Var}(N_i) &\overset{(4)}{=} \mathbb{E}\big[\mathrm{Var}(N_i \mid \delta_g)\big] + \mathrm{Var}\big(\mathbb{E}[N_i \mid \delta_g]\big) \overset{(5)}{=} \mathbb{E}\big[\bar{n}\Delta V(1+\delta_i)\big] + \mathrm{Var}\big(\bar{n}\Delta V \delta_i\big) \\ &\overset{(6)}{=} \bar{n}\Delta V + (\bar{n}\Delta V)^2 \sigma^2. \end{aligned} \tag{11.3}$$

(1) Average first over the galaxies at fixed field, then over fields (the law of iterated expectations).
(2) (11.1): the conditional mean of a Poisson count is its rate.
(3) δ_i has zero mean.
(4) The law of total variance.
(5) A Poisson variable has variance equal to its mean; the constant $\bar{n}\Delta V$ in the conditional mean does not vary and drops out of the second variance.
(6) $\mathbb{E}[\delta_i] = 0$ and $\mathrm{Var}(\delta_i) = \sigma^2$.

The variance of a count has two parts. The first, $\bar{n}\Delta V$, is what a perfectly uniform field would give: pure Poisson scatter, the shot noise. The second grows as the square of the mean count and is the clustering. Their ratio is $\bar{n}\Delta V \sigma^2$, the mean count times the field variance. In a sparse survey or a small cell, shot noise wins; in a dense survey or a big cell, clustering wins.

For two different cells $i \neq j$ the same steps give

$$\begin{aligned}\mathrm{Cov}(N_i, N_j) &\overset{(1)}{=} \mathbb{E}\big[\mathrm{Cov}(N_i, N_j \mid \delta_g)\big] + \mathrm{Cov}\big(\mathbb{E}[N_i \mid \delta_g], \mathbb{E}[N_j \mid \delta_g]\big) \\ &\overset{(2)}{=} 0 + (\bar{n}\Delta V)^2\, \mathbb{E}[\delta_i \delta_j] \overset{(3)}{=} (\bar{n}\Delta V)^2\, \xi(r_{ij}), \end{aligned} \tag{11.4}$$

$$\mathbb{E}[N_i N_j] \overset{(4)}{=} \mathrm{Cov}(N_i, N_j) + \mathbb{E}[N_i]\,\mathbb{E}[N_j] = (\bar{n}\Delta V)^2 \big[1 + \xi(r_{ij})\big]. \tag{11.5}$$

(1) The two-variable law of total variance.
(2) (G2): at fixed field the two counts are independent, so their conditional covariance is zero. The conditional means are $\bar{n}\Delta V(1+\delta_i)$ and $\bar{n}\Delta V(1+\delta_j)$, whose covariance is $(\bar{n}\Delta V)^2\,\mathrm{Cov}(\delta_i, \delta_j)$.
(3) The covariance of the field at two points a distance r_{ij} apart is the two-point function.

(4) Definition of covariance, and (11.2).

Read (11.5) with cells so small that each holds at most one galaxy. Then $N_i N_j$ is 1 when both cells hold a galaxy and 0 otherwise, so $\mathbb{E}[N_i N_j]$ is the probability of finding a galaxy in each of two volume elements a distance r apart:

$$\mathrm{d}\,\mathrm{P}_{\mathrm{pair}} = \bar{n}^2 \left[1 + \xi(r)\right] \mathrm{d}V_1\, \mathrm{d}V_2. \tag{11.6}$$

This is the classical definition of the galaxy correlation function [217]: $\xi(r)$ is the excess probability, over a random distribution, of finding a pair at separation r. Put as a question, $\xi(r)$ asks how strongly the density at one place is tied to the density at another place a distance r away. Under (G1)–(G3) it is not a separate definition. The pair excess of the *points* is the two-point function of the *field*, exactly, with no shot-noise correction, as long as the two volume elements are distinct. Shot noise lives only at zero separation. That one fact explains the designs of all the estimators of Sections 11.2 and 11.3.

Example: counts in cells. Take a cubic cell of side R, so that $\bar{N} = \bar{n}R^3$, and let σ_R^2 be the variance of the field averaged over that cube. Summing (11.3) and (11.4) over all the small cells inside the cube,

$$\mathrm{Var}(N) \overset{(1)}{=} \sum_i \mathrm{Var}(N_i) + \sum_{i\neq j} \mathrm{Cov}(N_i, N_j) \overset{(2)}{=} \bar{N} + (\bar{n}\Delta V)^2 \sum_{i,j} \xi(r_{ij}) \overset{(3)}{=} \bar{N} + \bar{N}^2 \sigma_R^2, \tag{11.7}$$

(1) The variance of a sum is the sum of all variances and covariances.

(2) (11.3) and (11.4); the diagonal σ^2 terms join the off-diagonal ξ terms because $\xi(0) = \sigma^2$, and the $\bar{n}\Delta V$ terms add up to $\bar{N}$.

(3) The variance of the cube average, $\sigma_R^2 = \mathrm{Var}\big(\frac{1}{n_c}\sum_i \delta_i\big) = \frac{1}{n_c^2}\sum_{i,j}\xi(r_{ij})$ with n_c cells in the cube, and $(\bar{n}\Delta V)^2 n_c^2 = \bar{N}^2$.

So the ratio of variance to mean, the Fano factor, is $1 + \bar{N}\sigma_R^2$. A cube of side $6.25\,h^{-1}$Mpc in a survey with $\bar{n} = 10^{-3}\,h^3\mathrm{Mpc}^{-3}$ holds on average $\bar{N} = 0.244$ galaxies. For the field of our simulation below, the predicted Fano factor there is 2.173: the clustering term is only a little larger than the shot noise. A cube of side $50\,h^{-1}$Mpc holds 125.1 galaxies on average, and the predicted ratio is 19.62: the clustering term is about twenty times the shot noise. The same field looks Poisson-dominated in small cells and clustering-dominated in big ones. In contrast form, with $\delta_N = N/\bar{N} - 1$, (11.7) reads $\mathrm{Var}(\delta_N) = 1/\bar{N} + \sigma_R^2$ [Bernardeau2002 eq. (374)].

11.1.3 The power spectrum of points: the shot-noise term

The same reasoning in Fourier space gives the most used formula in galaxy clustering. We work on the periodic grid of Section 7.4.1: a box of volume V cut into cells of volume ΔV, Fourier modes $\delta_{\boldsymbol{k}} = \Delta V \sum_{\boldsymbol{x}} \delta(\boldsymbol{x}) e^{-i\boldsymbol{k}\cdot\boldsymbol{x}}$, and the convention $\mathbb{E}\big[\delta_{\boldsymbol{k}}\delta^*_{\boldsymbol{k}'}\big] = VP(k)$ when $\boldsymbol{k} = \boldsymbol{k}'$ and zero otherwise. The density contrast of the galaxies in cell $\boldsymbol{x}$ is $\delta_n(\boldsymbol{x}) = N(\boldsymbol{x})/(\bar{n}\Delta V) - 1$. Dividing (11.3) and (11.4) by $(\bar{n}\Delta V)^2$, its covariance between cells $\boldsymbol{x}$ and $\boldsymbol{y}$ is $\xi(\boldsymbol{x}-\boldsymbol{y}) + \delta^K_{\boldsymbol{xy}}/(\bar{n}\Delta V)$, where the Kronecker δ^K is 1 for the same cell and 0 otherwise. Then

$$\begin{aligned}\mathbb{E}\big[|\delta_{n,\boldsymbol{k}}|^2\big] &\overset{(1)}{=} \Delta V^2 \sum_{\boldsymbol{x},\boldsymbol{y}} \left[\xi(\boldsymbol{x}-\boldsymbol{y}) + \frac{\delta^K_{\boldsymbol{xy}}}{\bar{n}\Delta V}\right] e^{-i\boldsymbol{k}\cdot(\boldsymbol{x}-\boldsymbol{y})} \\ &\overset{(2)}{=} \Delta V^2 \frac{V}{\Delta V} \sum_{\boldsymbol{r}} \left[\xi(\boldsymbol{r}) + \frac{\delta^K_{\boldsymbol{r0}}}{\bar{n}\Delta V}\right] e^{-i\boldsymbol{k}\cdot\boldsymbol{r}} \overset{(3)}{=} V\left[P(k) + \frac{1}{\bar{n}}\right].\end{aligned} \tag{11.8}$$

(1) Insert the definition of $\delta_{n,\boldsymbol{k}}$ twice and take the expectation inside the double sum ($\boldsymbol{k} \neq \boldsymbol{0}$, so the mean of δ_n drops out).
(2) Statistical homogeneity: the bracket depends only on $\boldsymbol{r} = \boldsymbol{x} - \boldsymbol{y}$, so the sum over $\boldsymbol{y}$ at fixed $\boldsymbol{r}$ gives the number of cells, $V/\Delta V$.
(3) $\Delta V \sum_{\boldsymbol{r}} \xi(\boldsymbol{r}) e^{-i\boldsymbol{k}\cdot\boldsymbol{r}} = P(k)$ is the grid Wiener–Khinchin relation (Section 7.3.2); the Kronecker term picks out $\boldsymbol{r} = \boldsymbol{0}$ and leaves $\Delta V \cdot 1/(\bar{n}\Delta V) = 1/\bar{n}$.

Key idea: Shot noise

For galaxies that are a Cox sample of a field with spectrum $P(k)$, the power spectrum of the galaxy density contrast is

$$P_{\mathrm{gal}}(k) = P(k) + \frac{1}{\bar{n}}.$$

The added constant is the power spectrum of pure randomness: it comes entirely from each galaxy being paired with itself, and it is the same at every k.

Why is the shot noise flat? In real space it lives at a single point, the zero separation of a galaxy with itself, and the Fourier transform of a spike is a constant. The size of the spike, $1/\bar{n}$, is the volume per galaxy: the sparser the sample, the larger the noise. The quantity that decides whether a survey measures the field or the noise at wavenumber k is therefore $\bar{n}P(k)$, the signal-to-noise per mode. It recurs in every formula of Sections 11.3 and 11.4.

Example: a field with no structure. If $\delta_g = 0$ everywhere, the catalogue is a homogeneous Poisson process and (11.8) gives $P_{\mathrm{gal}} = 1/\bar{n}$ at every k. With $\bar{n} = 10^{-3}\,h^3\mathrm{Mpc}^{-3}$ that is $1000\,h^{-3}\mathrm{Mpc}^3$. A uniform random catalogue is not "unclustered" in its measured power spectrum: it has exactly this flat power.

Example: how sparse is too sparse? Eisenstein and collaborators chose luminous red galaxies at $\bar{n} \approx 10^{-4}\,h^3\mathrm{Mpc}^{-3}$, so $1/\bar{n} = 10^4\,h^{-3}\mathrm{Mpc}^3$. They quote the galaxy power as $10^4\,h^{-3}\mathrm{Mpc}^3$ at $k \approx 0.15\,h\,\mathrm{Mpc}^{-1}$ and 4×10^4 at $k \approx 0.05$ [Eisenstein2005 §2]. So $\bar{n}P \approx 1$ at the smallest scales of the acoustic feature and $\bar{n}P \approx 4$ on the larger scales. A survey ten times denser would gain little on large scales, where the error is already dominated by the finite number of Fourier modes in the volume, and a survey ten times sparser would lose most of its signal to noise. Sparse, luminous, and spread over a huge volume is the right design for large-scale clustering, which is what Section 11.3 will make precise with the effective volume.

11.1.4 Simulating a Cox process: lognormal fields

To test (11.8) we need a field to sample. A Gaussian field from Section 7.4 will not do. Galaxy fields on the scale of a few Mpc have variance of order one, so a Gaussian δ_g dips below -1 in a sizeable fraction of cells, and the input $1 + \delta_g \geq 0$ fails: a Poisson count cannot have a negative mean. The standard fix is the **lognormal field** [55]. Take a Gaussian field G with zero mean, variance σ_G^2 and two-point function $\xi_G(r)$, and define

$$1 + \delta_g(\boldsymbol{x}) = \exp\!\left[G(\boldsymbol{x}) - \tfrac{1}{2}\sigma_G^2\right], \tag{11.9}$$

which is positive by construction. What two-point function does δ_g have? For two points with $G_1 = G(\boldsymbol{x}_1)$, $G_2 = G(\boldsymbol{x}_2)$,

$$\mathbb{E}\big[(1+\delta_1)(1+\delta_2)\big] \overset{(1)}{=} e^{-\sigma_G^2}\,\mathbb{E}\big[e^{G_1+G_2}\big] \overset{(2)}{=} e^{-\sigma_G^2}\,e^{\frac{1}{2}\mathrm{Var}(G_1+G_2)} \overset{(3)}{=} e^{-\sigma_G^2}\,e^{\sigma_G^2+\xi_G(r)} = e^{\xi_G(r)},$$
$$\xi(r) \overset{(4)}{=} \mathbb{E}\big[(1+\delta_1)(1+\delta_2)\big] - 1 = e^{\xi_G(r)} - 1. \tag{11.10}$$

(1) Multiply out the two exponentials of (11.9).
(2) $G_1 + G_2$ is Gaussian with zero mean, and for a Gaussian Y with zero mean, $\mathbb{E}[e^Y] = e^{\mathrm{Var}(Y)/2}$ (the moment generating function of Chapter 2 at $t = 1$).
(3) $\mathrm{Var}(G_1 + G_2) = \sigma_G^2 + \sigma_G^2 + 2\xi_G(r)$.
(4) $\mathbb{E}[\delta_1] = \mathbb{E}[\delta_2] = 0$ (set $\boldsymbol{x}_1 = \boldsymbol{x}_2$ in the first line with one factor: $\mathbb{E}[1 + \delta] = e^{-\sigma_G^2/2} e^{\sigma_G^2/2} = 1$), so the cross terms vanish.

So to make a lognormal field with a target $\xi(r)$, we build the Gaussian field with $\xi_G = \ln(1 + \xi)$ [Xavier2016 eq. (7)], compute its spectrum P_G by a Fourier transform on the grid, and draw it as in Section 7.4.2. When the target is a realistic galaxy spectrum, the transformed P_G can be slightly negative at a few high wavenumbers (no Gaussian field has exactly the required ξ_G); we set those to zero. For the fields below this touches 1.78% of the modes. The lognormal is not the true distribution of galaxies, but it has the right two-point function, it is positive, and it is skewed in the same direction as the real density, which is all a test of (11.8) requires.

Listing 11.1: Lognormal field and Cox sampling (`code/ch11/lib_lss.py`)

```
class Lognormal:
    """Lognormal field 1 + delta = exp(G - sigma_G^2/2) with a prescribed grid two-point function.

    xi_target is the two-point function the lognormal field should have on the grid. Because
    <(1+d1)(1+d2)> = exp(xi_G(r)) for jointly Gaussian G (the mean of exp of a Gaussian), the
    Gaussian field needs xi_G = ln(1 + xi_target) [Xavier et al. 2016, eq. 7 with shift 1].
    Its spectrum P_G is the FFT of xi_G; rare negative values are clipped to zero.
    """

    def __init__(self, xi_target_grid, N, L, d):
        self.N, self.L, self.d = N, L, d
        D = L / N
        xg = np.log1p(xi_target_grid)
        pg = np.fft.rfftn(xg).real * D ** d          # P_G on the rfft grid
        self.clipped = float(np.mean(pg < 0))
        self.pg = np.clip(pg, 0, None)
        self.pg.flat[0] = 0.0
        w = rfft_weights(N, d)
        self.sigma2 = float(np.sum(w * self.pg) / L ** d)   # variance of G on the grid

    def draw(self, rng, nsim=None, return_gauss=False):
        g = gaussian_from_pgrid(self.pg, self.N, self.L, self.d, rng, nsim)
        out = np.expm1(g - 0.5 * self.sigma2)
        return (out, g) if return_gauss else out


def xi_box_from_pk(Pfun, N, L, d):
    """Exact grid two-point function (1/V) sum_k P(k) e^{ik.r} of a field with spectrum P on the grid."""
    D = L / N
    kmag, _ = kgrid(N, L, d, real=False)
    with np.errstate(divide="ignore", invalid="ignore"):
        pk = np.where(kmag > 0, Pfun(np.where(kmag > 0, kmag, 1.0)), 0.0)
    return np.real(np.fft.ifftn(pk)) / D ** d


def poisson_sample(density, nbar, L, rng, window=None):
    """Counts per cell of a Cox process: N_cell ~ Poisson(nbar * dV * W * (1 + delta)).

    density : 1 + delta on the grid (must be >= 0); window: optional W(x) in [0, 1] (or n(x)/nbar).
    """
```

```
    d = density.ndim
    dV = (L / density.shape[0]) ** d
    lam = nbar * dV * density * (1.0 if window is None else window)
    return rng.poisson(np.clip(lam, 0, None))
```

The class `Lognormal` does the three steps just described: line 142 takes $\ln(1+\xi)$, line 143 Fourier transforms it into P_G (the factor ΔV turns the grid sum into the continuum normalisation of Section 7.4.1), and lines 144–146 clip and record the negative modes. The method `draw` makes a Gaussian field with that spectrum and applies (11.9). `poisson_sample` is (11.1) in one line: every cell gets a Poisson count with mean $\bar{n}\Delta V(1+\delta)$, optionally multiplied by a survey window, which we use in the next two sections.

11.1.5 Checking the shot noise and the counts in cells

The script `code/ch11/01_poisson_sampling.py` draws 10 lognormal fields in a box of side $600\,h^{-1}$Mpc with 96^3 cells. Their target spectrum is $b^2P_{\rm lin}(k)$ with $b = 1.5$ and $P_{\rm lin}$ the CAMB linear spectrum at $z = 0$ (Section T1.4). Each field is sampled at three densities, $\bar{n} = 10^{-4}, 10^{-3}, 10^{-2}\,h^3\mathrm{Mpc}^{-3}$.

Listing 11.2: Sampling one field at three densities (`code/ch11/01_poisson_sampling.py`)

```
for s in range(NSIM):
    delta = ln.draw(rng)
    p_field.append(kb.avg(power_raw(delta, L)))
    for nb in NBARS:
        c = poisson_sample(1 + delta, nb, L, rng)
        dn = c / (nb * dV) - 1                  # density contrast of the points
        p_pts[nb].append(kb.avg(power_raw(dn, L)))
        p_eps[nb].append(kb.avg(power_raw(dn - delta, L)))   # the sampling noise alone
        if nb == 1e-3:
            for m in cic:                       # counts in cubes of m^3 cells
                cs = c.reshape(N // m, m, N // m, m, N // m, m).sum(axis=(1, 3, 5))
                ds = (1 + delta).reshape(N // m, m, N // m, m, N // m, m).mean(axis=(1, 3, 5)) - 1
                cic[m].append((cs.mean(), cs.var(), ds.var()))
    cu = poisson_sample(np.ones_like(delta), 1e-3, L, rng)
    p_unif.append(kb.avg(power_raw(cu / (1e-3 * dV) - 1, L)))
```

Line 42 turns the counts into the galaxy contrast δ_n. Line 43 stores its shell-averaged power, line 44 the power of $\delta_n - \delta_g$, which is the sampling noise alone, and lines 46–49 add up the counts in cubes of 1, 2, 4 and 8 cells on a side for the counts-in-cells test. Lines 50–51 sample a perfectly uniform field for comparison.

Figure 11.3 shows the result. At every density the galaxy power is the field power lifted by $1/\bar{n}$, and at $\bar{n} = 10^{-4}$ the lift is so large that beyond $k \approx 0.1\,h\,\mathrm{Mpc}^{-1}$ the galaxies tell us almost nothing about the field. The power of the sampling noise alone, multiplied by $\bar{n}$, averages 0.987, 0.981 and 1.000 at the three densities, against the predicted 1; the uniform field gives 0.982.

Figure 11.4 checks (11.7). In the smallest cubes the measured ratio is 2.170 against the predicted 2.173; in the largest, 19.59 against 19.62. Both differ from the prediction by less than the scatter between realisations.

Where the inputs break. Each input can fail, and each failure has a signature in the data.

- (G2) fails on small scales. Two galaxies cannot overlap, and two massive dark matter haloes cannot sit closer than their own radii. That exclusion makes counts in small cells *less* variable than Poisson, and the effective shot noise smaller than $1/\bar{n}$ [Bernardeau2002 p. 136, footnote 60].

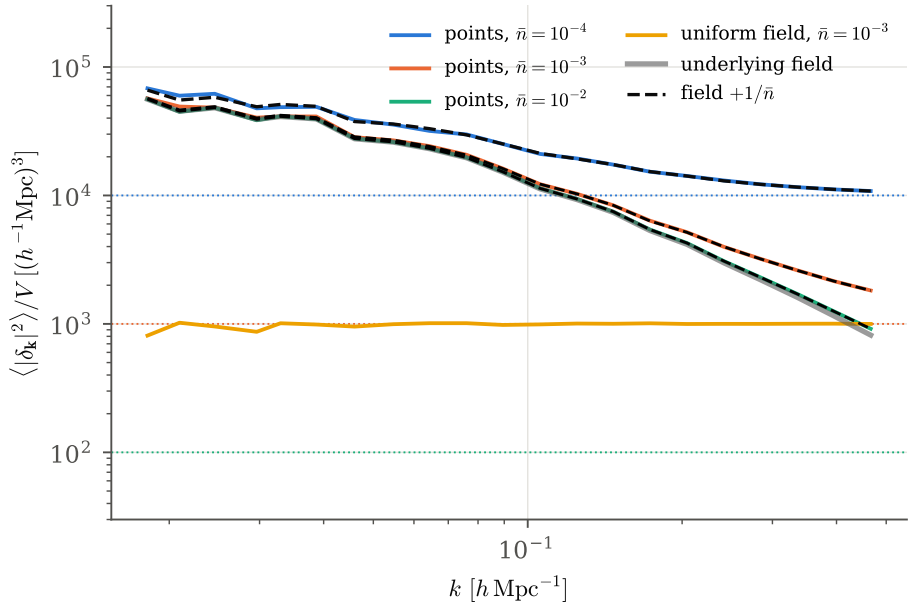


Figure 11.3: Power spectra of Poisson samples of one lognormal field. The grey band is the field itself. Each coloured curve is the galaxy power at one mean density, and the dashed black curves are the field plus $1/\bar{n}$; the dotted horizontal lines mark $1/\bar{n}$. The sparsest sample (blue) is dominated by the flat noise beyond $k \approx 0.1\,h\,\mathrm{Mpc}^{-1}$. The yellow curve is a uniform field sampled at $\bar{n} = 10^{-3}$: flat at $1000\,h^{-3}\mathrm{Mpc}^3$.

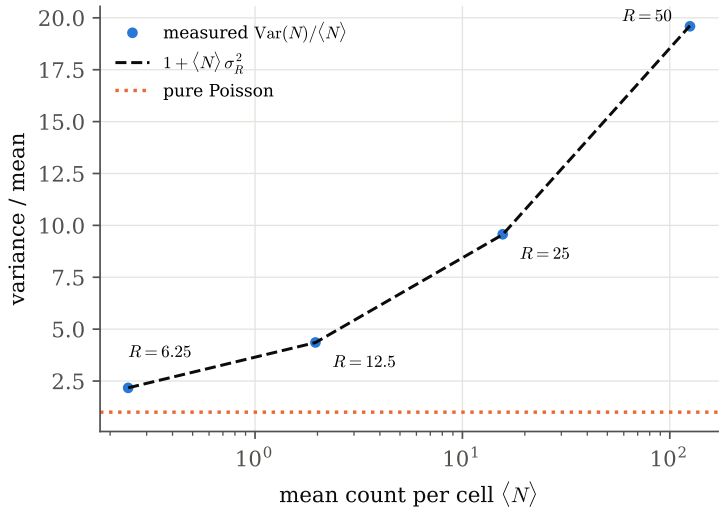


Figure 11.4: Variance over mean of the counts in cubes of side $R = 6.25$ to $50\,h^{-1}\mathrm{Mpc}$, for $\bar{n} = 10^{-3}\,h^3\mathrm{Mpc}^{-3}$. The dots are the measured ratios, the dashed curve is $1 + \bar{N}\sigma_R^2$ from (11.7) with σ_R^2 measured from the field in the same cubes, and the dotted line is the pure Poisson value 1.

- (G2) also fails in the opposite direction. Several galaxies often share one halo, so finding one makes another nearby more likely, beyond what the smooth field predicts. In the halo model this appears as an extra, nearly constant term in $P(k)$ that looks like shot noise but is not $1/\bar{n}$, and the number of galaxies in low-mass haloes is closer to binomial than Poisson [107]. Describing galaxies through the haloes they live in is the job of the halo model.

- (G1) assumes the rate depends only on the local density. If galaxy formation also depends on the environment on larger scales, the relation $\delta_g = b\delta$ becomes scale-dependent, which shows up as a bias that changes with k.

In practice the measured shot noise of a sample is a free parameter of the fit, allowed to differ from $1/\bar{n}$ by tens of percent. On the scales of the acoustic peak, where the following sections work, the Cox model is an excellent description.

11.2 Estimating the correlation function from pair counts

The Sloan survey covers an irregular patch of sky, with holes around bright stars and boundaries set by the telescope's coverage. Its observed galaxy density also falls with distance as faint galaxies drop out of the sample. The pair probability (11.6) of Section 11.1 tells us what to measure: count the pairs of galaxies at separation r, and compare with the number we would have found if the galaxies had no clustering at all. The difficulty is the comparison. In this oddly shaped survey, with its holes and its falling density, how many pairs at separation r would an *unclustered* set of galaxies have produced? Near an edge a galaxy has fewer partners at large r simply because part of its surroundings was never observed, and that loss has nothing to do with cosmology.

Nobody computes that number analytically for a real survey. Instead we *simulate* it: scatter a large number of points uniformly through the same survey volume, with exactly the same angular mask and the same fall-off with distance, and count their pairs. Such a set of points is a **random catalogue**. It is Monte Carlo integration (Chapter 6) of the survey's pair geometry. The rest of the section asks how best to combine the galaxy–galaxy, galaxy–random and random–random pair counts. Four combinations are in use, and they behave very differently near edges; the reason fits in one expansion.

11.2.1 Pair counts and four estimators

Let the data catalogue hold N_d galaxies and the random catalogue N_r points. For a separation bin $[r, r+\Delta r)$ define the **normalised pair counts**

$$DD(r) = \frac{\#\{\text{ordered galaxy pairs in the bin}\}}{N_d(N_d-1)}, \qquad DR(r) = \frac{\#\{\text{galaxy–random pairs in the bin}\}}{N_d N_r},$$
$$RR(r) = \frac{\#\{\text{ordered random pairs in the bin}\}}{N_r(N_r-1)}. \tag{11.11}$$

Each is the fraction of all possible pairs of its kind that fall in the bin. Dividing by the number of possible pairs makes the three comparable even though N_r is several times N_d.

Working definition 11.2 (pair-count estimators of $\xi(r)$)

With DD, DR, RR as in (11.11) [fractions of all possible pairs that fall in the separation bin],

$$\hat{\xi}_{\mathrm{N}} = \frac{DD}{RR} - 1 \quad \text{natural [218]}, \qquad \hat{\xi}_{\mathrm{DP}} = \frac{DD}{DR} - 1 \quad \text{Davis–Peebles [219]},$$

$$\hat{\xi}_{\mathrm{H}} = \frac{DD\,RR}{DR^2} - 1 \quad \text{Hamilton [220]}, \qquad \hat{\xi}_{\mathrm{LS}} = \frac{DD - 2DR + RR}{RR} \quad \text{Landy–Szalay [221]}.$$

Operationally: build a random catalogue with the survey's mask and selection, count the three kinds of pairs in each bin with a tree code, and combine.

All four are motivated by the same idea. In an unclustered survey DD, DR and RR would agree on average, because galaxies and randoms would then be samples of the same distribution. Clustering raises DD above the other two by the factor $1 + \xi$. The estimators differ only in which of the noisy, edge-affected counts they divide by, and that turns out to matter a great deal.

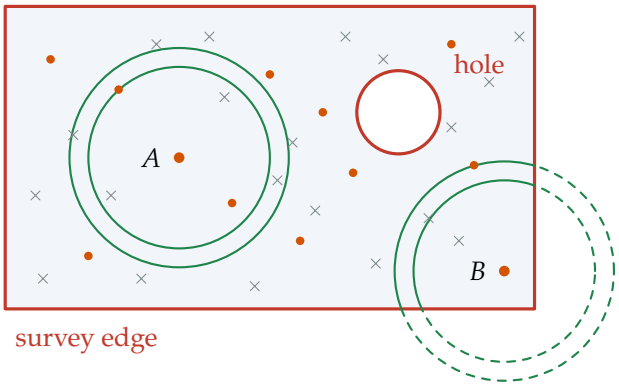


Figure 11.5: Why edges matter. Orange dots are galaxies, grey crosses are random points, the blue region is the observed survey and the red curves are its boundary. Galaxy A sits in the interior: its whole shell of separations $[r, r + \Delta r)$ (the green annulus) is observed. Galaxy B sits in a corner: most of its shell (dashed) lies outside the survey, so it contributes few pairs at this separation whatever the clustering. Random points near B lose the same fraction of their shell, which is what the random catalogue is for.

11.2.2 What each estimator does to first order

To compare the estimators we write every count as its noise-free value plus a fluctuation, and keep track of which fluctuations survive in each combination. This is the calculation of [221], set out here in one notation for all four estimators.

The fields. Let $W(\boldsymbol{x})$ be the survey window, 1 inside the observed region and 0 outside, and $A = \int W \,\mathrm{d}\boldsymbol{x}$ its area (or volume). The galaxy number density, smoothed over a scale much smaller than the bin, is $n_d(\boldsymbol{x}) = (N_d/A)\,W(\boldsymbol{x})\,[1 + \delta_d(\boldsymbol{x})]$. This defines the **data fluctuation** δ_d, which holds the clustering, the Poisson noise of the sampling, and nothing else. Because the normalisation uses the actual number N_d, the window average of δ_d is exactly zero: $\int W\delta_d \,\mathrm{d}\boldsymbol{x} = 0$. The randoms are written the same way, with a fluctuation δ_r that holds only their own Poisson noise.

The pair geometry. Let $\Theta(\boldsymbol{x},\boldsymbol{y})$ be 1 when $|\boldsymbol{x}-\boldsymbol{y}|$ falls in the bin and 0 otherwise, and let

$$G=\frac{1}{A^2}\int\!\!\int W(\boldsymbol{x})W(\boldsymbol{y})\Theta(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y}$$

be the fraction of all pairs of points in the survey that fall in the bin. For any two functions u, v write the **pair average**

$$[u\,v]=\frac{1}{A^2G}\int\!\!\int W(\boldsymbol{x})W(\boldsymbol{y})\Theta(\boldsymbol{x},\boldsymbol{y})\,u(\boldsymbol{x})\,v(\boldsymbol{y})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y},$$

and the **bin-weighted mean** $a_d=[\delta_d\,1]$, which is the average of δ_d over all pair members in the bin; $a_r=[\delta_r\,1]$ likewise.

The counts are then pair averages of products of densities (the self-pairs do not enter, since $r>0$):

$$\begin{aligned}DD&\overset{(1)}{=}\frac{1}{A^2}\int\!\!\int WW\Theta\,(1+\delta_d)(1+\delta_d)\overset{(2)}{=}G\big(1+2a_d+[\delta_d\delta_d]\big),\\ DR&\overset{(3)}{=}G\big(1+a_d+a_r+[\delta_d\delta_r]\big),\qquad RR\overset{(4)}{=}G\big(1+2a_r+[\delta_r\delta_r]\big).\end{aligned}\tag{11.12}$$

(1) The fraction of galaxy pairs in the bin is the pair integral of the normalised density $n_d/(N_d/A)=W(1+\delta_d)$ at both ends.
(2) Multiply out: the $1\cdot1$ term gives G; the two cross terms are each $G\,a_d$, by symmetry of Θ; the last is $G[\delta_d\delta_d]$.
(3) The same with $1+\delta_r$ at one end.
(4) The same with $1+\delta_r$ at both ends.

The first-order quantities are a_d and a_r; the pair averages $[\delta\,\delta]$ are second order. The one we want is $[\delta_d\delta_d]$, whose expectation is $\xi(r)$ up to the small correction derived below. Now insert (11.12) into each estimator and expand to first order in a_d, a_r, keeping the second-order pair averages:

$$\begin{aligned}\hat\xi_{\mathrm{N}}&\overset{(1)}{=}\frac{1+2a_d+[\delta_d\delta_d]}{1+2a_r+[\delta_r\delta_r]}-1\approx[\delta_d\delta_d]-[\delta_r\delta_r]+2(a_d-a_r),\\ \hat\xi_{\mathrm{DP}}&\overset{(2)}{=}\frac{1+2a_d+[\delta_d\delta_d]}{1+a_d+a_r+[\delta_d\delta_r]}-1\approx[\delta_d\delta_d]-[\delta_d\delta_r]+(a_d-a_r),\\ \hat\xi_{\mathrm{H}}&\overset{(3)}{=}\frac{(1+2a_d+[\delta_d\delta_d])(1+2a_r+[\delta_r\delta_r])}{(1+a_d+a_r+[\delta_d\delta_r])^2}-1\approx\big[(\delta_d-\delta_r)(\delta_d-\delta_r)\big]-(a_d-a_r)^2,\\ \hat\xi_{\mathrm{LS}}&\overset{(4)}{=}\frac{[\delta_d\delta_d]-2[\delta_d\delta_r]+[\delta_r\delta_r]}{1+2a_r+[\delta_r\delta_r]}=\frac{\big[(\delta_d-\delta_r)(\delta_d-\delta_r)\big]}{1+2a_r+[\delta_r\delta_r]}.\end{aligned}\tag{11.13}$$

(1) The common factor G cancels; to first order $1/(1+\epsilon)\approx1-\epsilon$.
(2) The same, with DR in the denominator: the first-order terms are $2a_d-(a_d+a_r)=a_d-a_r$.
(3) Keeping only the a's, the numerator minus the denominator is $1+2a_d+2a_r+4a_da_r-(1+a_d+a_r)^2=-(a_d-a_r)^2$: the first-order terms cancel identically. The pair averages enter as $[\delta_d\delta_d]+[\delta_r\delta_r]-2[\delta_d\delta_r]$.
(4) In $DD-2DR+RR$ every term of $G(1+\dots)$ that is zeroth or first order cancels: $1-2+1=0$ and $2a_d-2(a_d+a_r)+2a_r=0$. What is left is the pair average of the product of the two *differences* $\delta_d-\delta_r$.

So the natural estimator carries the first-order term $2(a_d-a_r)$, the Davis–Peebles estimator carries half of it, and the Hamilton and Landy–Szalay estimators carry none. What is a_d, physically? It is the average of the data fluctuation over the members of all pairs in the bin.

Every point of the survey is weighted by the fraction of its shell that lies inside the survey. Interior points (like A in Figure 11.5) get full weight; points near an edge (like B) get less. The plain survey average of δ_d is zero by construction, but this *edge-weighted* average is not. If the interior of the survey happens to be overdense and the edges underdense, a_d is positive. It is a large-scale fluctuation, of first order in δ, and at the separations where ξ is small it is far larger than the signal.

Key idea: Why Landy–Szalay wins

The Landy–Szalay numerator is the pair average of $(\delta_d - \delta_r)(\delta_d - \delta_r)$: the product of the data fluctuation and the random fluctuation at the two ends of each pair. It contains no term linear in either fluctuation, so the edge-weighted large-scale fluctuations a_d and a_r that dominate the error of $DD/RR - 1$ and $DD/DR - 1$ at large separations cancel exactly. What remains is the scatter of the pair average itself, which no pair-count estimator can avoid.

In the weak-correlation limit Landy and Szalay showed that this residual scatter is just the Poisson scatter of the pair counts, the smallest possible, while the variances of DD/RR and DD/DR are considerably larger than Poisson [LandySzalay1993 §4]. The Hamilton estimator also cancels the first-order term, only through a product rather than a difference, and it performs almost as well; a systematic comparison of nine estimators on large simulations ranks the two together at the top [222] [Bernardeau2002 §6.4.1, p. 139].

How much does the edge term cost? The claim that LS reaches the smallest scatter can be checked in the simplest case, where both errors can be computed by hand: galaxies with no clustering at all ($\xi = 0$), and a random catalogue so large that its own noise vanishes ($\delta_r = 0$, so $a_r = 0$). Then LS reduces to $[\delta_d\delta_d]$ and the natural estimator to $[\delta_d\delta_d] + 2a_d$. Two questions remain. How much does the pair average $[\delta_d\delta_d]$ scatter? And how much does a_d?

The pair term. The bin contains on average $N_p = \frac{1}{2}N_d^2 G$ distinct galaxy pairs. With no clustering, whether a given pair lands in the bin is a rare event that hardly depends on the other pairs, so the count of pairs scatters like a Poisson count (Section 2.5.1), with fractional standard deviation $N_p^{-1/2}$. Since $DD = G(1 + [\delta_d\delta_d])$ when a_d is set aside, that fractional scatter is the scatter of $[\delta_d\delta_d]$: its variance is $1/N_p$.

The edge term. Write $e(\boldsymbol{x})$ for the fraction of the survey that lies in the shell around $\boldsymbol{x}$, divided by its average over the survey; $e = 1$ on average, $e < 1$ near edges and holes. Then

$$a_d \overset{(1)}{=} \frac{1}{A}\int W(\boldsymbol{x})\,\delta_d(\boldsymbol{x})\,e(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \overset{(2)}{=} \frac{1}{N_d}\sum_{i=1}^{N_d} e(\boldsymbol{x}_i) - 1,$$
$$\mathrm{Var}(a_d) \overset{(3)}{=} \frac{\sigma_e^2}{N_d}, \qquad \sigma_e^2 = \frac{1}{A}\int W\,(e-1)^2\,\mathrm{d}\boldsymbol{x}. \tag{11.14}$$

(1) In the definition of $a_d = [\delta_d\,1]$, do the integral over the partner $\boldsymbol{y}$ first: it gives the area of $\boldsymbol{x}$'s shell inside the survey, which is $A\,G\,e(\boldsymbol{x})$ by the definitions of e and of G.

(2) Replace $W\delta_d$ by $(A/N_d)\,n_d - W$, from the definition of δ_d. The density n_d is a sum of spikes at the galaxies, so its integral against e is $\sum_i e(\boldsymbol{x}_i)$; the integral of We is A, since e averages to 1.

(3) Unclustered galaxies are N_d independent uniform points in the survey, so the $e(\boldsymbol{x}_i)$ are independent with mean 1 and variance σ_e^2; the variance of their mean is σ_e^2/N_d.

So a_d is simply the average edge weight of the galaxies we happened to get, minus its expected

value. Comparing the two contributions to the natural estimator,

$$\frac{\mathrm{Var}(2a_d)}{\mathrm{Var}([\delta_d\delta_d])} = \frac{4\sigma_e^2/N_d}{1/N_p} = 2\,\sigma_e^2\,N_dG. \tag{11.15}$$

The product N_dG is the average number of neighbours a galaxy has in the bin. For a square survey of side $700\,h^{-1}$Mpc and $r = 100\,h^{-1}$Mpc, a direct numerical average of the shell fraction over the square gives $\sigma_e^2 \approx 0.064$. A galaxy with 1000 neighbours in the bin then makes the edge term ≈ 130 times the pair noise in variance, about eleven times in standard deviation. LS removes that term and keeps only the $1/N_p$ that no pair-count estimator can avoid. In a periodic box $e = 1$ everywhere, so $\sigma_e^2 = 0$ and the edge term is absent from every estimator.

Example: a periodic box. In a periodic box there are no edges: every point has its whole shell inside the box, so the edge weighting is uniform and a_d equals the plain average of δ_d, which is zero. The same holds for a_r. The first-order terms of the natural and Davis–Peebles estimators then vanish, and all four estimators are equivalent to first order. Simulations of a periodic box should show four nearly identical scatters; any difference must come from the survey's edges.

Example: how big is the edge weight? Take a square survey of side $700\,h^{-1}$Mpc and $r = 100\,h^{-1}$Mpc. A point at the centre has its whole shell inside. A point in a corner has a quarter of it. A point on an edge, away from the corners, has exactly half: a straight edge through the centre of a circle cuts it into two equal halves. So the edge weight varies by a factor of four across the survey. Within $100\,h^{-1}$Mpc of the boundary lies a fraction $1 - (500/700)^2 \approx 0.49$ of the area: half of the survey's galaxies take part in pairs with reduced weight. If that outer half happens to be 1% underdense, which is typical of fluctuations on scales of hundreds of Mpc, then a_d is of order 10^{-3}, a sizeable fraction of ξ itself at $r = 100\,h^{-1}$Mpc, which is a few $\times 10^{-3}$.

Example: how many randoms? The random catalogue's own Poisson noise enters through δ_r. In the natural estimator it appears at first order, through $-2a_r$, whose variance falls only as $1/N_r$; so the natural estimator needs a random catalogue far larger than the data to be competitive. In Landy–Szalay δ_r enters only through the difference $\delta_d - \delta_r$, inside a product, and a random catalogue a few times larger than the data is enough. Eisenstein and collaborators used at least sixteen times as many randoms as galaxies [Eisenstein2005 §3.1]; our simulation uses five.

11.2.3 The integral constraint

Removing the first-order terms does not remove every bias. The trouble is the normalisation by the observed N_d. In a survey that happens to be overdense as a whole, dividing by the observed mean density makes the survey look average, and pairs everywhere look less correlated than they are. We can say how much less.

Write the true field as δ and its plain survey average as $\bar\delta_W = \frac{1}{A}\int W\delta\,\mathrm{d}\boldsymbol{x}$. Leaving out the Poisson noise, which does not correlate distinct points (Section 11.1.2), the normalised data fluctuation is $1 + \delta_d = (1+\delta)/(1+\bar\delta_W)$, so $\delta_d \approx \delta - \bar\delta_W$ to first order. Then

$$\begin{aligned}\mathbb{E}\big[\,[\delta_d\delta_d]\,\big] &\overset{(1)}{=} \mathbb{E}\big[\,[(\delta-\bar\delta_W)(\delta-\bar\delta_W)]\,\big] \overset{(2)}{=} \mathbb{E}\big[[\delta\delta]\big] - 2\,\mathbb{E}\big[[\delta\,1]\,\bar\delta_W\big] + \mathbb{E}\big[\bar\delta_W^2\big] \\ &\overset{(3)}{=} \xi(r) - 2\,\mathbb{E}\big[[\delta\,1]\,\bar\delta_W\big] + \bar\xi \overset{(4)}{=} \xi(r) - \bar\xi, \qquad \bar\xi \equiv \frac{1}{A^2}\int\!\!\int W(\boldsymbol{x})W(\boldsymbol{y})\,\xi(|\boldsymbol{x}-\boldsymbol{y}|)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y}.\end{aligned} \tag{11.16}$$

(1) $\delta_d \approx \delta - \bar{\delta}_W$.
(2) Multiply out the product; $\bar{\delta}_W$ is a number, the same at both ends of every pair, and $[1\,1] = 1$.
(3) The expectation of $\delta(\boldsymbol{x})\delta(\boldsymbol{y})$ is ξ at their separation, which is fixed within the bin; $\mathbb{E}[\bar{\delta}_W^2]$ is the double window average of ξ, called $\bar{\xi}$.
(4) $\mathbb{E}\big[[\delta\,1]\,\bar{\delta}_W\big]$ is the average of ξ between a pair member and a random point of the survey. Ignoring the edge weighting it equals $\bar{\xi}$ as well, and $-2\bar{\xi} + \bar{\xi} = -\bar{\xi}$.

All four estimators share this bias, because all of them normalise by the observed number of galaxies. It is called the **integral constraint**: the measured correlation function, integrated over the survey, is forced to be close to zero, because the survey's own mean density was taken as the truth. It shifts $\hat{\xi}$ down by the constant $\bar{\xi}$, the variance of the mean density of the survey, and it cannot be removed by any combination of pair counts, only modelled [Bernardeau2002 eq. (393) and p. 141]. At small separations, where $\xi \gg \bar{\xi}$, it is negligible; at the scale of the acoustic peak it is not. A commonly used form that also holds when ξ is not small is $\mathbb{E}[\hat{\xi}] \approx (1+\xi)/(1+\bar{\xi}) - 1$, which reduces to $\xi - \bar{\xi}$ to first order.

Watch out: The randoms must have the survey's selection, exactly

Every estimator derivation so far assumes that the randoms are a uniform sample of the *same* window as the data, including the fall-off of density with distance and every masked region. If the randoms are a little too dense in one part of the survey (an error in the photometric calibration of one patch of sky, say), then δ_d acquires a large-scale pattern that is not cosmology, and its pair average adds a spurious ξ at every separation. Eisenstein and collaborators found that a 1% shift in the $r - i$ colour changes the LRG number density by 8–10% [Eisenstein2005 §3.2]. Real analyses spend much of their effort building the random catalogue and its weights.

11.2.4 Testing the estimators in a survey with holes

The script `code/ch11/02_xi_estimators.py` makes the question concrete in two dimensions, where pair counts are cheap. The field is a lognormal field (Section 11.1.4) of unit variance on a grid of 256^2 cells in a box of side $1000\,h^{-1}$Mpc; its two-point function is known exactly on the grid, and it is 0.212 at $r = 30\,h^{-1}$Mpc and -0.0089 at $r = 100\,h^{-1}$Mpc. The survey window is a 700×700 square with five circular holes and a notch, of area $448 \times 10^3\ h^{-2}\mathrm{Mpc}^2$ (Figure 11.6). Each of 300 mock catalogues is a Poisson sample with $\bar{n} = 0.012\,h^2\mathrm{Mpc}^{-2}$ inside the window, about 5379 galaxies, and each mock gets its own random catalogue of 26896 points.

Listing 11.3: Pair counts and the four estimators (`code/ch11/02_xi_estimators.py`)

```
def counts(t1, t2, edges):
    c = t1.count_neighbors(t2, edges)     # ordered pairs with separation <= edge
    return np.diff(c).astype(float)


def randoms(rng):
    cand = rng.random((int(nr / (W.mean()) * 1.3), 2)) * L
    inw = W[(cand[:, 0] // D).astype(int), (cand[:, 1] // D).astype(int)] > 0
    t = cKDTree(cand[inw][:nr])
    return t, counts(t, t, EDGES) / (nr * (nr - 1))


tr_fix, rr_fix = randoms(rng)          # one catalogue reused for every mock (the
    comparison)


def estimators(dd, dr, rr):
```

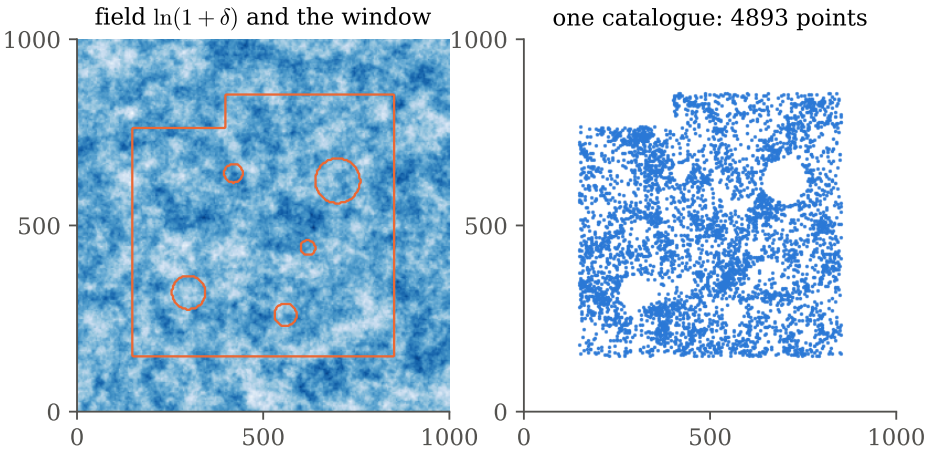


Figure 11.6: The test survey. Left: one lognormal field (shown as $\ln(1+\delta)$) with the survey boundary and holes in orange. Right: one Poisson catalogue drawn from a field inside that window.

```
80      return np.array([dd / rr - 1, dd / dr - 1, dd * rr / dr ** 2 - 1, (dd - 2 * dr + rr) /
        rr])
81
82
83  NAMES = ["natural $DD/RR-1$", "Davis--Peebles $DD/DR-1$", "Hamilton $DD\\,RR/DR^2-1$",
        "Landy--Szalay"]
84  t0 = time.time()
85  est, est_fix, ddpairs = [], [], []
86  for s in range(NSIM):
87      delta = ln.draw(rng)
88      c = poisson_sample(1 + delta, NBAR, L, rng, window=W)
89      g = points_from_counts(c, L, rng)
90      ng = len(g)
91      tg = cKDTree(g)
92      ddc = counts(tg, tg, EDGES)
93      dd = ddc / (ng * (ng - 1))
94      tr_r, rr = randoms(rng)              # a fresh random catalogue for this mock
95      dr = counts(tg, tr_r, EDGES) / (ng * nr)
96      est.append(estimators(dd, dr, rr))
97      dr_fix = counts(tg, tr_fix, EDGES) / (ng * nr)
98      est_fix.append(estimators(dd, dr_fix, rr_fix))
99      ddpairs.append(ddc / 2)
100     if s == 0:
101         g_example = g
```

`counts` uses `cKDTree.count_neighbors`, which returns, for each bin edge, the number of ordered pairs closer than that edge; differencing gives the pairs per bin. A tree visits only nearby branches, so the cost grows much more slowly than the N^2 of a double loop. Lines 73, 93 and 95 divide by the numbers of possible pairs, (11.11), and line 80 forms the four estimators. The same estimators are also computed in a periodic box with the same number of points and no edges, as the control of the first example above. Finally, the integral constraint $\bar{\xi}$ of (11.16) is evaluated exactly on the grid (lines 56–58 of the script) by a Fourier transform: the double window integral is the window's autocorrelation times ξ, summed. It is $\bar{\xi} = 0.0016$ for this window.

Figure 11.7 answers the question. On the right, in the survey with holes, the natural estimator's scatter at $r = 100\,h^{-1}$Mpc is 1.87 times that of Landy–Szalay, Davis–Peebles is 1.29 times, and Hamilton is 0.99 times: exactly the ordering of the first-order terms in (11.13), coefficient 2, 1, 0 and 0. At $r = 30\,h^{-1}$Mpc, where ξ is large and the first-order term is

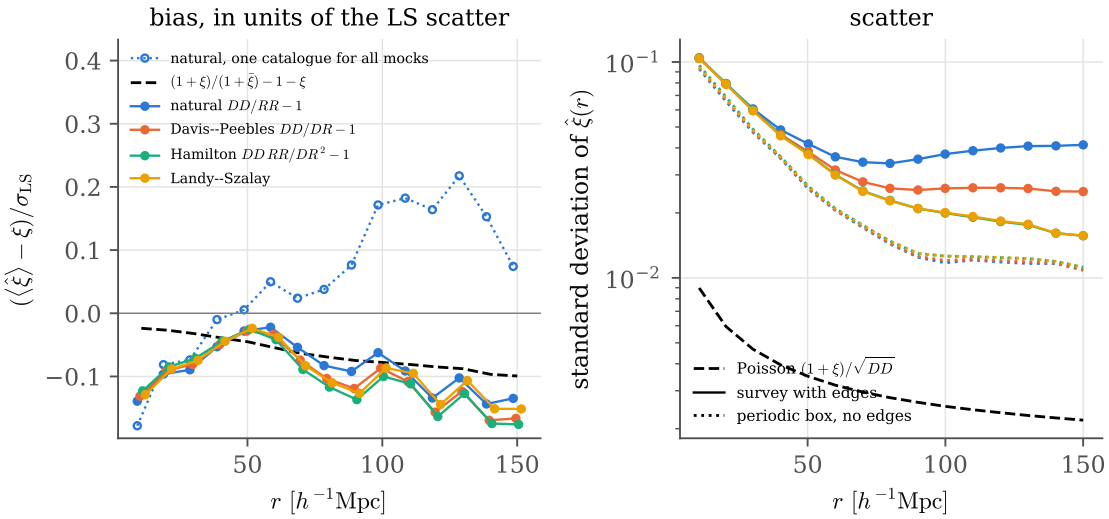


Figure 11.7: The four estimators over 300 mock surveys. Left: the mean minus the true $\xi(r)$, in units of the Landy–Szalay scatter; the dashed curve is the integral-constraint expectation $(1+\xi)/(1+\bar{\xi})-1-\xi$; the open circles are the natural estimator when one random catalogue is reused for every mock. Right: the standard deviation of each estimator in the survey with holes (solid) and in a periodic box with the same number of points (dotted), with the Poisson floor $(1+\xi)/\sqrt{DD}$ (dashed black).

relatively unimportant, the ratios shrink to 1.02 and 1.00. In the periodic control the four scatters nearly coincide (the natural-to-LS ratio is 0.93 at $100\,h^{-1}$Mpc): without edges there is nothing for Landy–Szalay to cancel. On the left, all four means follow the integral-constraint curve. At $r = 100\,h^{-1}$Mpc the prediction is -0.0105 (the true ξ is -0.0089); Landy–Szalay gives -0.0107 ± 0.0012, Hamilton -0.0109, Davis–Peebles -0.0107, and the natural estimator -0.0102 ± 0.0022. On average, then, none of them is biased beyond the integral constraint; they differ in how much they scatter.

That agreement needs one precaution, and it is easy to miss. Suppose we save time and make a single random catalogue for all 300 mocks. Landy–Szalay and Hamilton do not notice: their means become -0.0104 and -0.0106. But the natural estimator's mean moves, in this run to -0.0055, and Davis–Peebles' to -0.0082. How far they move, and in which direction, depends on which random catalogue happened to be drawn; another catalogue would shift them by a different amount. The reason is the first-order random term of the previous example. Each random catalogue has its own small lumps, and the natural estimator carries them straight into $\hat{\xi}$. With a fresh catalogue for every mock those lumps differ from mock to mock and average away. With one shared catalogue they are the same in every mock, so averaging 300 mocks cannot remove them, and the error bar computed from the scatter of the mocks does not even know they are there. Landy–Szalay cancels the random lumps by construction, which is one more reason to prefer it.

Notice also how far the scatter is above the Poisson floor: at $r = 100\,h^{-1}$Mpc the LS standard deviation is 0.0201 and the Poisson floor is 0.0025. Landy and Szalay's optimality is a statement about the noise their estimator *adds*. The scatter that remains is mostly the sample variance of the field: the survey holds only a few independent patches of size $100\,h^{-1}$Mpc, and the field in those patches is one draw. No estimator removes that; only a bigger survey does.

11.3 The power spectrum of a survey: the FKP estimator

A flux-limited survey is dense nearby and sparse far away. Close to us we see faint galaxies as well as bright ones, so the mean density $\bar{n}(\boldsymbol{x})$ is high; at the far edge only the most luminous galaxies make it into the catalogue, and $\bar{n}$ may be a hundred times lower. Both parts of the survey carry information about $P(k)$, but of different quality. The nearby part is limited by its small volume: it holds few independent Fourier modes, and its error is sample variance. The distant part has plenty of volume but few galaxies: its error is shot noise. If we Fourier transform the galaxy density with every galaxy counted equally, the crowded nearby region dominates the answer and the huge distant volume is wasted.

How much weight should a galaxy get, depending on where it sits, so that the estimated power spectrum has the smallest possible error? Feldman, Kaiser and Peacock answered it in 1994, and the answer, the FKP weight, is used in essentially every galaxy power-spectrum measurement since [223]. Its derivation needs only the Cox model of Section 11.1 and the Gaussian statistics of Fourier modes from Chapter 7.

11.3.1 The weighted field and its mean power

Write the galaxy catalogue as a sum of spikes, $n_g(\boldsymbol{x}) = \sum_i \delta_D(\boldsymbol{x} - \boldsymbol{x}_i)$, and do the same for a random catalogue with the same selection but $1/\alpha$ times as many points, $n_s(\boldsymbol{x})$. Here $\bar{n}(\boldsymbol{x})$ is the expected galaxy density at $\boldsymbol{x}$ in the absence of clustering, set by the survey's angular mask and its fall-off with distance, and $w(\boldsymbol{x})$ is a weight still to be chosen. FKP's weighted fluctuation field is

$$F(\boldsymbol{x}) = \frac{w(\boldsymbol{x})\,\big[n_g(\boldsymbol{x}) - \alpha\, n_s(\boldsymbol{x})\big]}{I^{1/2}}, \qquad I = \int \bar{n}^2(\boldsymbol{x})\, w^2(\boldsymbol{x})\, \mathrm{d}^3x \tag{11.17}$$

[FKP1994 eq. (2.1.3)]. The randoms subtract the mean density, so F is zero on average; the weight shapes how much each place counts; and I is chosen so that $\mathbb{E}\,|F_{\boldsymbol{k}}|^2$ comes out, in the calculation below, as $P(k)$ itself.

The two-point functions of the catalogues follow from Section 11.1.2, now with a mean density that varies with position. For galaxies, the pair term (11.5) becomes $\bar{n}(\boldsymbol{x})\bar{n}(\boldsymbol{y})[1+\xi(\boldsymbol{x}-\boldsymbol{y})]$, and the Poisson variance of a single cell, $\bar{n}\Delta V$, becomes the self-pair spike $\bar{n}(\boldsymbol{x})\delta_D(\boldsymbol{x}-\boldsymbol{y})$ in the continuum. Randoms are an unclustered Cox sample with density $\bar{n}/\alpha$, independent of the galaxies:

$$\begin{aligned} \mathbb{E}[n_g(\boldsymbol{x})n_g(\boldsymbol{y})] &= \bar{n}(\boldsymbol{x})\bar{n}(\boldsymbol{y})\big[1+\xi(\boldsymbol{x}-\boldsymbol{y})\big] + \bar{n}(\boldsymbol{x})\,\delta_D(\boldsymbol{x}-\boldsymbol{y}), \\ \mathbb{E}[n_s(\boldsymbol{x})n_s(\boldsymbol{y})] &= \alpha^{-2}\bar{n}(\boldsymbol{x})\bar{n}(\boldsymbol{y}) + \alpha^{-1}\bar{n}(\boldsymbol{x})\,\delta_D(\boldsymbol{x}-\boldsymbol{y}), \qquad \mathbb{E}[n_g(\boldsymbol{x})n_s(\boldsymbol{y})] = \alpha^{-1}\bar{n}(\boldsymbol{x})\bar{n}(\boldsymbol{y}) \end{aligned} \tag{11.18}$$

[FKP1994 eq. (2.1.5)]. With the Fourier convention $F_{\boldsymbol{k}} = \int F(\boldsymbol{x})e^{-i\boldsymbol{k}\cdot\boldsymbol{x}}\mathrm{d}^3x$, so that an infinite field has $\mathbb{E}[\delta_{\boldsymbol{k}}\delta^*_{\boldsymbol{k}'}] = (2\pi)^3\delta_D(\boldsymbol{k}-\boldsymbol{k}')P(k)$,

$$\begin{aligned} \mathbb{E}\,|F_{\boldsymbol{k}}|^2 &\overset{(1)}{=} \frac{1}{I}\int\!\!\int w(\boldsymbol{x})w(\boldsymbol{y})\;\mathbb{E}\big[(n_g-\alpha n_s)_{\boldsymbol{x}}(n_g-\alpha n_s)_{\boldsymbol{y}}\big]e^{-i\boldsymbol{k}\cdot(\boldsymbol{x}-\boldsymbol{y})}\mathrm{d}^3x\,\mathrm{d}^3y \\ &\overset{(2)}{=} \frac{1}{I}\int\!\!\int w(\boldsymbol{x})w(\boldsymbol{y})\Big[\bar{n}(\boldsymbol{x})\bar{n}(\boldsymbol{y})\,\xi(\boldsymbol{x}-\boldsymbol{y}) + (1+\alpha)\,\bar{n}(\boldsymbol{x})\,\delta_D(\boldsymbol{x}-\boldsymbol{y})\Big]e^{-i\boldsymbol{k}\cdot(\boldsymbol{x}-\boldsymbol{y})}\mathrm{d}^3x\,\mathrm{d}^3y \\ &\overset{(3)}{=} \int \frac{\mathrm{d}^3k'}{(2\pi)^3}\,P(k')\,\big|\tilde{G}(\boldsymbol{k}-\boldsymbol{k}')\big|^2 + P_{\rm shot}, \\ \tilde{G}(\boldsymbol{q}) &\equiv \frac{1}{I^{1/2}}\int \bar{n}(\boldsymbol{x})w(\boldsymbol{x})\,e^{-i\boldsymbol{q}\cdot\boldsymbol{x}}\,\mathrm{d}^3x, \qquad P_{\rm shot} \equiv \frac{(1+\alpha)}{I}\int \bar{n}(\boldsymbol{x})\,w^2(\boldsymbol{x})\,\mathrm{d}^3x. \end{aligned} \tag{11.19}$$

(1) Square the Fourier transform of (11.17) and take the expectation inside the integrals.

(2) Expand the bracket with (11.18): the four $\bar{n}\bar{n}$ terms are $1-1-1+1=0$, leaving the clustering term $\bar{n}\bar{n}\,\xi$; the spikes come in with coefficients 1 (galaxies) and $\alpha^2\cdot\alpha^{-1}=\alpha$ (randoms).

(3) In the first term write $\xi(\boldsymbol{x}-\boldsymbol{y})=\int\frac{\mathrm{d}^3k'}{(2\pi)^3}P(k')e^{i\boldsymbol{k}'\cdot(\boldsymbol{x}-\boldsymbol{y})}$; the $\boldsymbol{x}$ and $\boldsymbol{y}$ integrals then factorise into $\tilde{G}(\boldsymbol{k}-\boldsymbol{k}')$ and its complex conjugate. In the second term the spike sets $\boldsymbol{y}=\boldsymbol{x}$ and the exponential becomes 1.

The first term is the true spectrum blurred by the window. The function $\left|\tilde{G}\right|^2$ is the power spectrum of the weighted survey shape $\bar{n}w$; for a survey of size D it is a blob of width about $1/D$ around $\boldsymbol{q}=0$, and by Parseval's theorem its integral $\int\frac{\mathrm{d}^3q}{(2\pi)^3}\left|\tilde{G}\right|^2=\frac{1}{I}\int\bar{n}^2w^2\,\mathrm{d}^3x=1$. So on scales much smaller than the survey, $k\gg1/D$, where P hardly changes across the blob, (11.19) becomes

$$\mathbb{E}\,|F_{\boldsymbol{k}}|^2\approx P(k)+P_{\rm shot},\qquad \hat{P}(k)=\frac{1}{V_k}\int_{\rm shell}\left(|F_{\boldsymbol{k}'}|^2-P_{\rm shot}\right)\mathrm{d}^3k' \tag{11.20}$$

[FKP1994 eqs. (2.1.8)–(2.1.11)], where V_k is the volume of the shell of wavevectors averaged over. This is the **FKP estimator**: Fourier transform the weighted field, square, subtract the known shot noise, average over a shell.

Two corrections are hidden in the "$\approx$". On scales comparable to the survey, $k\sim1/D$, the window convolution in (11.19) is not negligible: it smooths out features and suppresses power at the lowest k. It is the same mode coupling met for masked skies in Section 8.10, and it is handled the same way, by convolving the *model* with the window before comparing it with the data. And the shot noise subtracted is the Poisson value; Section 11.1 listed why real galaxies need not be exactly Poisson.

Example: a uniform survey. For constant $\bar{n}$ and w inside a volume V, $I=\bar{n}^2w^2V$ and $P_{\rm shot}=(1+\alpha)\bar{n}w^2V/(\bar{n}^2w^2V)=(1+\alpha)/\bar{n}$. With a random catalogue much larger than the data, $\alpha\to0$, and we recover the $1/\bar{n}$ of (11.8). A finite random catalogue adds its own shot noise, $\alpha/\bar{n}$: with twenty times as many randoms as galaxies that is a 5% addition, which is why the randoms are made large.

11.3.2 The error, and the weight that minimises it

To choose w we need the variance of $\hat{P}(k)$. Assume the Fourier modes $F_{\boldsymbol{k}}$ are Gaussian; for the long-wavelength modes of a survey this is a good approximation, because each $F_{\boldsymbol{k}}$ is a sum over a huge number of galaxies (the central limit theorem of Section 3.3.4). Write $C(\boldsymbol{k},\boldsymbol{k}')=\mathbb{E}[F_{\boldsymbol{k}}F^*_{\boldsymbol{k}'}]$. Repeat the first two steps of (11.19), but keep the two wavevectors different, so that the phases are $e^{-i\boldsymbol{k}\cdot\boldsymbol{x}+i\boldsymbol{k}'\cdot\boldsymbol{y}}$. The spike term then gives, after the δ_D sets $\boldsymbol{y}=\boldsymbol{x}$, the number $\tilde{S}(\boldsymbol{q})=\frac{1+\alpha}{I}\int\bar{n}w^2e^{-i\boldsymbol{q}\cdot\boldsymbol{x}}\mathrm{d}^3x$ at $\boldsymbol{q}=\boldsymbol{k}-\boldsymbol{k}'$. The clustering term, with ξ written as the integral of P as before, is $\int\frac{\mathrm{d}^3k''}{(2\pi)^3}P(k')\,\tilde{G}(\boldsymbol{k}-\boldsymbol{k}'')\,\tilde{G}^*(\boldsymbol{k}'-\boldsymbol{k}'')$. Both $\tilde{G}$ factors are blobs of width $\sim1/D$, so only $\boldsymbol{k}''$ within $1/D$ of $\boldsymbol{k}$ contributes, and there P is nearly the constant $P(k)$. Pulling it out leaves the convolution of $\tilde{G}$ with its own conjugate, which is the Fourier transform of $|G(\boldsymbol{x})|^2=\bar{n}^2w^2/I$: that is $\tilde{Q}(\boldsymbol{q})$. So $C(\boldsymbol{k},\boldsymbol{k}')\approx P(k)\,\tilde{Q}(\boldsymbol{k}-\boldsymbol{k}')+\tilde{S}(\boldsymbol{k}-\boldsymbol{k}')$ [FKP1994 eqs. (2.2.2)–(2.2.5)], with $\tilde{Q}$ the Fourier transform of $\bar{n}^2w^2/I$. Both $\tilde{Q}$ and $\tilde{S}$ are blobs of width $\sim1/D$ around zero, and the covariance of two modes dies away once they are more than $1/D$ apart. Then

$$\begin{aligned}\operatorname{Var}\hat{P}(k)&\overset{(1)}{=}\frac{1}{V_k^2}\int_{\rm shell}\int_{\rm shell}\Big[|C(\boldsymbol{k},\boldsymbol{k}')|^2+|C(\boldsymbol{k},-\boldsymbol{k}')|^2\Big]\mathrm{d}^3k\,\mathrm{d}^3k'\\&\overset{(2)}{=}\frac{2}{V_k}\int\left|P\,\tilde{Q}(\boldsymbol{q})+\tilde{S}(\boldsymbol{q})\right|^2\mathrm{d}^3q\overset{(3)}{=}\frac{2(2\pi)^3}{V_k\,I^2}\int\left[\bar{n}^2w^2P+(1+\alpha)\bar{n}w^2\right]^2\mathrm{d}^3x,\end{aligned}$$

$$\frac{\operatorname{Var}\hat{P}(k)}{P^2(k)} \overset{(4)}{=} \frac{2(2\pi)^3}{V_k} \frac{\int \bar{n}^4 w^4 \left[1 + 1/(\bar{n}P)\right]^2 \mathrm{d}^3x}{\left[\int \bar{n}^2 w^2 \,\mathrm{d}^3x\right]^2}. \tag{11.21}$$

(1) The covariance of $|F_{\boldsymbol{k}}|^2$ and $|F_{\boldsymbol{k}'}|^2$ for complex Gaussian modes, by Isserlis' theorem (Section 7.2.1). The second term is there because the field is real, $F_{-\boldsymbol{k}'} = F^*_{\boldsymbol{k}'}$, so a mode near $-\boldsymbol{k}$ is the same mode as one near $\boldsymbol{k}$.
(2) C depends only on $\boldsymbol{q} = \boldsymbol{k} - \boldsymbol{k}'$ and is narrow compared with the shell, so for each $\boldsymbol{k}$ the $\boldsymbol{k}'$ integral covers the whole blob; the remaining $\boldsymbol{k}$ integral gives V_k. The second term gives the same.
(3) Parseval: $\int |\tilde{g}(\boldsymbol{q})|^2 \,\mathrm{d}^3q = (2\pi)^3 \int |g(\boldsymbol{x})|^2 \,\mathrm{d}^3x$, with $g = [\bar{n}^2 w^2 P + (1+\alpha)\bar{n}w^2]/I$.
(4) Divide by P^2, take $\alpha \to 0$, and insert $I = \int \bar{n}^2 w^2$.

This is FKP's eq. (2.3.2), up to the factor 2.[1]

Now the question becomes a small optimisation problem. Write $u = \bar{n}^2 w^2 \geq 0$ and $c = 1 + 1/(\bar{n}P)$. The fractional variance is proportional to $R[u] = \int u^2 c^2 / \left(\int u\right)^2$, and we want the u that makes it smallest. The Cauchy–Schwarz inequality does it in one line:

$$\left(\int u \,\mathrm{d}^3x\right)^2 \overset{(1)}{=} \left(\int (u\,c)\,\frac{1}{c}\,\mathrm{d}^3x\right)^2 \overset{(2)}{\leq} \int u^2c^2\,\mathrm{d}^3x \int \frac{1}{c^2}\,\mathrm{d}^3x,$$
$$R[u] \overset{(3)}{\geq} \frac{1}{\int c^{-2}\,\mathrm{d}^3x} \overset{(4)}{=} \frac{1}{\int \left[\frac{\bar{n}P}{1+\bar{n}P}\right]^2 \mathrm{d}^3x}. \tag{11.22}$$

(1) Multiply and divide by $c > 0$.
(2) Cauchy–Schwarz for integrals, $\left(\int fg\right)^2 \leq \int f^2 \int g^2$, with $f = uc$ and $g = 1/c$. Equality holds exactly when $f \propto g$.
(3) Divide both sides by $(\int u)^2 \int c^{-2}$.
(4) $1/c = \bar{n}P/(1+\bar{n}P)$.

The bound is reached when $uc \propto 1/c$, that is $\bar{n}^2 w^2 \propto [\bar{n}P/(1+\bar{n}P)]^2$, or

$$w_{\mathrm{FKP}}(\boldsymbol{x}) = \frac{1}{1 + \bar{n}(\boldsymbol{x})\,P(k)} \tag{11.23}$$

[FKP1994 eq. (2.3.5)], up to an irrelevant constant. At the minimum, (11.21) becomes

$$\frac{\operatorname{Var}\hat{P}(k)}{P^2(k)} = \frac{2(2\pi)^3}{V_k\,V_{\mathrm{eff}}(k)}, \qquad V_{\mathrm{eff}}(k) \equiv \int \left[\frac{\bar{n}(\boldsymbol{x})P(k)}{1+\bar{n}(\boldsymbol{x})P(k)}\right]^2 \mathrm{d}^3x. \tag{11.24}$$

Working definition 11.3 (effective volume)

The **effective volume** $V_{\mathrm{eff}}(k)$ of a survey is the volume of a noise-free survey that would measure $P(k)$ as well [with the same number of independent Fourier modes] as the real one does with optimal weights. Every volume element counts with the factor $[\bar{n}P/(1+\bar{n}P)]^2$: fully where the signal-to-noise per mode $\bar{n}P \gg 1$, hardly at all where $\bar{n}P \ll 1$. *Operationally:* integrate $[\bar{n}P/(1+\bar{n}P)]^2$ over the survey; the fractional error of P in a shell of volume V_k is then $\sqrt{2(2\pi)^3/(V_k V_{\mathrm{eff}})}$.

Read (11.24) next to the sample variance of a periodic box (Section 7.5.2): there a shell holds $N_{\mathrm{modes}} = VV_k/(2\pi)^3$ modes, half of them independent, and $\operatorname{Var}\hat{P}/P^2 = 2/N_{\mathrm{modes}}$. The survey

[1] FKP keep only the neighbourhood $\boldsymbol{k}' \approx \boldsymbol{k}$ in their eq. (2.3.1). For a real field the neighbourhood $\boldsymbol{k}' \approx -\boldsymbol{k}$ is the same set of modes counted again, which doubles the variance, as in the mode counting of [SeoEisenstein2003 eq. (8)]. The factor does not change the optimal weight.

formula is the same with V replaced by V_{eff}. And read the weight (11.23) in its two limits. Where galaxies are plentiful, $\bar{n}P \gg 1$, the weight is $1/(\bar{n}P) \propto 1/\bar{n}$: each galaxy is down-weighted so that each *volume* counts equally, because the error there is sample variance, which depends on volume, not on galaxies. Where galaxies are scarce, $\bar{n}P \ll 1$, the weight is 1: each *galaxy* counts equally, because the error there is shot noise. The weight on a region's contribution to the power, $\bar{n}^2 w^2 \propto [P/(P + 1/\bar{n})]^2$, is the inverse of that region's fractional variance per mode: it is inverse-variance weighting, the same rule that gave the best combination of measurements in Section 3.7.

Example: a uniform survey. With constant $\bar{n}$ the weight is a constant, so weighting changes nothing, and $V_{\text{eff}} = V[\bar{n}P/(1 + \bar{n}P)]^2$. At $\bar{n}P = 1$ the survey is worth a quarter of its volume; at $\bar{n}P = 3$, 56%; at $\bar{n}P = 10$, 83%. Beyond $\bar{n}P \approx 3$ more galaxies buy little: the error is already close to the sample-variance floor of the volume.

Example: Eisenstein's survey, rebuilt. The SDSS LRG sample has $\bar{n} \approx 10^{-4}\, h^3\text{Mpc}^{-3}$ out to $z = 0.36$, falling beyond, with $\bar{n}P_w \approx 4$ at $z < 0.36$ and ≈ 1 at $z = 0.47$ for $P_w = 4 \times 10^4\, h^{-3}\text{Mpc}^3$, over $3816\,\text{deg}^2$ and $0.16 < z < 0.47$ [Eisenstein2005 §2, §3.1]. Those numbers fix $\bar{n}(z)$: 10^{-4} up to $z = 0.36$, then falling to 2.5×10^{-5} at $z = 0.47$ (we interpolate in $\ln \bar{n}$). The script `code/ch11/05_bao_fisher.py` builds this survey and evaluates (11.24). It gives a total volume of $0.72\, h^{-3}\text{Gpc}^3$ (paper: 0.72) and effective volumes of 0.13, 0.38 and $0.54\, h^{-3}\text{Gpc}^3$ for $P = 10^4$, 4×10^4 and $10^5\, h^{-3}\text{Mpc}^3$ (paper: 0.13, 0.38, 0.55). The paper's effective volumes follow from its stated densities through (11.24). On the scales of the acoustic peak, $P \approx 10^4$, the survey is worth less than a fifth of its volume.

Example: one weight for all scales. The optimal weight depends on k through $P(k)$. In practice one picks a single P_w for all scales, so that the weighting and hence the effective bias of the sample do not change with k. Eisenstein and collaborators chose $P_w = 4 \times 10^4\, h^{-3}\text{Mpc}^3$, optimal at $k \approx 0.05\, h\,\text{Mpc}^{-1}$ and within 5% of the optimal error at every $k \lesssim 0.15\, h\,\text{Mpc}^{-1}$ [Eisenstein2005 §3.1]. The minimum in (11.22) is flat: a weight that is only roughly right costs little.

11.3.3 Checking the FKP weight on a simulated survey

The script `code/ch11/03_fkp.py` makes a survey that is dense in the middle and sparse at the edge: a sphere of radius $450\, h^{-1}\text{Mpc}$ around an observer at the centre of a $1000\, h^{-1}\text{Mpc}$ box, with $\bar{n}(r) = 0.003\, e^{-(r/200)^2}\, h^3\text{Mpc}^{-3}$. It holds about 131×10^3 galaxies in a geometric volume of $0.38\, h^{-3}\text{Gpc}^3$ (Figure 11.8). The galaxies are Cox samples of 100 lognormal fields with spectrum $b^2 P_{\text{lin}}$, $b = 1.5$. Each catalogue is analysed twice, once with $w = 1$ and once with the FKP weight for $P_0 = 10000\, h^{-3}\text{Mpc}^3$.

Listing 11.4: The FKP estimator, its expectation and its predicted variance (`code/ch11/03_fkp.py`)

```
def fkp_setup(w):
    norm = np.sum(nbar ** 2 * w ** 2) * dV
    pshot = np.sum(nbar * w ** 2) * dV / norm              # FKP eq. (2.1.9) with
    infinitely many randoms
    # exact expectation: FT of xi(r) times the window autocorrelation Q(r) of n_bar w
    g = np.fft.rfftn(nbar * w)
    Q = np.fft.irfftn(np.abs(g) ** 2, s=nbar.shape) * dV / norm      # Q(0) = 1
    pconv = kb.avg(np.fft.rfftn(Q * xi_grid).real * dV)
    # FKP variance (their 2.3.2) evaluated at each shell's P: (2pi)^3/V_k * int n^4 w^4
    (1+1/nP)^2 / norm^2,
```

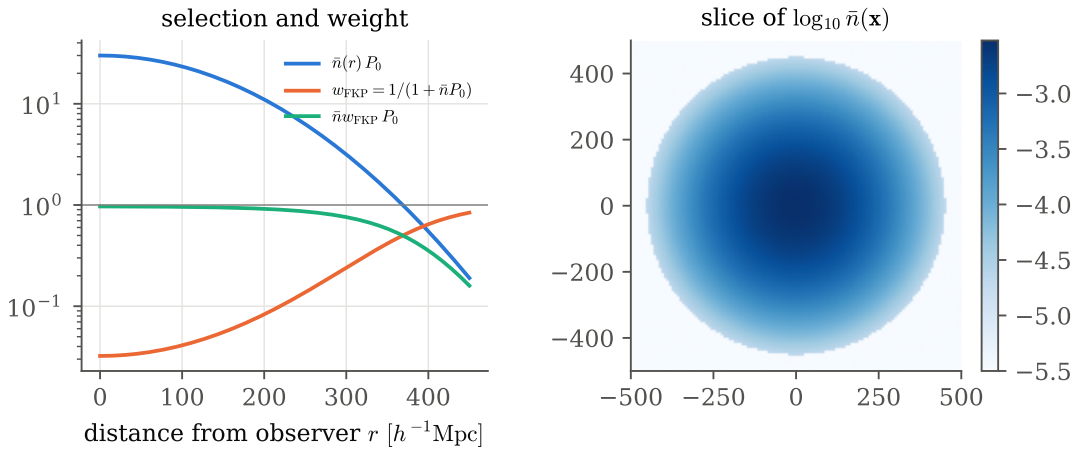


Figure 11.8: The simulated survey. Left: against distance from the observer, the signal-to-noise per mode $\bar{n}P_0$ (blue), the FKP weight (orange) and the weighted density $\bar{n}w\,P_0$ (green). Near the observer $\bar{n}P_0 \approx 30$ and the weight suppresses the crowded centre; beyond $r \approx 370\,h^{-1}$Mpc, where $\bar{n}P_0 < 1$, the weight is close to one. Right: a slice through the box showing $\log_{10}\bar{n}$.

```
56      # times 2 because the shell holds every independent mode twice (k and -k carry the same
        |F_k|^2)
57      vk = 4 * np.pi * k ** 2 * np.diff(edges)
58      sel = nbar > 0
59      frac = np.array([np.sum(nbar[sel] ** 4 * w[sel] ** 4 * (1 + 1 / (nbar[sel] * p)) ** 2)
        * dV / norm ** 2
60                       for p in ptrue]) * 2 * (2 * np.pi) ** 3 / vk
61      return norm, pshot, pconv, frac
62
63
64  setups = {name: fkp_setup(w) for name, w in weights.items()}
65  est = {name: [] for name in weights}
66  for s in range(NSIM):
67      delta = ln.draw(rng)
68      ng = poisson_sample(1 + delta, 1.0, L, rng, window=nbar) / dV      # galaxies per unit
        volume
69      for name, w in weights.items():
70          norm, pshot, _, _ = setups[name]
71          F = w * (ng - nbar) / np.sqrt(norm)
72          pk = np.abs(np.fft.rfftn(F) * dV) ** 2
73          est[name].append(kb.avg(pk) - pshot)
```

`fkp_setup` computes, for one weight, the normalisation I (line 49), the shot noise $P_{\rm shot}$ with $\alpha \to 0$ (line 50), and the exact expectation (11.19) (lines 52–54). The last uses the real-space form of the convolution: the window term is the Fourier transform of $\xi(\boldsymbol{r})$ times the autocorrelation of $\bar{n}w$, which is cheaper on a grid than the convolution itself. Lines 57–60 evaluate the predicted variance (11.21) for every shell. The main loop (lines 66–73) is (11.17) and (11.20) in four lines; the random catalogue is replaced by the exact $\bar{n}$, which is the limit $\alpha \to 0$.

Figure 11.9 shows three things. First, the estimator is unbiased once the window is accounted for: with the FKP weight the mean at $k = 0.10\,h\,\mathrm{Mpc}^{-1}$ is 1.001 times the window-convolved expectation, and at the lowest $k = 0.02$ the window itself pulls the expectation down to 0.98 of the true P. Second, the FKP weight about halves the error: the ratio of the FKP scatter to the $w = 1$ scatter is 0.46 at $k = 0.04$ and 0.54 at $k = 0.10\,h\,\mathrm{Mpc}^{-1}$, against 0.46 and 0.51 predicted by (11.21). The gain comes despite a shot-noise term that is five times larger with the FKP weight ($P_{\rm shot} = 4794$ against $926\,h^{-3}\mathrm{Mpc}^3$), because the weight spends shot noise to buy volume. Third, the absolute scatter is larger than the Gaussian prediction, by a factor 1.13 at $k = 0.04$ and

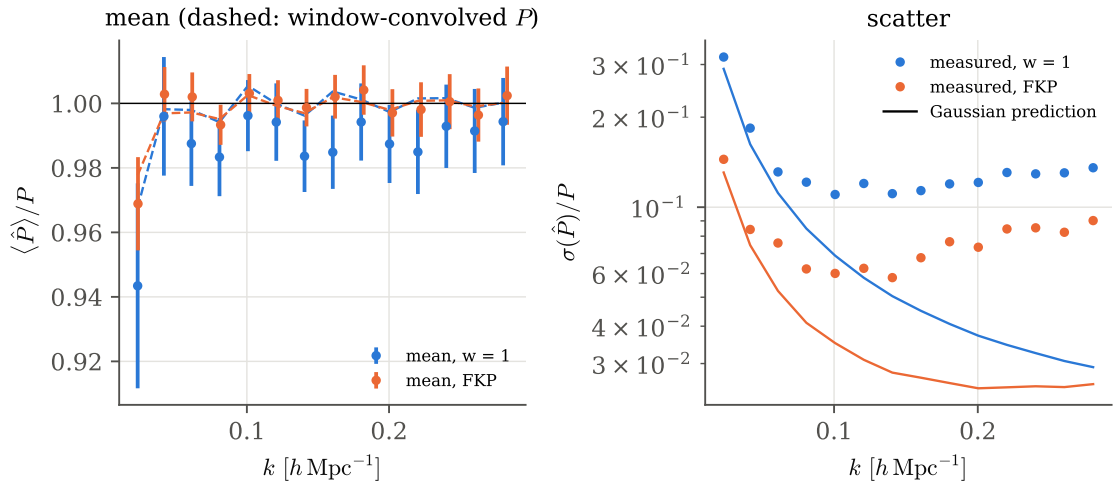


Figure 11.9: The FKP estimator over 100 simulated surveys. Left: the mean of $\hat{P}$ divided by the true spectrum, for $w = 1$ (blue) and the FKP weight (orange); the dashed curves are the window-convolved expectation (11.19). Right: the scatter of $\hat{P}/P$ (points) against the Gaussian prediction (11.21) (lines).

1.60 at $k = 0.10\,h\,\mathrm{Mpc}^{-1}$ for $w = 1$. The Gaussian assumption of (11.21) ignores the four-point function of the field, and a lognormal field at $z = 0$ has a large one on these scales. The excess affects both weightings alike, so the *ratio* still follows the Gaussian prediction. For the absolute error, real analyses do not trust (11.21); they measure the covariance from mock catalogues (Section 11.6). For the effective volume of this survey, (11.24) gives 0.23 $h^{-3}\mathrm{Gpc}^3$ at $k = 0.04$ and 0.16 at $k = 0.10\,h\,\mathrm{Mpc}^{-1}$, out of 0.38.

11.4 The acoustic peak: a ruler in the galaxy distribution

Before the Universe was 400,000 years old, photons and baryons formed a single fluid, and every overdensity sent out a spherical sound wave through it. When the baryons were released from the photons, at the drag epoch $z_d \approx 1060$, the waves stopped where they were. Each initial lump of matter is therefore surrounded by a faint shell of extra baryons at a fixed comoving radius, the **sound horizon at the drag epoch**

$$r_d = \int_{z_d}^{\infty} \frac{c_s(z)\,\mathrm{d}z}{H(z)}, \qquad c_s = \frac{c}{\sqrt{3(1+R)}}, \qquad R = \frac{3\rho_b}{4\rho_\gamma} \tag{11.25}$$

[DESI2024VI eqs. (2.3)–(2.4)]; this is the sound horizon $r_s(\eta)$ of Section T1.4 written as an integral over redshift (see the cosmology notes). For the Planck fiducial model CAMB gives $r_d = 147.10\,\mathrm{Mpc}$. Galaxies form where matter is dense, so galaxy pairs separated by r_d are slightly more common than their neighbours: the bump in $\xi(r)$ near 100 $h^{-1}\mathrm{Mpc}$ that we met in Figure T1.5 and that [42] first detected.

What makes the bump valuable is that its length is known. The CMB measures the physical densities that enter (11.25) so well that r_d is fixed to a fraction of a percent. A galaxy survey that locates the bump therefore measures how far away the galaxies are, the way a known ruler seen at an unknown distance measures the distance through its apparent size. Turning that idea into a measurement raises three statistical questions. What exactly does a shifted bump tell us? How significant is a bump in a noisy correlation function? And how precisely can a survey of given size locate it?

11.4.1 What a misplaced bump means: the dilation α

The event. We turned angles and redshifts into comoving positions using a fiducial cosmology, computed $\xi(r)$, and found the bump not at $r_d^{\rm fid}$, the fiducial model's sound horizon, but a few percent away.

What could have put it there? The ruler's true length is r_d, which may differ from the fiducial value. And the distances we used to convert the catalogue may be wrong. The second effect has two parts, because angles and redshifts are converted with different distances. Two galaxies at redshift z separated by a small angle $\Delta\theta$ are a transverse distance $r_\perp = D_M(z)\,\Delta\theta$ apart, where D_M is the comoving angular diameter distance; two galaxies on the same line of sight separated by Δz are a radial distance $r_\parallel = c\,\Delta z/H(z) \equiv D_H(z)\,\Delta z$ apart. (The definitions of D_M and $H(z)$ are in Section T1.2.) We computed both with the fiducial model, so

$$r_\perp^{\rm fid} \overset{(1)}{=} D_M^{\rm fid}\,\Delta\theta \overset{(2)}{=} \frac{D_M^{\rm fid}}{D_M}\,r_\perp \equiv \frac{r_\perp}{\alpha_\perp}, \qquad r_\parallel^{\rm fid} \overset{(3)}{=} \frac{D_H^{\rm fid}}{D_H}\,r_\parallel \equiv \frac{r_\parallel}{\alpha_\parallel}. \tag{11.26}$$

(1) The observed angle converted with the fiducial distance.
(2) The true separation is $r_\perp = D_M\Delta\theta$, so $\Delta\theta = r_\perp/D_M$.
(3) The same for the radial direction, with $D_H = c/H$ in place of D_M.

A true shell of radius r_d therefore appears in the fiducial coordinates with transverse radius $r_d/\alpha_\perp$ and radial radius $r_d/\alpha_\parallel$. Measuring both positions relative to the fiducial ruler gives the two numbers that modern surveys report: $\alpha_\perp = (D_M/r_d)/(D_M/r_d)^{\rm fid}$ and $\alpha_\parallel = (D_H/r_d)/(D_H/r_d)^{\rm fid}$, once the ruler's own length is included [DESI2024VI §2.3]. Only the ratios D_M/r_d and D_H/r_d are measured. A survey cannot tell a longer ruler from a smaller distance.

The angle average. The spherically averaged $\xi(r)$ of a sample with few galaxies cannot separate the two directions, and sees one overall stretch. A small volume element is stretched by $\alpha_\perp$ in two directions and $\alpha_\parallel$ in one, so its volume changes by $\alpha_\perp^2\alpha_\parallel$, and the isotropic stretch with the same effect on volume is the cube root,

$$\alpha = \left(\alpha_\perp^2\alpha_\parallel\right)^{1/3} = \frac{D_V/r_d}{(D_V/r_d)^{\rm fid}}, \qquad D_V(z) \equiv \left[z\,D_M^2(z)\,D_H(z)\right]^{1/3} \tag{11.27}$$

[Eisenstein2005 eq. (2)], [Anderson2012 eq. (21)]; the factor z is a convention that gives D_V the units and low-redshift limit of a distance. Finally, what does a dilation do to the measured correlation function? A pair with true separation s is placed at $r = s/\alpha$ in the fiducial coordinates, so

$$\xi_{\rm obs}(r) = \xi_{\rm true}(\alpha r), \qquad P_{\rm obs}(k) = \alpha^{-3}P_{\rm true}(k/\alpha). \tag{11.28}$$

The second form follows from the first by the substitution $\boldsymbol{s} = \alpha\boldsymbol{r}$ in the Fourier integral; the α^{-3} is the Jacobian.

Key idea: The acoustic peak measures a ratio of distances

A galaxy survey that finds the acoustic peak at a fraction $1/\alpha$ of its fiducial position measures $D_V(z)/r_d = \alpha\,(D_V/r_d)^{\rm fid}$ at the survey's redshift. Splitting the clustering by direction measures D_M/r_d and D_H/r_d separately. The ruler's length r_d comes from the early Universe; the distances come from the late Universe; BAO compares the two.

The Alcock–Paczyński idea. Even without knowing r_d, the *shape* of the clustering carries information. Structures that are statistically round in the true coordinates look squashed or stretched along the line of sight if $\alpha_\parallel \neq \alpha_\perp$, that is, if the fiducial cosmology got the product D_MH wrong. Testing for that apparent anisotropy is the Alcock–Paczyński test [DESI2024VI §2.3].

Its difficulty is that galaxy motions also squash the clustering along the line of sight (Section 11.5), so the two effects must be modelled together; for the acoustic peak, a sharp feature at a known scale, they separate cleanly. Figure 11.10 shows the three cases.

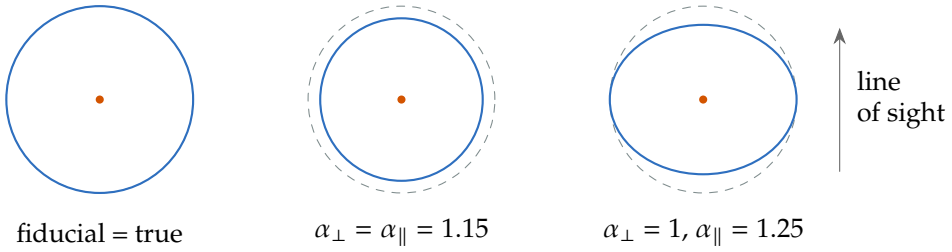


Figure 11.10: The acoustic shell around a galaxy, drawn in the fiducial coordinates. The dashed circle is where it would appear if the fiducial distances were right. Middle: true distances 15% larger in every direction make the shell look 15% smaller, an isotropic dilation. Right: a radial distance 25% too small in the fiducial model, with the transverse one right, squashes the shell along the line of sight; that anisotropy is the Alcock–Paczyński signal.

11.4.2 A model for the peak, and a fit

To locate the bump we fit a template whose bump position is controlled by α. Two physical effects decide what the template looks like.

The bump is smeared. Since the drag epoch, galaxies have moved by several h^{-1}Mpc, pushed by large-scale flows. Model each pair separation as shifted by a Gaussian random displacement with variance $\Sigma_{\rm nl}^2$ per component. Then the observed ξ is the true ξ convolved with a Gaussian, and by the convolution theorem (Section 7.6.1) the power spectrum is multiplied by the Fourier transform of that Gaussian, $e^{-k^2\Sigma_{\rm nl}^2/2}$. Only the oscillating part of $P(k)$ needs this damping, because the smooth part is handled by the free broad-band terms below. Writing $P_{\rm nw}$ for a smooth "no-wiggle" spectrum with the same broad shape [109], the template is

$$P_t(k) = \left[P_{\rm lin}(k) - P_{\rm nw}(k)\right] e^{-k^2\Sigma_{\rm nl}^2/2} + P_{\rm nw}(k) \tag{11.29}$$

[Anderson2012 eq. (26)], with $\Sigma_{\rm nl} = 8\,h^{-1}$Mpc for galaxies as observed and about $4\,h^{-1}$Mpc after "reconstruction", a step that estimates the flows from the galaxy field and moves the galaxies back [DESI2024VI §2.3]. Its correlation function $\xi_t(r)$ follows from the Fourier integral of Section T1.4.

The broad band is uncertain. Galaxy bias, non-linear growth and the integral constraint all change the smooth shape of ξ by amounts we do not want to model. So the fit includes an amplitude and a few smooth nuisance terms:

$$\xi_{\rm fit}(r) = B^2\,\xi_t(\alpha r) + a_0 + \frac{a_1}{r} + \frac{a_2}{r^2} \tag{11.30}$$

[Anderson2012 eqs. (23), (25)]. For a fixed α the model is linear in (B^2, a_0, a_1, a_2), so for each α on a grid we solve for them by generalised least squares with the covariance matrix $\mathbf{C}$ of the binned $\hat{\xi}$ (Section 3.10.5), and record the minimum $\chi^2(\alpha) = [\hat{\xi} - \xi_{\rm fit}]^{\rm T}\mathbf{C}^{-1}[\hat{\xi} - \xi_{\rm fit}]$. This is profiling the nuisance parameters out (Section 3.14).

What the fit asks. The observed event is the measured vector $\hat{\xi}$. The possible explanations are the values of α. For one trial value, the model predicts a vector $\xi_{\rm fit}$, and the measurement differs from that prediction by noise, a Gaussian scatter with covariance $\mathbf{C}$ (the mock covariance). So for a given α the question is: how plausible is the particular noise realisation $\hat{\xi} - \xi_{\rm fit}$ that would turn the model's prediction into the data we actually have? A Gaussian assigns that realisation

a density proportional to $e^{-\chi^2/2}$, and the same density read as a function of α, with the data held fixed, is the likelihood (Chapter 3). The best α is where the required noise is least surprising, that is, where $\chi^2(\alpha)$ is smallest. With a flat prior in α the posterior is $p(\alpha) \propto e^{-\chi^2(\alpha)/2}$, and its width is the uncertainty [Anderson2012 eqs. (27)–(29)].

11.4.3 How significant is a bump? $\Delta\chi^2$ and the number of sigma

Eisenstein and collaborators quoted their detection as "3.4σ", from the difference in χ^2 between the best model without the acoustic feature and the best model with it, $\Delta\chi^2 = 11.7$ [Eisenstein2005 §4.3]. Before reproducing the number we should see why $\sqrt{\Delta\chi^2}$ is a number of sigma.

Give the bump an amplitude A: the model is the smooth part plus A times the difference between the templates with and without the feature, so $A = 1$ is the full acoustic signal and $A = 0$ is none. The model is linear in A, and with Gaussian errors χ^2 is then exactly quadratic in A (Section 3.10.3): $\chi^2(A) = \chi^2_{\min} + (A - \hat{A})^2/\sigma_A^2$, where $\hat{A}$ is the best fit and σ_A its standard error. Then

$$\Delta\chi^2 \equiv \chi^2(0) - \chi^2(1) \overset{(1)}{=} \frac{\hat{A}^2 - (1-\hat{A})^2}{\sigma_A^2} \overset{(2)}{=} \frac{2\hat{A} - 1}{\sigma_A^2}, \qquad \text{and if } \hat{A} = 1: \quad \sqrt{\Delta\chi^2} \overset{(3)}{=} \frac{\hat{A}}{\sigma_A}. \tag{11.31}$$

(1) Insert $A = 0$ and $A = 1$ into the quadratic; $\chi^2_{\min}$ cancels.
(2) $\hat{A}^2 - (1 - 2\hat{A} + \hat{A}^2) = 2\hat{A} - 1$.
(3) With $\hat{A} = 1$, $\Delta\chi^2 = 1/\sigma_A^2 = (\hat{A}/\sigma_A)^2$.

So when the data prefer the full acoustic amplitude, $\sqrt{\Delta\chi^2}$ is the number of standard errors by which the bump's amplitude differs from zero: a signal-to-noise ratio.

The same number can be read as a test. Let A float and take as the null hypothesis "there is no acoustic bump", $A = 0$. It is the hypothesis whose sampling distribution we know: if it were true, $\chi^2(0) - \chi^2(\hat{A})$ would follow a χ^2 distribution with one degree of freedom (Wilks' theorem, Section 4.7). The question for the data is then: if there were no bump, would a $\Delta\chi^2$ as large as 11.7, or larger, be something we would reasonably expect from noise alone? The tail probability of a χ^2_1 beyond 11.7 is 6×10^{-4}, which is the two-sided Gaussian probability of lying more than $\sqrt{11.7} = 3.4$ standard deviations out (Section 4.9.1). So noise alone would produce a bump this strong about once in sixteen hundred surveys. Both readings give the same rule of thumb: for one extra parameter, the square root of the drop in χ^2 counts the sigmas.

11.4.4 Reproducing the first detection on mock surveys

What the paper compared. Eisenstein and collaborators measured $\xi(s)$ in redshift space, spherically averaged, with the Landy–Szalay estimator (Section 11.2) and FKP-like weights with $P_w = 4{\times}10^4$, in 20 bins out to 180 h^{-1}Mpc. The covariance came from 1278 mock catalogues. They fitted models of ξ with a free amplitude and free $\Omega_m h^2$ and a dilation of the scale, (11.28). The best model with baryons had $\chi^2 = 16.1$ for 17 degrees of freedom; the best pure cold-dark-matter model, which has no acoustic feature, had $\chi^2 = 27.8$, so $\Delta\chi^2 = 11.7$, "3.4σ" by (11.31) [Eisenstein2005 §3.1, §4.3, Fig. 6]. The quantity to reproduce is $\Delta\chi^2$: how strongly a survey of that size prefers a template with the peak over one without it.

Our nerfed version. The script `code/ch11/04_bao_mocks.py` replaces the survey by a periodic box of the same volume, side 900 h^{-1}Mpc (0.73 h^{-3}Gpc^3), holding the same number of galaxies, $\bar{n} = 6.49 \times 10^{-5}\ h^3\text{Mpc}^{-3}$ (about 47329 per box). The galaxy field is lognormal with spectrum $b^2 D^2(z)\, K_0\, P_t(k)$, with $z = 0.35$, $b = 2$, growth factor $D = 0.831$, damping $\Sigma_{\rm nl} = 8\ h^{-1}$Mpc, and

$K_0 = 1.260$ the angle-averaged boost of the clustering amplitude by galaxy motions (Section 11.5). Its amplitude, $\sigma_{8,g} = 1.51$, is close to the value 1.8 that the paper assumed for the LRGs. In the box there are no edges, so ξ is estimated by Fourier-transform pair counting on the grid (96^3 cells), which is what all four estimators of Section 11.2 reduce to without edges. We use 16 bins of 10 h^{-1}Mpc from 20 to 180 h^{-1}Mpc. 1500 mocks give the covariance, with the Hartlap factor 0.989 (Section 8.13.2); 500 further mocks are each fitted with (11.30), once with the template (11.29) and once with the no-wiggle template, on a grid of $\alpha \in [0.8, 1.2]$.

Listing 11.5: Templates, mocks and the fit (`code/ch11/04_bao_mocks.py`)

```
def template_bins(P, alpha):
    """Bin-averaged grid two-point function of the spectrum P(k/alpha)/alpha^3, i.e.
    xi(alpha r)."""
    return rb(xi_box_from_pk(lambda k: P(k / alpha) / alpha ** 3, N, L, 3))


t0 = time.time()
T_w = np.array([template_bins(P_w, a) for a in ALPHAS])     # (n_alpha, n_bins)
T_nw = np.array([template_bins(P_nw, a) for a in ALPHAS])
print(f"templates: {time.time()-t0:.0f} s")

# ---- mocks: lognormal galaxy field with the template spectrum, Poisson galaxies, xi by FFT
ln = Lognormal(xi_box_from_pk(P_w, N, L, 3), N, L, 3)
xis, ngal = [], []
for s in range(NCOV + NFIT):
    delta = ln.draw(rng)
    c = poisson_sample(1 + delta, NBAR, L, rng)
    dn = c / c.mean() - 1                       # density contrast of the galaxies, mean
    from the box
    xis.append(rb(xi_grid_fft(dn)))
    ngal.append(c.sum())
xis = np.array(xis)
print(f"mocks: {time.time()-t0:.0f} s")
xc, xf = xis[:NCOV], xis[NCOV:]
C = np.cov(xc, rowvar=False)
nb = len(r)
hartlap = (NCOV - nb - 2) / (NCOV - 1)
Ci = hartlap * np.linalg.inv(C)


def design(t, poly=True):
    """Columns of the linear part of the model: B^2 times the template, then a0, a1/r,
    a2/r^2."""
    return np.column_stack([t, np.ones(nb), 1 / r, 1 / r ** 2]) if poly else t[:, None]


def fit(d, T, poly=True):
    """chi^2(alpha) with the linear coefficients solved by generalised least squares at
    each alpha."""
    chi2 = np.empty(len(ALPHAS))
    for i, t in enumerate(T):
        X = design(t, poly)
        A = X.T @ Ci @ X
        beta = np.linalg.solve(A, X.T @ Ci @ d)
        res = d - X @ beta
        chi2[i] = res @ Ci @ res
```

`template_bins` builds the model for any α through the Fourier form of (11.28): the spectrum $P(k/\alpha)/\alpha^3$ is turned into the grid correlation function and averaged in the same bins as the data, so that the model and the measurement see exactly the same grid. The mock loop (lines 67–72) is Section 11.1 in four lines: lognormal field, Poisson galaxies, density contrast with the

box's own mean (so the integral constraint is present), pair counting by FFT. Lines 76–79 form the covariance and its Hartlap-corrected inverse. In `fit`, line 93 is the generalised least-squares solution $(X^{\mathsf{T}}\mathbf{C}^{-1}X)^{-1}X^{\mathsf{T}}\mathbf{C}^{-1}\hat{\boldsymbol{\xi}}$ for the four linear coefficients at each α.

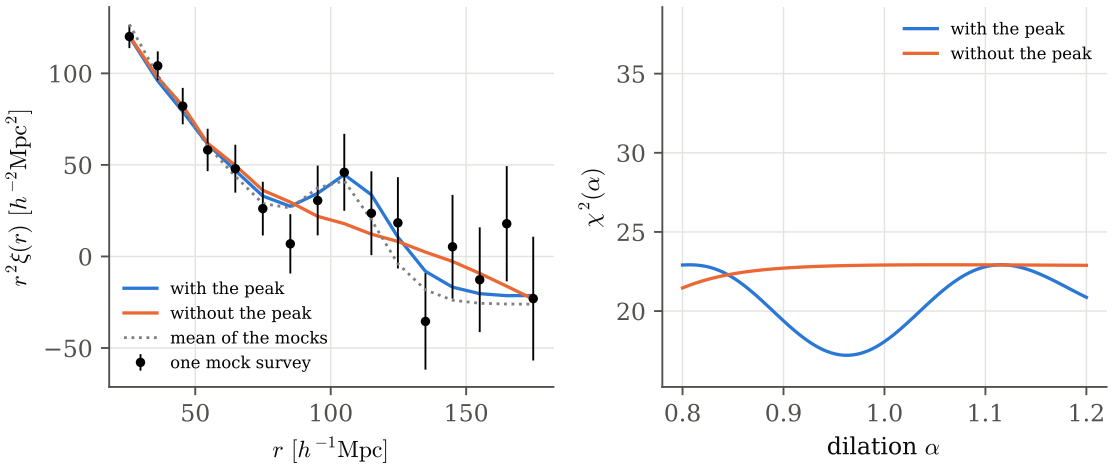


Figure 11.11: One mock survey, chosen as the one with the median $\Delta\chi^2$ among the 500 fitted. Left: $r^2\hat{\xi}(r)$ (points, with the diagonal errors of the mock covariance; neighbouring points are correlated), the mean over mocks (dotted) and the best fits with and without the acoustic feature. Right: $\chi^2(\alpha)$ for both templates. The template with the peak has a minimum near $\alpha = 1$; the one without is almost flat, because nothing in it marks a scale.

Figure 11.11 shows a typical mock. Its best fit with the peak has $\chi^2 = 17.2$ for 11 degrees of freedom and $\hat{\alpha} = 0.963$; the no-wiggle template is worse by $\Delta\chi^2 = 4.3$. Over all fitted mocks the mean χ^2 of the peak template is 10.8, close to the 11 expected if the covariance is right, and the mean $\hat{\alpha}$ is 0.9971 ± 0.0028: the fit is unbiased.

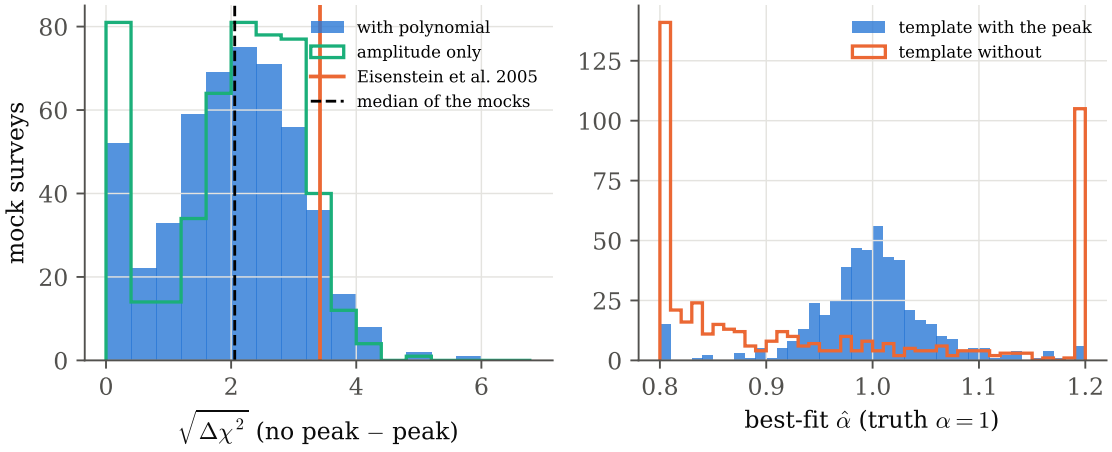


Figure 11.12: Left: the detection significance $\sqrt{\Delta\chi^2}$ over 500 mock surveys of the size of the SDSS LRG sample, with the broad-band polynomial (filled) and with the amplitude as the only free coefficient (green outline). The dashed line is the median, the orange line the 3.4σ of the 2005 paper. Right: the best-fit dilation with the peak template (filled) and with the no-wiggle template (outline), which wanders to the edges of the allowed range.

Figure 11.12 is the reproduction. A survey of that volume and density typically detects the acoustic peak at $\sqrt{\Delta\chi^2} = 2.1$ (the middle 68% of mocks lie between 0.9 and 3.1), and 7% of mock surveys reach the paper's $\Delta\chi^2 = 11.7$. Fitting as the paper did, with the amplitude

as the only free coefficient, changes little: median 2.2, and 6% reach 3.4σ. So the qualitative result reproduces: a survey of $0.7\,h^{-3}\mathrm{Gpc}^3$ at $\bar{n}P \sim 1$ does see the peak, at a modest significance. The 2005 detection sits in the upper tail. Part of that is the luck of one realisation, which the paper's own Figure 2 hints at (the measured points sit above every model). Part is that their comparison model, cold dark matter without baryons, differs from the baryonic model in its broad shape as well as in the peak, and their fit reached down to $10\,h^{-1}\mathrm{Mpc}$, where that shape difference adds to $\Delta\chi^2$; our no-wiggle template differs from the peak template *only* in the peak.

The dilation is measured to 6.3% (standard deviation of $\hat{\alpha}$ over mocks; the median per-mock error from $p(\alpha)$ is 5.9%), against the paper's 4.7% on $D_V(0.35)$, which also includes the shape information. Without the peak, the dilation is barely constrained (15.6%, limited by the edges of the grid). The paper found the same contrast: its pure-CDM fit constrained the scale a factor of two worse [Eisenstein2005 §4.3]. The distance information is in the peak.

11.4.5 How well can a survey locate the peak? A Fisher forecast

The mocks answer the question for one survey at the cost of two thousand simulations. The Fisher matrix of Chapter 10 answers it for any survey in a fraction of a second. The data are the Fourier modes $\delta_{\boldsymbol{k}}$ of the galaxy field; the parameter is α. For a Gaussian mode whose variance depends on the parameter, the Fisher information is $F = \frac{1}{2}\,\mathrm{tr}[\mathbf{C}^{-1}\mathbf{C}_{,\alpha}\mathbf{C}^{-1}\mathbf{C}_{,\alpha}]$ (Section 10.3; [HobsonBMC §5.3.1, eq. (5.17)]). A complex mode $\delta_{\boldsymbol{k}}$ is two real Gaussian numbers, its real and imaginary parts, each with variance $\frac{V}{2}[P(k) + 1/\bar{n}]$ in a box of volume V with uniform density (Section 7.3.3). Modes at $\boldsymbol{k}$ and $-\boldsymbol{k}$ are the same mode, so a shell of width $\mathrm{d}k$ holds $\frac{1}{2}\,V\,4\pi k^2\mathrm{d}k/(2\pi)^3$ independent complex modes. Then

$$F_{\alpha\alpha} \overset{(1)}{=} \sum_{\text{modes}} 2 \times \frac{1}{2}\Big(\frac{\partial_\alpha P}{P + 1/\bar{n}}\Big)^2 \overset{(2)}{=} \int \frac{V\,4\pi k^2\,\mathrm{d}k}{2(2\pi)^3}\Big(\frac{\bar{n}P}{1+\bar{n}P}\Big)^2\Big(\frac{\partial \ln P}{\partial\alpha}\Big)^2$$
$$\overset{(3)}{=} \int \frac{k^2\,\mathrm{d}k}{4\pi^2}\,V_{\mathrm{eff}}(k)\Big(\frac{\partial \ln P}{\partial\alpha}\Big)^2. \tag{11.32}$$

(1) For one real Gaussian variable of variance $v(\alpha)$, $\frac{1}{2}(v_{,\alpha}/v)^2$; each complex mode contributes two such variables, and the shot noise does not depend on α.
(2) Replace the sum by the mode count of the shell, and write $\partial_\alpha P/(P + 1/\bar{n}) = \frac{\bar{n}P}{1+\bar{n}P}\,\partial_\alpha \ln P$.
(3) $4\pi/(2(2\pi)^3) = 1/(4\pi^2)$, and $V[\bar{n}P/(1 + \bar{n}P)]^2 = V_{\mathrm{eff}}(k)$, which for a survey with varying $\bar{n}$ becomes the integral (11.24).

This is the forecast formula of Seo and Eisenstein [SeoEisenstein2003 eq. (9)], and it is the same effective volume that measured the error of $\hat{P}$ in Section 11.3. The derivative must contain only the information in the peak, since the broad band is marginalised in the fit. With $P(k;\alpha) = P_{\mathrm{nw}}(k)\,[1 + O(k/\alpha)]$, where $O = (P_t - P_{\mathrm{nw}})/P_{\mathrm{nw}}$ is the damped wiggle pattern,

$$\frac{\partial \ln P}{\partial\alpha}\Big|_{\alpha=1} = -\frac{k\,O'(k)}{1 + O(k)}. \tag{11.33}$$

The error forecast is $\sigma_\alpha = F_{\alpha\alpha}^{-1/2}$.

Listing 11.6: The Fisher error on α (`code/ch11/05_bao_fisher.py`)

```
def dlnp_dalpha(sig_nl):
    """d ln P(k/alpha)/d alpha at alpha = 1 with the broad band held fixed: -k O'(k)/(1 + O)."""
    O = (p_lin / p_nw - 1) * np.exp(-0.5 * (k * sig_nl) ** 2)
    return -k * np.gradient(O, k) / (1 + O)


```

```python
def sigma_alpha(veff_of_P, P, sig_nl, kmin=2e-3, kmax=0.3):
    """Fisher error on alpha. veff_of_P maps the galaxy power P(k) to V_eff(k)
    [(Mpc/h)^3]."""
    sel = (k >= kmin) & (k <= kmax)
    integrand = k ** 2 / (4 * np.pi ** 2) * veff_of_P(P) * dlnp_dalpha(sig_nl) ** 2
    F = np.trapezoid(integrand[sel], k[sel])
    return 1 / np.sqrt(F), integrand
```

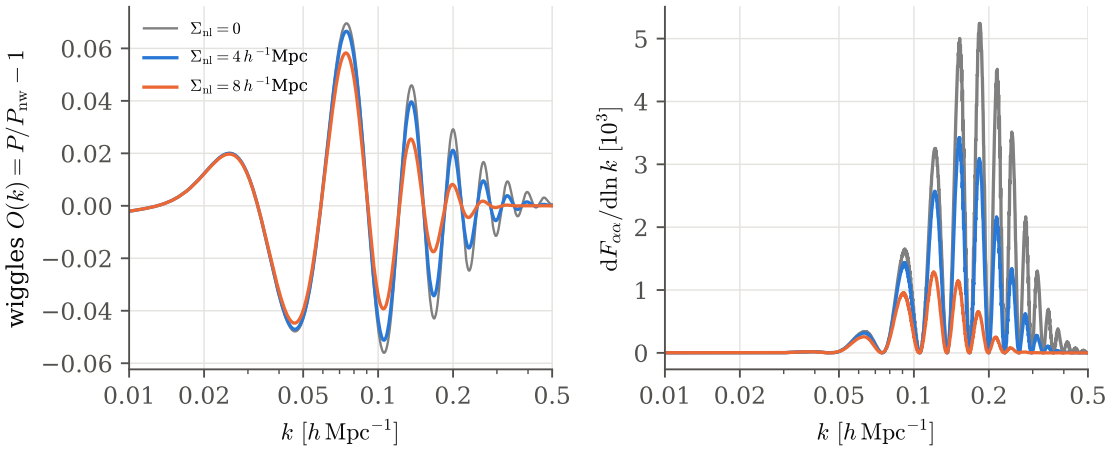


Figure 11.13: Where the information on α comes from. Left: the wiggle pattern $O(k)$ undamped (grey) and damped with $\Sigma_{\rm nl} = 4$ (blue) and $8\,h^{-1}$Mpc (orange). Right: the Fisher integrand of (11.32) per $\ln k$ for the mock box, with the same colours. The information sits in the wiggles between $k \approx 0.05$ and $0.3\,h\,\mathrm{Mpc}^{-1}$; damping erases the higher harmonics.

Example: the mock box. For the box of Section 11.4.4, (11.32) predicts $\sigma_\alpha = 4.1\%$. The mocks gave 6.3% (scatter of $\hat{\alpha}$) and 5.9% (median per-mock error). The forecast is optimistic, and it is supposed to be: the inverse Fisher matrix is a lower bound (Section 10.2), reached only when the likelihood for α is close to Gaussian. At a typical detection of 2σ it is not: $\chi^2(\alpha)$ in Figure 11.11 is shallow, with secondary minima where a noise fluctuation mimics the peak, and in some mocks the fit wanders to the edge of the grid. The bound becomes accurate only for high signal-to-noise. The lognormal field's non-Gaussian covariance, seen already in Section 11.3.3, adds to the gap.

Example: the 2005 survey. For the survey rebuilt from the paper's numbers in Section 11.3 (with $\bar{n}(z)$, the paper's galaxy amplitude $\sigma_{8,g} = 1.8$, and $\Sigma_{\rm nl} = 8\,h^{-1}$Mpc), the forecast is $\sigma_\alpha = 3.4\%$. The paper reports $D_V(0.35) = 1370 \pm 64$ Mpc, a 4.7% distance. Compared with the distance to the last-scattering surface it is the ratio $R_{0.35} = D_V(0.35)/D_M(1089) = 0.0979 \pm 0.0036$, a 3.7% measurement. And the dimensionless combination $A = D_V(0.35)\sqrt{\Omega_m H_0^2}/(0.35\,c) = 0.469 \pm 0.017$ (3.6%) is the distance measured in units of $c/(\sqrt{\Omega_m} H_0)$, a combination that depends on the other parameters only weakly [Eisenstein2005 Table 1].

Example: DESI's bright galaxies. DESI's first data release measured, from 300,017 bright galaxies at $0.1 < z < 0.4$, $D_V/r_d = 7.93 \pm 0.15$ at $z_{\rm eff} = 0.295$, a 1.9% measurement, after reconstruction, with a quoted effective volume of $1.7\,\mathrm{Gpc}^3$ [DESI2024VI Table 1]. In our units that is $0.52\,h^{-3}\mathrm{Gpc}^3$. Taking that $V_{\rm eff}$ as if it held at every k, (11.32) forecasts 1.1% with the post-reconstruction damping $\Sigma_{\rm nl} = 4\,h^{-1}$Mpc and 2.2% with the pre-reconstruction $8\,h^{-1}$Mpc. The measured error lies between the two. The simple forecast is optimistic for the same reasons

as before, plus one: a single $V_{\rm eff}$ quoted at one reference wavenumber overstates the information at higher k, where $\bar{n}P$ is smaller.

How this is used in real research

Our mock boxes play the role of a *verification*: the Fisher formula (11.32) is a known answer for Gaussian modes, and the mocks check it and show where it fails (low signal-to-noise, non-Gaussian covariance). In a survey analysis the same machinery becomes the *research instrument*. For a real footprint, with its mask, its varying $\bar{n}$, fibre collisions and reconstruction, there is no formula for the covariance of $\hat{\xi}$ or for the scatter of $\hat{\alpha}$; the only way to obtain them is to push a thousand or so mock catalogues through the full pipeline, or to calibrate a semi-analytic covariance against them (Section 11.6).

11.4.6 Comparing with the data: DESI distances against Planck

Figure 11.14 puts the DESI DR1 distances next to the prediction of the Planck 2018 ΛCDM model. For the anisotropic measurements we form $D_V/r_d = (z\,(D_M/r_d)^2\,(D_H/r_d))^{1/3}$ from the published D_M/r_d and D_H/r_d, with the error from the delta method (Section 3.6.3): for $\ln D_V = \frac{1}{3}(2\ln D_M + \ln D_H + \ln z)$,

$$\frac{\sigma_{D_V}}{D_V} = \frac{1}{3}\sqrt{4\left(\frac{\sigma_{D_M}}{D_M}\right)^2 + \left(\frac{\sigma_{D_H}}{D_H}\right)^2 + 4\rho\,\frac{\sigma_{D_M}}{D_M}\frac{\sigma_{D_H}}{D_H}}, \tag{11.34}$$

where ρ is the published correlation coefficient between the two.

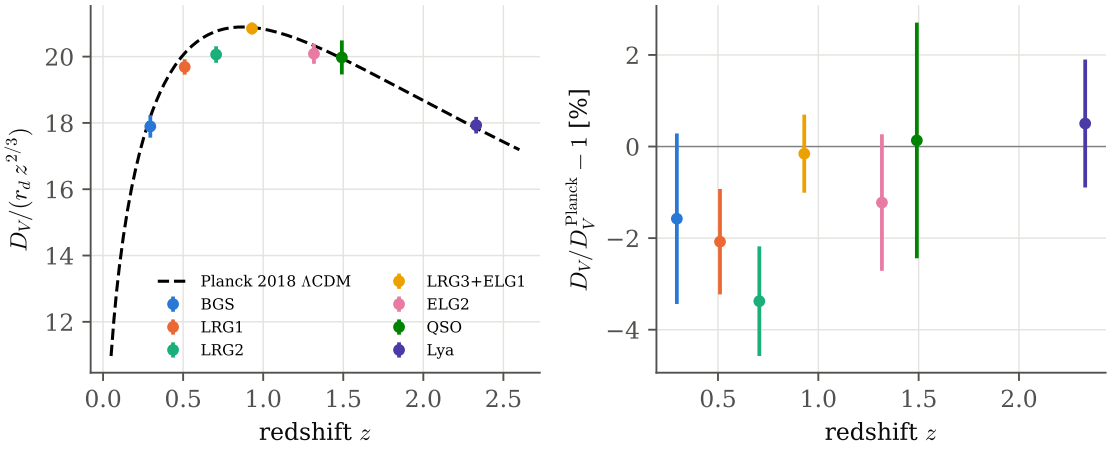


Figure 11.14: DESI DR1 BAO distances (points; [106], Table 1) against the Planck 2018 ΛCDM prediction (dashed). Left: D_V/r_d, divided by $z^{2/3}$ to flatten the curve. Right: the fractional difference from the prediction.

The bright-galaxy point, 7.93 ± 0.15, sits -0.8σ from the Planck prediction 8.06. The largest deviation is the luminous red galaxy bin at $z = 0.71$, 15.91 ± 0.20 against 16.46, a 2.8σ pull. Taken together, the 7 points give $\chi^2 = 12.8$ for 7 degrees of freedom, a p-value of 0.08. How surprised should we be by one 2.8σ point? Among 7 independent points, the chance that at least one deviates by that much is 0.03, which is the look-elsewhere correction of Section 4.14. The data are compatible with the Planck model at the level of this simple comparison; the cosmological fits in the DESI paper, which use the two directions separately and add other data, are where the interesting tensions are argued.

Where we have met H_0 before. These are the same twelve DESI numbers that Section T2.15 fitted on a grid in (Ω_m, hr_d), finding $hr_d = 101.8 \pm 1.3$ Mpc; with the sound horizon $r_d = 147.09$ Mpc

taken from the CMB, that became an inverse distance ladder, $H_0 = 69.2 \pm 0.9$, near Planck's 67.36 ± 0.54 and well below the 73.04 ± 1.04 of the supernova ladder (Section T2.12.3). The degeneracy is the same here: every D/r_d depends on H_0 and r_d only through their product, as the supernova magnitudes of Section 3.17.4 depended on H_0 and M only through $M - 5\log_{10} H_0$. What differs is that nothing is fitted. Planck's model, with its $H_0 = 67.36$ and $r_d = 147.10\,\mathrm{Mpc}$, is held fixed, so the comparison tests their product, $hr_d = 99.1\,\mathrm{Mpc}$, and the shape of the distance curve, never H_0 on its own; and each bin is reduced to the single number D_V/r_d. The verdicts differ for that reason: there, $\Delta\chi^2 = 8.3$ between the Planck point and the best fit, with two degrees of freedom, gave $p = 0.016$; here the absolute χ^2 spreads the same disagreement over 7 degrees of freedom and gives $p = 0.08$. Section 11.20 sets this route beside the other routes to H_0.

11.5 Galaxy motions: clustering in redshift space

A survey never measures how far away a galaxy is. It measures a redshift, and turns it into a distance with the Hubble law of a fiducial cosmology. The redshift, though, is not only the stretching of light by the expansion. Every galaxy also moves through space with its own **peculiar velocity**, a few hundred kilometres per second, and the part of that velocity along our line of sight adds a Doppler shift. The survey therefore places each galaxy a little in front of or a little behind its true position, and only along the line of sight.

If these velocities were random, the displacement would be a nuisance that blurs the map. They are not random. Galaxies fall towards overdense regions and stream out of empty ones, pulled by the same gravity that makes the overdensities grow. On large scales a galaxy in front of a supercluster falls away from us, towards the supercluster, and is placed further away; a galaxy behind it falls towards us and is placed nearer. Both move towards the centre in the map. The supercluster looks squashed along the line of sight and denser than it is. The size of the squashing is set by how fast the infall is, and that is set by how fast structure grows. A direction-dependent clustering pattern thus becomes a measurement of the growth rate of structure, the second number, after the acoustic scale of Section 11.4, that a galaxy survey hands to cosmology.

The question, in words:

> *If every galaxy is moved along the line of sight by its own velocity, and the velocities are the ones that gravity produces while structure is still small, how does the power spectrum of the moved galaxies depend on the angle between a Fourier mode and the line of sight?*

11.5.1 From a redshift to a misplaced galaxy

The event. A galaxy is observed at redshift z_{obs}, and we put it at the comoving distance $\chi(z_{\mathrm{obs}})$.

What produced that redshift? Two shifts, one after the other. Think of a comoving observer [one at rest with respect to the expansion] sitting right next to the galaxy. The galaxy moves away from this neighbour with line-of-sight velocity u_r, so the neighbour sees light of emitted frequency ν_e at $\nu' = \nu_e/(1 + u_r/c)$, the ordinary non-relativistic Doppler shift ($u_r \sim 300\,\mathrm{km\,s^{-1}}$ is a thousandth of c). That light then travels to us through the expanding Universe and is stretched by the cosmological factor $1 + z_c$, where z_c is the redshift the galaxy would have if it were at rest: $\nu_0 = \nu'/(1 + z_c)$. Frequency ratios multiply, so

$$1 + z_{\mathrm{obs}} = \frac{\nu_e}{\nu_0} = (1 + z_c)\Big(1 + \frac{u_r}{c}\Big). \tag{11.35}$$

(Our own motion adds a third factor of the same kind; it is the same for every galaxy in a given direction and is removed by quoting redshifts in the rest frame of the CMB.) The peculiar velocity is $\boldsymbol{u} = a\,\dot{\boldsymbol{x}}$, the rate of change of comoving position $\boldsymbol{x}$ times the scale factor a, which is how the cosmology notes define it; $u_r = \boldsymbol{u}\cdot\hat{\boldsymbol{r}}$ is its component along the line of sight $\hat{\boldsymbol{r}}$.

Where does the galaxy land? We convert $z_{\rm obs}$ with the comoving distance $\chi(z) = \int_0^z c\,\mathrm{d}z'/H(z')$. Calling the assigned distance s,

$$s \equiv \chi(z_{\rm obs}) \overset{(1)}{=} \chi(z_c) + \left.\frac{\mathrm{d}\chi}{\mathrm{d}z}\right|_{z_c}(z_{\rm obs}-z_c) \overset{(2)}{=} \chi + \frac{c}{H(z)}\,(1+z)\,\frac{u_r}{c} \overset{(3)}{=} \chi + \frac{u_r}{aH}. \tag{11.36}$$

(1) A first-order Taylor expansion; the shift $z_{\rm obs}-z_c$ is of order $10^{-3}(1+z)$, so the next term is a million times smaller than χ.
(2) $\mathrm{d}\chi/\mathrm{d}z = c/H(z)$ from the integral, and $z_{\rm obs}-z_c = (1+z_c)\,u_r/c$ from (11.35). We drop the subscript on z_c in factors that multiply a small quantity.
(3) $1+z = 1/a$.

The true comoving distance is χ; the galaxy is placed at s, displaced by $u_r/(aH)$ along the line of sight. In vector form, $\boldsymbol{s} = \boldsymbol{x} + \left(u_r/aH\right)\hat{\boldsymbol{r}}$. The space of the assigned positions $\boldsymbol{s}$ is called **redshift space**; the space of the true positions $\boldsymbol{x}$ is **real space**.

Example: how far is a galaxy moved? Near us, $a = 1$ and $H = H_0 = 100\,h\,\mathrm{km\,s^{-1}Mpc^{-1}}$. A galaxy falling at $300\,\mathrm{km\,s^{-1}}$ is moved by $300/100 = 3\,h^{-1}\mathrm{Mpc}$. A galaxy orbiting inside a rich cluster at $1000\,\mathrm{km\,s^{-1}}$ is moved by $10\,h^{-1}\mathrm{Mpc}$. The first number is small against the $100\,h^{-1}\mathrm{Mpc}$ acoustic scale, but not against the 10–30 $h^{-1}\mathrm{Mpc}$ separations where most pairs sit; the second is larger than the cluster itself.

For the rest of the section we look at a patch of the survey far from us and small on the sky, so that every line of sight in it points in nearly the same direction. Call that direction $\hat{\boldsymbol{z}}$. This is the **distant-observer approximation**. The mapping is then a shift along one fixed axis,

$$\boldsymbol{s} = \boldsymbol{x} + v_z(\boldsymbol{x})\,\hat{\boldsymbol{z}}, \qquad v_z \equiv \frac{u_z}{aH}, \tag{11.37}$$

where v_z is the line-of-sight velocity expressed as a comoving distance. For a Fourier mode with wavevector $\boldsymbol{k}$, let $\mu = \hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{z}}$ be the cosine of its angle with the line of sight: $\mu = \pm 1$ for a wave whose crests are perpendicular to the line of sight, $\mu = 0$ for crests along it.

11.5.2 Counting galaxies twice: the Kaiser factor

We want the density contrast of the galaxies in redshift space, $\delta_s(\boldsymbol{s})$, in terms of the real-space field. We state the inputs first, as in Section 11.1.1.

Physical inputs:

(K1) *Galaxies are conserved by the mapping.* Each galaxy appears exactly once in the redshift-space catalogue. This is bookkeeping, and exact.

(K2) *Linear theory.* The density contrast and the velocities are small, and the matter velocity field obeys the linear continuity equation of the cosmology notes,

$$\theta \equiv \frac{\nabla\cdot\boldsymbol{u}}{aH} = -f\,\delta, \qquad f \equiv \frac{\mathrm{d}\ln D}{\mathrm{d}\ln a}, \tag{11.38}$$

with δ the matter contrast, $D(a)$ the linear growth factor [how much a small fluctuation has grown by scale factor a] and f the **growth rate**. The velocity field has no curl (linear rotational modes decay; see the cosmology notes).

(K3) *Galaxies trace the matter.* Their density contrast is $\delta_g = b\,\delta$ with a constant linear bias b, and they move with the matter, because gravity accelerates a galaxy and the dark matter around it in the same way.

Geometric input: the distant-observer approximation of (11.37). *Mathematical input:* the mapping $\boldsymbol{x} \mapsto \boldsymbol{s}$ is smooth and one-to-one, which holds while v_z changes by much less than the distance over which it changes, that is while $|\partial_z v_z| \ll 1$. (In a collapsing cluster it fails: two galaxies at different true positions can land at the same $\boldsymbol{s}$.)

The galaxies in a small volume, counted twice. The galaxies inside a small real-space volume d^3x around $\boldsymbol{x}$ end up inside a redshift-space volume d^3s around $\boldsymbol{s}$, and by (K1) they are the same galaxies. With $\bar{n}$ the mean density (the same in both spaces, since the shifts average to zero), $\bar{n}[1+\delta_s(\boldsymbol{s})]\,\mathrm{d}^3s = \bar{n}[1+\delta_g(\boldsymbol{x})]\,\mathrm{d}^3x$. The ratio of the volumes is the Jacobian of the mapping (11.37). Its matrix $\partial s_i/\partial x_j$ has rows $(1,0,0)$, $(0,1,0)$ and $(\partial_x v_z, \partial_y v_z, 1+\partial_z v_z)$; it is triangular, so its determinant is the product of the diagonal entries, $1+\partial_z v_z$. Hence

$$1+\delta_s(\boldsymbol{s}) \overset{(1)}{=} \frac{1+\delta_g(\boldsymbol{x})}{1+\partial_z v_z(\boldsymbol{x})} \overset{(2)}{=} 1+\delta_g(\boldsymbol{x}) - \partial_z v_z(\boldsymbol{x}) + O(\delta^2),$$
$$\delta_s(\boldsymbol{x}) \overset{(3)}{=} \delta_g(\boldsymbol{x}) - \partial_z v_z(\boldsymbol{x}) + O(\delta^2). \tag{11.39}$$

(1) Number conservation, divided by $\bar{n}\,\mathrm{d}^3s$, with $\mathrm{d}^3x/\mathrm{d}^3s = 1/(1+\partial_z v_z)$.
(2) $1/(1+\epsilon) = 1-\epsilon+\epsilon^2-\ldots$, and the product $\delta_g\,\partial_z v_z$ is second order, so it is dropped with the rest.
(3) $\delta_s(\boldsymbol{s}) = \delta_s(\boldsymbol{x}+v_z\hat{\boldsymbol{z}}) = \delta_s(\boldsymbol{x}) + v_z\,\partial_z\delta_s + \ldots$; the correction is again a product of two small quantities.

So in linear theory the redshift-space contrast is the real-space contrast minus the line-of-sight gradient of the line-of-sight velocity. Where galaxies on the near side move away from us and those on the far side move towards us, $\partial_z v_z < 0$: the galaxies crowd together in the map and δ_s grows.

The velocity in terms of the density. Now use (K2). Since the velocity field has no curl it is a gradient, $\boldsymbol{u}/(aH) = \nabla\phi$ for some potential ϕ. In Fourier space a gradient is multiplication by $i\boldsymbol{k}$. Then

$$-k^2\phi_{\boldsymbol{k}} \overset{(1)}{=} i\boldsymbol{k}\cdot\left(i\boldsymbol{k}\,\phi_{\boldsymbol{k}}\right) \overset{(2)}{=} \theta_{\boldsymbol{k}} \overset{(3)}{=} -f\,\delta_{\boldsymbol{k}}, \qquad \text{so} \qquad \left(\frac{\boldsymbol{u}}{aH}\right)_{\boldsymbol{k}} = i\boldsymbol{k}\,\phi_{\boldsymbol{k}} = \frac{i\boldsymbol{k}}{k^2}\,f\,\delta_{\boldsymbol{k}},$$
$$\left(\partial_z v_z\right)_{\boldsymbol{k}} \overset{(4)}{=} (ik_z)\,\frac{ik_z}{k^2}\,f\,\delta_{\boldsymbol{k}} \overset{(5)}{=} -f\mu^2\,\delta_{\boldsymbol{k}}. \tag{11.40}$$

(1) $\boldsymbol{k}\cdot\boldsymbol{k} = k^2$ and $i^2 = -1$.
(2) The divergence of $\boldsymbol{u}/(aH) = \nabla\phi$ is θ by the definition in (11.38); in Fourier space a divergence is $i\boldsymbol{k}\cdot$.
(3) The continuity equation (11.38).
(4) v_z is the z component of $\boldsymbol{u}/(aH)$, and ∂_z is multiplication by ik_z.
(5) $i^2 = -1$ and $k_z/k = \mu$.

Putting (11.40) into (11.39) with $\delta_g = b\,\delta$ from (K3),

$$\delta_{s,\boldsymbol{k}} = \left(b + f\mu^2\right)\delta_{\boldsymbol{k}}, \qquad P_s(k,\mu) = \left(b+f\mu^2\right)^2 P(k) = b^2\left(1+\beta\mu^2\right)^2 P(k), \quad \beta \equiv \frac{f}{b}, \tag{11.41}$$

where $P(k)$ is the matter power spectrum and $P_s(k,\mu)$ is defined, as in Section 11.1, by $\mathbb{E}\,|\delta_{s,\boldsymbol{k}}|^2 = V\,P_s(k,\mu)$ for a box of volume V. This is the **Kaiser formula** [108]. The density term carries the bias; the velocity term does not, because galaxies move with the matter whatever

their bias. So the shape of P_s in μ measures the ratio $\beta = f/b$, and the μ^4 term on its own measures f times the amplitude of the matter field, with no bias in it. Shot noise, from Section 11.1, adds $1/\bar{n}$ to (11.41) unchanged: the mapping moves galaxies but does not make or destroy them.

Key idea: The Kaiser formula

In linear theory and for a distant observer, infall moves galaxies along the line of sight in such a way that a Fourier mode at angle $\arccos\mu$ to the line of sight is amplified by $b + f\mu^2$ instead of b: $P_s(k,\mu) = (b + f\mu^2)^2 P(k)$. Modes across the line of sight ($\mu = 0$) are unchanged; modes along it are boosted by $(1+\beta)^2$.

Example: one wave along the line of sight. Take a single matter wave $\delta = A\cos kz$ with crests perpendicular to the line of sight ($\mu = 1$), $b = 1$ and a small amplitude A. For a wave that depends only on z, the divergence is ∂_z, so the continuity equation (11.38) reads $\partial_z v_z = -fA\cos kz$, and integrating, $v_z = -(fA/k)\sin kz$. Just above the crest at $z = 0$, $\sin kz > 0$ and $v_z < 0$: galaxies are moved down, towards the crest. Just below it they are moved up. Around a trough, at $kz = \pi$, the signs reverse and galaxies are moved away. Crests are piled up and troughs are emptied, and (11.39) gives the size:

$$1 + \delta_s \overset{(1)}{=} \frac{1 + A\cos kz}{1 - fA\cos kz} \overset{(2)}{=} 1 + (1+f)\,A\cos kz + O(A^2).$$

(1) $\partial_z v_z = -fA\cos kz$ in the denominator of (11.39).
(2) Expand the denominator to first order in A.

The amplitude grows from A to $(1+f)A$, which is $b + f\mu^2$ at $b = 1$, $\mu = 1$. With $f = 0.8$ the power of this mode rises by $(1.8)^2 = 3.24$. Figure 11.15 draws it.

Example: one wave across the line of sight. Now take $\delta = A\cos kx$, crests running along the line of sight ($\mu = 0$). The velocity points along $\hat{\boldsymbol{x}}$, so $u_z = 0$, nobody is moved along the line of sight, and $\delta_s = \delta$. The factor $b + f\cdot 0^2 = b$ agrees.

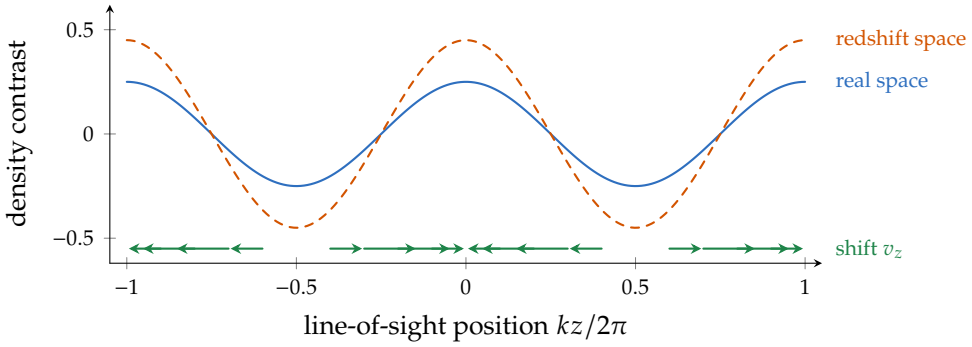


Figure 11.15: A density wave along the line of sight (blue) and the same wave in redshift space (orange, dashed), drawn for $f = 0.8$ and an exaggerated amplitude. The green arrows at the bottom are the line-of-sight shifts $v_z = -(fA/k)\sin kz$: they point towards the crests at $kz/2\pi = 0, \pm 1$ and away from the troughs at $\pm\frac{1}{2}$. Galaxies pile up on the crests, and the wave's amplitude grows by the factor $1 + f$.

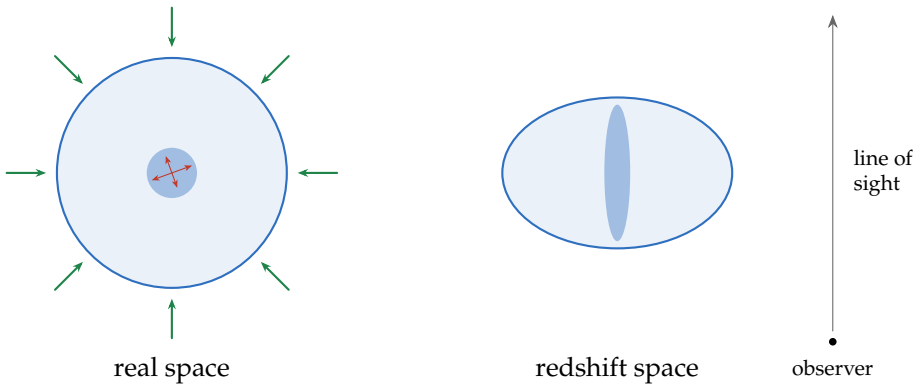


Figure 11.16: Left: a spherical overdensity in real space. On large scales matter falls in coherently (green arrows); in the virialised core galaxies orbit with large random velocities (red). Right: the same structure in redshift space. The coherent infall squashes the large shell along the line of sight (the Kaiser effect); the random motions in the core stretch it into a finger pointing at the observer (the finger of God, Section 11.5.5).

11.5.3 Multipoles: three numbers per wavenumber

The Kaiser formula turns the one-variable power spectrum $P(k)$ into a function of two variables, $P_s(k,\mu)$. Measuring a two-dimensional function in fine bins of both k and μ spreads the modes thin. But (11.41) is a polynomial in μ^2 of degree two, so a few numbers per k hold all of its μ dependence. The natural numbers are the coefficients in Legendre polynomials, because those are orthogonal on $-1 \le \mu \le 1$: $\int_{-1}^{1} L_\ell(\mu) L_{\ell'}(\mu)\,\mathrm{d}\mu = \frac{2}{2\ell+1}\delta_{\ell\ell'}$, with $L_0 = 1$, $L_2 = \frac{1}{2}(3\mu^2-1)$ and $L_4 = \frac{1}{8}(35\mu^4 - 30\mu^2 + 3)$.

Working definition 11.4 (Redshift-space multipoles)

The **multipoles** of the redshift-space power spectrum are

$$P_\ell(k) = \frac{2\ell+1}{2}\int_{-1}^{1} P_s(k,\mu)\, L_\ell(\mu)\,\mathrm{d}\mu, \qquad P_s(k,\mu) = \sum_\ell P_\ell(k)\, L_\ell(\mu),$$

the monopole P_0, quadrupole P_2 and hexadecapole P_4. *Operationally:* in a shell of modes of length near k, average $|\delta_{s,\boldsymbol{k}}|^2/V$ times $(2\ell+1)L_\ell(\mu)$ over the modes [each mode with its own $\mu = k_z/k$]; this replaces the integral by an average over the directions present in the shell.

The factor $(2\ell+1)/2$ is what orthogonality requires: multiply the expansion by L_ℓ, integrate, and only the ℓ-th term survives, with weight $2/(2\ell+1)$.

The integrals we need. Write $\langle g\rangle = \frac{1}{2}\int_{-1}^{1} g(\mu)\,\mathrm{d}\mu$ for the average over directions. Then $\langle \mu^{2n}\rangle = \frac{1}{2}\cdot\frac{2}{2n+1} = \frac{1}{2n+1}$: $\langle\mu^2\rangle = \frac{1}{3}$, $\langle\mu^4\rangle = \frac{1}{5}$, $\langle\mu^6\rangle = \frac{1}{7}$, $\langle\mu^8\rangle = \frac{1}{9}$. With $(b+f\mu^2)^2 = b^2 + 2bf\mu^2 + f^2\mu^4$,

$$\begin{aligned}
\frac{P_0}{P} &\overset{(1)}{=} \left\langle b^2 + 2bf\mu^2 + f^2\mu^4\right\rangle \overset{(2)}{=} b^2 + \frac{2bf}{3} + \frac{f^2}{5},\\
\frac{P_2}{P} &\overset{(3)}{=} \frac{5}{2}\left\langle (b^2 + 2bf\mu^2 + f^2\mu^4)(3\mu^2-1)\right\rangle \overset{(4)}{=} \frac{5}{2}\left[b^2\left(1-1\right) + 2bf\left(\frac{3}{5}-\frac{1}{3}\right) + f^2\left(\frac{3}{7}-\frac{1}{5}\right)\right]\\
&\overset{(5)}{=} \frac{4bf}{3} + \frac{4f^2}{7},\\
\frac{P_4}{P} &\overset{(6)}{=} \frac{9}{8}\left[b^2\left(7-10+3\right) + 2bf\left(5-6+1\right) + f^2\left(\frac{35}{9}-\frac{30}{7}+\frac{3}{5}\right)\right] \overset{(7)}{=} \frac{8f^2}{35}.
\end{aligned} \tag{11.42}$$

(1) The definition with $\ell = 0$, $L_0 = 1$: the monopole is the average over directions.
(2) $\langle\mu^2\rangle = \frac{1}{3}$, $\langle\mu^4\rangle = \frac{1}{5}$.
(3) $(2\ell+1)/2 \cdot 2\langle\cdot\rangle$ with $\ell = 2$ and $L_2 = \frac{1}{2}(3\mu^2 - 1)$: the factors give $5 \cdot \frac{1}{2} = \frac{5}{2}$.
(4) Multiply out and average each power: $b^2(3\langle\mu^2\rangle - 1)$, $2bf(3\langle\mu^4\rangle - \langle\mu^2\rangle)$, $f^2(3\langle\mu^6\rangle - \langle\mu^4\rangle)$.
(5) $\frac{3}{5} - \frac{1}{3} = \frac{4}{15}$ and $\frac{3}{7} - \frac{1}{5} = \frac{8}{35}$; then $\frac{5}{2} \cdot 2 \cdot \frac{4}{15} = \frac{4}{3}$ and $\frac{5}{2} \cdot \frac{8}{35} = \frac{4}{7}$.
(6) The same with $\ell = 4$: $9 \cdot \frac{1}{8}$ in front, and $35\langle\mu^{2n+4}\rangle - 30\langle\mu^{2n+2}\rangle + 3\langle\mu^{2n}\rangle$ for $n = 0, 1, 2$: $35 \cdot \frac{1}{5} - 30 \cdot \frac{1}{3} + 3 = 0$, $35 \cdot \frac{1}{7} - 30 \cdot \frac{1}{5} + 3 \cdot \frac{1}{3} = 0$.
(7) $\frac{35}{9} - \frac{30}{7} + \frac{3}{5} = \frac{1225-1350+189}{315} = \frac{64}{315}$, and $\frac{9}{8} \cdot \frac{64}{315} = \frac{8}{35}$.

Higher multipoles vanish. L_ℓ is orthogonal to every polynomial of degree below ℓ, and $(b + f\mu^2)^2$ has degree 4, so $P_\ell = 0$ for $\ell \geq 6$; the odd multipoles vanish because P_s is even in μ. Dividing by b^2 puts them in Kaiser's form: $P_0 = b^2(1 + \frac{2}{3}\beta + \frac{1}{5}\beta^2)P$, $P_2 = b^2(\frac{4}{3}\beta + \frac{4}{7}\beta^2)P$, $P_4 = b^2\,\frac{8}{35}\beta^2 P$.

Example: the boost used in the acoustic-peak mocks. The mock surveys of Section 11.4.4 were isotropic lognormal boxes whose amplitude was raised by the angle-averaged redshift-space boost. For their luminous red galaxies at $z = 0.35$ with $b = 2$ and $f = 0.706$, $\beta = 0.353$ and the monopole factor is $1 + \frac{2}{3}\beta + \frac{1}{5}\beta^2 = 1.260$. The quadrupole, which a box with an isotropic spectrum cannot have, was left out; for the position of the acoustic peak in the monopole that does not matter.

Example: β without knowing the amplitude. The ratio of the quadrupole to the monopole,

$$\frac{P_2}{P_0} = \frac{\frac{4}{3}\beta + \frac{4}{7}\beta^2}{1 + \frac{2}{3}\beta + \frac{1}{5}\beta^2}, \tag{11.43}$$

contains neither b^2 nor the amplitude of $P(k)$: both cancel. A sample with $b = 2$ and $f = 0.7$ has $\beta = 0.35$, and $P_2/P_0 = (0.467 + 0.070)/(1 + 0.233 + 0.0245) = 0.427$. Conversely, a measured ratio of 0.43 gives back $\beta \approx 0.35$ by solving the quadratic. This is the oldest use of redshift-space distortions: a ratio of two measured multipoles gives β with no knowledge of the bias. To go further, from β to f, needs the bias, which is the subject of Section 11.5.4.

11.5.4 What the anisotropy measures: $f\sigma_8$

The Kaiser formula has three unknowns, b, f and the amplitude of $P(k)$. A galaxy survey cannot tell all three apart, and it pays to see exactly which combinations it does measure.

Split the spectrum into shape and amplitude. The shape of the linear $P(k)$ is fixed by the CMB to high precision; its amplitude at the survey's redshift is conveniently measured by $\sigma_8(z)$, the rms linear matter fluctuation in spheres of $8\,h^{-1}$Mpc (Section T1.2). Write $P(k, z) = \sigma_8^2(z)\,\hat{P}(k)$, where $\hat{P}$ is the spectrum normalised to unit variance in such spheres. Then (11.41) becomes

$$P_s(k, \mu) = \left[b\sigma_8 + f\sigma_8\,\mu^2\right]^2 \hat{P}(k) = \left[(b\sigma_8)^2 + 2\mu^2(b\sigma_8)(f\sigma_8) + \mu^4(f\sigma_8)^2\right]\hat{P}(k). \tag{11.44}$$

Only the products $b\sigma_8$ and $f\sigma_8$ appear. Rescaling $(b, f, \sigma_8) \to (b/\lambda, f/\lambda, \lambda\sigma_8)$ for any $\lambda > 0$ leaves every term unchanged, so no measurement of galaxy clustering alone can separate them. Percival and White wrote the same statement for the variance in $8\,h^{-1}$Mpc spheres seen at a fixed angle μ: integrating (11.44) against the top-hat window gives $\sigma^2_{8,g}(\mu) = (b\sigma_8)^2 + 2\mu^2(b\sigma_8)(f\sigma_8) + \mu^4(f\sigma_8)^2$ [PercivalWhite2009 eq. (14)], because $\hat{P}$ integrates to one by construction. The monopole-to-quadrupole ratio (11.43) fixes $\beta = f\sigma_8/b\sigma_8$; the monopole's amplitude then fixes $b\sigma_8$; and together they give $f\sigma_8$.

Why $f\sigma_8$ is the natural number. With D normalised to $D = 1$ today, the amplitude grows as $\sigma_8(z) = \sigma_8\, D(z)$, and

$$f\sigma_8(z) \overset{(1)}{=} \frac{\mathrm{d}\ln D}{\mathrm{d}\ln a}\,\sigma_8\, D \overset{(2)}{=} \frac{\mathrm{d}(\sigma_8 D)}{\mathrm{d}\ln a} \overset{(3)}{=} \frac{\mathrm{d}\sigma_8(z)}{\mathrm{d}\ln a}.$$

(1) The definition of f in (11.38), times $\sigma_8(z) = \sigma_8 D$.
(2) $\mathrm{d}\ln D/\mathrm{d}\ln a = (1/D)\,\mathrm{d}D/\mathrm{d}\ln a$; the D cancels and σ_8 is a constant.
(3) The definition of $\sigma_8(z)$.

So $f\sigma_8$ is the rate at which the amplitude of structure grows per e-fold of expansion. In general relativity with ΛCDM, f follows from the expansion history alone (close to a power of $\Omega_m(z)$; see the cosmology notes). A theory of gravity that changes how fast structure grows, while leaving the expansion the same, changes $f\sigma_8$ and not the acoustic distances. That is why the two features of Figure 11.1 are complementary tests.

Example: what a 10% bias error would do. Suppose we knew $\sigma_8(z)$ perfectly from the CMB and tried to extract f from a measured β using a bias taken from somewhere else. With $f = \beta b$, a 10% error in b is a 10% error in f. Measuring $f\sigma_8$ directly from (11.44) avoids that external bias entirely; it uses the μ^4 term, which contains no bias at all. The price is that the quoted number is a product, and comparing it with a model needs the model's $\sigma_8(z)$ as well as its $f(z)$.

11.5.5 Fingers of God

Inside a galaxy cluster, the linear picture fails. The galaxies are not falling in; they orbit in a virialised halo with random line-of-sight velocities of several hundred to a thousand kilometres per second. The mapping (11.36) spreads such a cluster along the line of sight by $\sigma_v/(aH)$, up to $10\,h^{-1}$Mpc, into a long thin streak pointing at the observer, the **finger of God** (Figure 11.16, the red core). On large scales these streaks do not move the centre of any structure, but they smear the pattern along the line of sight and lower the power of modes with μ near ±1 at high k. They act against the Kaiser boost and grow in importance as k increases.

A model from first principles, in its simplest form. Give every galaxy, in addition to its coherent shift, an independent random line-of-sight shift Δ, Gaussian with mean zero and standard deviation σ (in comoving distance, $\sigma = \sigma_v/aH$). The power spectrum is the Fourier transform of the distribution of pair separations, and the shifts change every pair separation along $\hat{\boldsymbol{z}}$ by $\Delta_1 - \Delta_2$. A pair of distinct galaxies therefore contributes $e^{-i\boldsymbol{k}\cdot(\boldsymbol{r}_{12}+(\Delta_1-\Delta_2)\hat{\boldsymbol{z}})}$ instead of $e^{-i\boldsymbol{k}\cdot\boldsymbol{r}_{12}}$, and averaging over the shifts,

$$\mathbb{E}\left[e^{-ik_z(\Delta_1-\Delta_2)}\right] \overset{(1)}{=} e^{-\frac{1}{2}k_z^2(2\sigma^2)} \overset{(2)}{=} e^{-k^2\mu^2\sigma^2}, \qquad \text{so} \qquad P_s^{\mathrm{obs}}(k,\mu) = \left(b + f\mu^2\right)^2 P(k)\, e^{-k^2\mu^2\sigma^2} + \frac{1}{\bar{n}}. \tag{11.45}$$

(1) $\Delta_1 - \Delta_2$ is the difference of two independent Gaussians, hence Gaussian with variance $2\sigma^2$, and its characteristic function at k_z is $e^{-k_z^2\cdot 2\sigma^2/2}$ ((2.50)).
(2) $k_z = k\mu$.

The shot-noise term is untouched: it comes from each galaxy paired with itself, for which $\Delta_1 - \Delta_2 = 0$. The damping $e^{-k^2\mu^2\sigma^2}$ is the square of the single-galaxy factor $e^{-(k\sigma\mu)^2/2}$ used in the halo model [CooraySheth2002 eqs. (163)–(165)]. It is a phenomenological model: real velocity dispersions depend on the halo a galaxy lives in, and the shifts of galaxies in the same halo are not independent. Analyses therefore treat σ as a nuisance parameter and marginalise over it, and cut the fit at a maximum k beyond which no simple model is trusted.

Example: where the fingers matter. For $\sigma_v = 400\,\mathrm{km\,s^{-1}}$ near $z = 0$, $\sigma = 4\,h^{-1}\mathrm{Mpc}$. Along the line of sight the damping $e^{-k^2\sigma^2}$ reaches e^{-1} at $k = 1/\sigma = 0.25\,h\,\mathrm{Mpc}^{-1}$ and is $e^{-0.16} = 0.85$ already at $k = 0.1\,h\,\mathrm{Mpc}^{-1}$. The acoustic wiggles of Section 11.4 sit at 0.05–0.3 $h\,\mathrm{Mpc}^{-1}$; this is one more reason a redshift-space analysis must model the two effects together.

11.5.6 Checking the Kaiser formula on moving particles

To test (11.41) we need a catalogue of particles that carry both a position and a velocity, with the velocities obeying the continuity equation. The lognormal fields of Section 11.1.4 have no velocities. The simplest model that has them is the **Zel'dovich approximation**: every particle starts on a regular grid at its initial (Lagrangian) position $\boldsymbol{q}$ and is moved by a displacement that grows with the linear growth factor,

$$\boldsymbol{x}(\boldsymbol{q}, t) = \boldsymbol{q} + \boldsymbol{\Psi}(\boldsymbol{q}, t), \qquad \boldsymbol{\Psi}(\boldsymbol{q}, t) = D(t)\,\boldsymbol{\Psi}_0(\boldsymbol{q}). \tag{11.46}$$

Which displacement gives the right density? Mass conservation counts the particles twice, exactly as in (11.39): the particles of a small Lagrangian cell d^3q at uniform density $\bar\rho$ land in d^3x, so $1 + \delta = \mathrm{d}^3q/\mathrm{d}^3x = 1/\det(\mathbb{1} + \partial\boldsymbol{\Psi}/\partial\boldsymbol{q}) \approx 1 - \nabla\cdot\boldsymbol{\Psi}$ to first order. We want this to be the linear field δ, so $\nabla\cdot\boldsymbol{\Psi} = -\delta$, and a curl-free $\boldsymbol{\Psi}$ with that divergence is, by the same steps as (11.40), $\boldsymbol{\Psi}_{\boldsymbol{k}} = i\boldsymbol{k}\,\delta_{\boldsymbol{k}}/k^2$.

Which velocity does it give? Differentiating (11.46),

$$\boldsymbol{u} \overset{(1)}{=} a\,\dot{\boldsymbol{x}} \overset{(2)}{=} a\,\frac{\dot D}{D}\,\boldsymbol{\Psi} \overset{(3)}{=} a\,H\,\frac{\mathrm{d}\ln D}{\mathrm{d}\ln a}\,\boldsymbol{\Psi} \overset{(4)}{=} aHf\,\boldsymbol{\Psi}, \qquad \text{so} \qquad v_z = \frac{u_z}{aH} = f\,\Psi_z.$$

(1) The definition of the peculiar velocity; $\boldsymbol{q}$ does not change.
(2) $\boldsymbol{\Psi} = D\boldsymbol{\Psi}_0$ and $\boldsymbol{\Psi}_0$ is fixed, so $\dot{\boldsymbol{\Psi}} = \dot D\,\boldsymbol{\Psi}_0 = (\dot D/D)\boldsymbol{\Psi}$.
(3) $\dot D/D = \mathrm{d}\ln D/\mathrm{d}t = (\mathrm{d}\ln D/\mathrm{d}\ln a)(\mathrm{d}\ln a/\mathrm{d}t)$ and $\mathrm{d}\ln a/\mathrm{d}t = \dot a/a = H$.
(4) The definition of f.

As a check, $\nabla\cdot\boldsymbol{u}/(aH) = f\nabla\cdot\boldsymbol{\Psi} = -f\delta$, the continuity equation (11.38). So in redshift space each particle sits at $\boldsymbol{s} = \boldsymbol{q} + \boldsymbol{\Psi} + f\Psi_z\hat{\boldsymbol{z}}$: one extra line of code.

Listing 11.7: Zel'dovich particles in real and redshift space (`code/ch11/06_rsd_kaiser.py`)

```
for s in range(NSIM):
    dk = np.fft.rfftn(rng.standard_normal((N, N, N))) * amp          # linear delta_k on the grid
    psi = np.stack([np.fft.irfftn(1j * c * inv_k2 * dk, s=(N, N, N), axes=(0, 1, 2)).ravel() for c in comps], axis=1)
    psi_rms.append(np.sqrt(np.mean(psi ** 2)))
    x = (q + psi) % L
    sz = q + psi
    sz[:, 2] += f * psi[:, 2]                                        # s = x + (u_z/aH) z_hat, u = aHf Psi
    sz %= L
    pr, ps = power(cic(x)), power(cic(sz))
    mult_r.append(kb.multipoles(pr))
    mult_s.append(kb.multipoles(ps))
    mult_k.append(kb.multipoles((1 + f * mu2_grid) ** 2 * pr))    # Kaiser applied to this box's own modes
    sel = (kmag > 0) & (kmag < 0.08)
    ratio_mu.append(np.sqrt(ps[sel] / pr[sel]))                      # should be 1 + f mu^2 mode by mode
    if mus is None:
        mus = np.abs(kb.mu.reshape(kmag.shape)[sel])
    if s == 0:
        x0, s0 = x, sz
```

The script draws 24 Gaussian linear fields at $z = 1$ (growth factor $D = 0.607$, growth rate $f = 0.877$ for the Planck 2018 model) in boxes of side 1000 h^{-1}Mpc with 128^3 particles. Line 74 draws δ_k with the recipe of Section 7.3.3; line 75 is $\boldsymbol{\Psi}_k = i\boldsymbol{k}\,\delta_k/k^2$, transformed back to the grid; lines 77–80 place the particles in real space and in redshift space, the only difference being the term $f\Psi_z$ in line 79. Both catalogues are assigned to the mesh with cloud-in-cell weights, and `power` divides out the cloud-in-cell window. Line 86 keeps, for every mode with $k < 0.08\,h\,\mathrm{Mpc}^{-1}$, the square root of the ratio of its redshift-space to its real-space power, which by (11.41) with $b = 1$ should be $1 + f\mu^2$.

Taking the ratio mode by mode is what lets a few dozen boxes suffice. Both powers come from the same δ_k, so the random amplitude of each mode, the cosmic variance of Section 7.5.2, cancels in the ratio, and what remains scatters only through the non-linear terms of the Zel'dovich mapping. A real survey has no real-space catalogue to divide by; this is a verification of the formula, not a way to measure f.

Estimating f. With $r_j = \sqrt{P_s/P_r}$ for mode j and the model $r_j - 1 = f\mu_j^2$, least squares minimises $\sum_j (r_j - 1 - f\mu_j^2)^2$. Setting the derivative in f to zero, $-2\sum_j \mu_j^2 (r_j - 1 - f\mu_j^2) = 0$, gives $\hat{f} = \sum_j \mu_j^2 (r_j - 1)/\sum_j \mu_j^4$ (line 98 of the script).

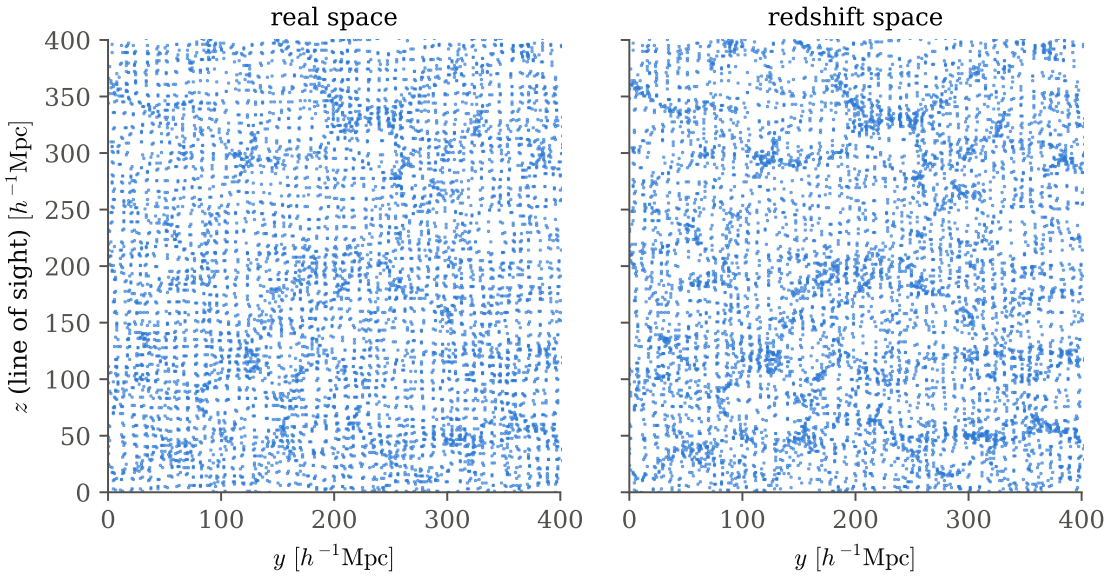


Figure 11.17: A slab 20 h^{-1}Mpc thick from one Zel'dovich box at $z = 1$, in real space (left) and redshift space (right), with the line of sight vertical. The faint grid pattern is the starting lattice, not yet erased by the small displacements. Every particle on the right is moved vertically only, by $f\Psi_z$. By eye the difference is slight at this amplitude; the power spectra of Figure 11.18 measure it.

What the boxes show. The left panel of Figure 11.18 follows $(1 + f\mu^2)^2$ from $\mu = 0$, where the two powers agree, to $\mu = 1$, where the redshift-space power is more than three times larger. The least-squares fit gives $\hat{f} = 0.864$ against the input $f = 0.877$. The scatter of $\hat{f}$ between boxes is 0.014, so the mean of 24 boxes is uncertain by 0.003; the difference $\hat{f} - f = -0.013$ is -4.6 times that error, too large to be chance. What could make the boost slightly too small? Not the linear physics, which the derivation covers exactly, but the part of the displacements that is not coherent on large scales. The rms Zel'dovich displacement per axis is $\sigma_\Psi = 3.4\,h^{-1}$Mpc, so the redshift-space shift $f\Psi_z$ has an rms of 3.0 h^{-1}Mpc, and its small-scale part smears the field along the line of sight much like the random shifts of (11.45): a finger of God in miniature. If that is the cause, the deficit must grow with k and hit the quadrupole hardest, since the

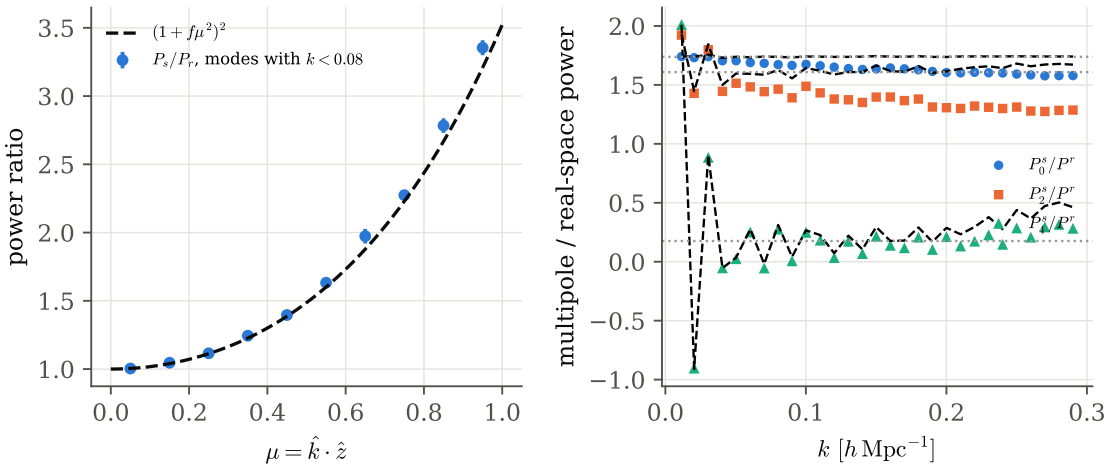


Figure 11.18: Left: the ratio of redshift-space to real-space power for modes with $k < 0.08\,h\,\mathrm{Mpc}^{-1}$, averaged in bins of μ, against the Kaiser factor $(1 + f\mu^2)^2$ (dashed). Right: the monopole (blue), quadrupole (orange) and hexadecapole (green) of the redshift-space power, each divided by the real-space power. Dotted grey lines are the continuum Kaiser values $1 + \frac{2}{3}f + \frac{1}{5}f^2$, $\frac{4}{3}f + \frac{4}{7}f^2$ and $\frac{8}{35}f^2$; the black dashed curves apply $(1 + f\mu^2)^2$ to the boxes' own real-space modes and then take the multipoles with the same discrete average.

damping acts at μ near ±1. At $k = 0.2\,h\,\mathrm{Mpc}^{-1}$ the exponent $(kf\sigma_\Psi)^2$ is already 0.35, and the quadrupole in the right panel falls below the Kaiser value there. At the small k of the fit the effect is small but visible.

For the multipoles averaged over $k < 0.06\,h\,\mathrm{Mpc}^{-1}$, the monopole ratio is 1.724, against 1.738 from (11.42) and 1.741 for the Kaiser factor applied to the same discrete modes. The quadrupole ratio is 1.622, against 1.664 for the same discrete modes, but 1.608 for the continuum formula. The measured quadrupole falls 2.5% below the discrete-mode prediction, the same small shortfall from the smearing. The two predictions differ by 3%, and how much depends on the boxes drawn, since the few modes at low k carry random amplitudes. So a measured quadrupole can land near the continuum value or far from it by chance. The gap between the two predictions is not physics but counting, and it is a trap.

Watch out: Multipoles of a few modes

The continuum results (11.42) assume that the directions in a k shell are spread uniformly in μ. At low k in a box, a shell holds a few dozen modes, and $\mu^2 = k_z^2/k^2$ takes only a handful of values set by the integer grid (for the smallest shell, values such as 0, $\frac{1}{5}$, $\frac{1}{3}$, $\frac{1}{2}$, $\frac{4}{5}$, 1). Then $\langle\mu^4\rangle$ and the Legendre averages are not $\frac{1}{5}$, $\frac{1}{3}$, ..., and even a perfect Kaiser field gives a quadrupole far from $\frac{4}{3}f + \frac{4}{7}f^2$ in the first bins (the jagged dashed curves in Figure 11.18). Always compare a measured multipole with the model evaluated on the same modes, or with a model convolved with the survey's window, which does the same thing for a survey (Section 11.3).

11.5.7 Comparing with the data: the growth rate from BOSS

Their question. By 2016 the Baryon Oscillation Spectroscopic Survey had measured redshifts for 1.2 million massive galaxies over 9329 deg^2 [Alam2017 abstract]. Besides the acoustic peak, the collaboration asked whether structure at $z \approx 0.4$–0.6 had grown at the rate predicted by the ΛCDM model that fits the CMB, a test of gravity on scales of tens of megaparsecs.

Their setup. The galaxies were divided into three overlapping redshift slices with effective redshifts 0.38, 0.51 and 0.61. Four teams fitted the full anisotropic clustering: the monopole and quadrupole of ξ between 25 and 150 h^{-1}Mpc; three μ-wedges of ξ; the power-spectrum multipoles P_0, P_2 up to $k = 0.15\,h\,\mathrm{Mpc}^{-1}$ and P_4 up to 0.1; and power-spectrum wedges up to $0.2\,h\,\mathrm{Mpc}^{-1}$ [Alam2017 §7]. Each used a perturbation-theory model of the non-linear redshift-space clustering, which reduces to (11.41) on large scales and contains a velocity damping in the spirit of (11.45), with the bias, the nuisance parameters and the two dilations of Section 11.4.1 free, and a covariance from one to two thousand mock catalogues. The four results were combined into one consensus value of $f\sigma_8$ per slice. The quantity compared with theory is exactly the $f\sigma_8$ of (11.44), because, as we derived, nothing else in the redshift-space spectrum can be separated from the bias.

What we compute. The prediction is $f(z)\,\sigma_8\,D(z)$ for the Planck 2018 model, with D and f from the linear growth equation of the cosmology notes (`growth` in `lib_lss.py`) and σ_8 from CAMB. The data are the consensus values with statistical and systematic errors added in quadrature.

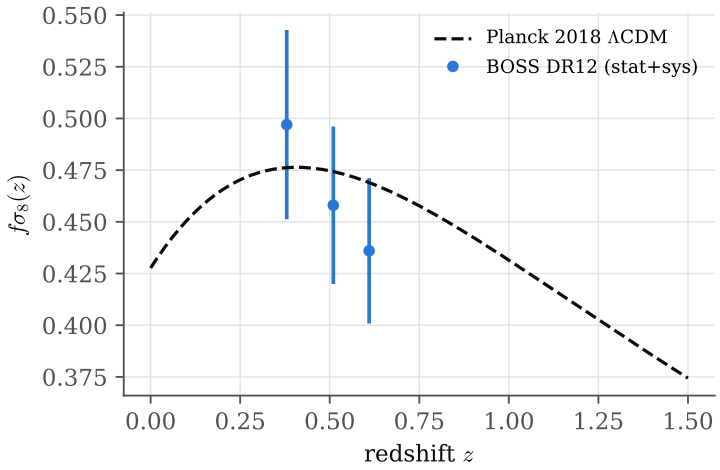


Figure 11.19: The growth rate $f\sigma_8(z)$ of the Planck 2018 ΛCDM model (dashed) and the BOSS DR12 consensus measurements in three redshift slices (points, statistical and systematic errors combined; [224], Table 7).

z_{eff}	BOSS DR12 $f\sigma_8$ (stat, sys)	Planck ΛCDM	pull
0.38	$0.497 \pm 0.039 \pm 0.024$	0.476	0.5
0.51	$0.458 \pm 0.035 \pm 0.015$	0.474	−0.4
0.61	$0.436 \pm 0.034 \pm 0.009$	0.469	−0.9

Side by side. All three points lie within one standard deviation of the prediction (Figure 11.19). The pulls are not three independent draws: the slices overlap in redshift, so neighbouring measurements share galaxies and are positively correlated, and a full test would use the published covariance between them rather than adding three squared pulls. Even so, at 8–9% precision per slice, growth at $z \approx 0.5$ is what ΛCDM with the Planck parameters predicts.

The lesson. Of the three quantities in the Kaiser formula, the survey can deliver only the product $f\sigma_8$; once that is understood, a direction-dependent clustering measurement becomes a test of how fast gravity builds structure.

11.6 How surveys do it, and what our shortcuts cost

Table 1 of the DESI first-year cosmology paper has one line per galaxy sample: a redshift range, a number of objects, an effective volume and a distance ratio such as $D_V/r_d = 7.93 \pm 0.15$ [DESI2024VI Table 1]. Behind each line sits the chain of ideas developed above: points treated as a Poisson sample of a field (Section 11.1), clustering measured against a random catalogue (Section 11.2) or through a weighted Fourier field (Section 11.3), a template fitted for the dilation α (Section 11.4), and the line-of-sight squashing of Section 11.5. Our scripts ran every one of those steps on a laptop. The survey runs each of them at full scale, on data with real defects, with an error budget that has to survive referees who did not take part. The question, in words:

> *At each step from the catalogue to a published distance or growth rate, what does a survey do that our scaled-down versions left out, and what does leaving it out cost?*

11.6.1 The pipeline of a modern survey

Figure 11.20 shows the steps. We take them in order.

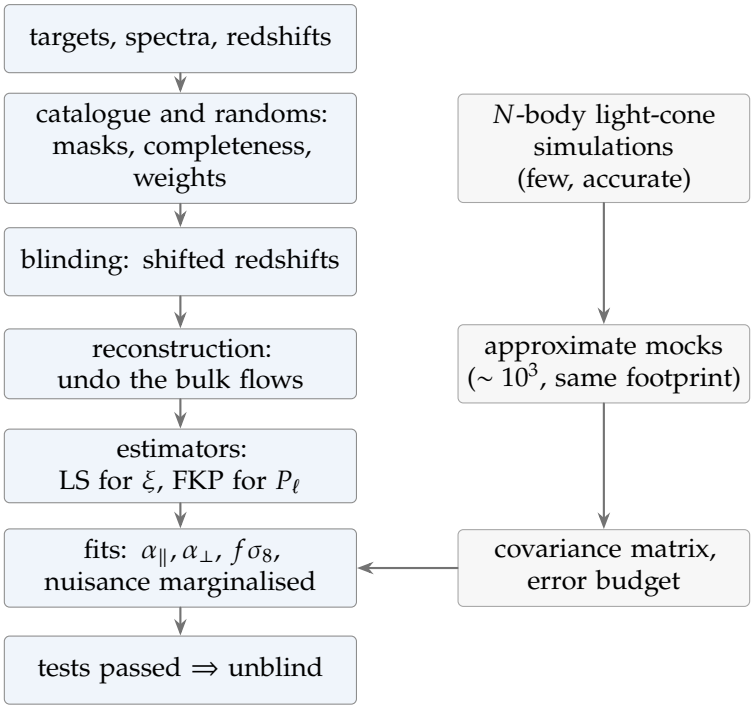


Figure 11.20: The steps from a spectroscopic survey to a distance or growth rate (left column) and the simulations that support them (right). Approximate mock catalogues, calibrated on a few accurate N-body simulations, are run through the same pipeline as the data; their scatter becomes the covariance matrix used in the fits.

The catalogue and its randoms. A real footprint is not a square with a few holes. In BOSS the veto masks, for bright stars, poor imaging, plate centre-posts and similar defects, removed 6.6% and 9.3% of the area in the two galactic caps, leaving a completeness-weighted area of 9329 deg^2; the galaxies carry weights that undo the dependence of the target density on stellar density and seeing, and corrections for fibre collisions and failed redshifts [Alam2017 §2]. In DESI's first year only 35% of the emission-line galaxy targets had received a fibre [DESI2024VI §2.1]. Every one of these effects modulates the observed density without any cosmological cause. The random

catalogue of Section 11.2 must carry every one of them, and the pitfall there (the randoms must have the survey's selection, exactly) is the reason survey teams spend more effort on the randoms than on the estimator. The FKP weights of Section 11.3 multiply these systematic weights.

Blinding. *The problem.* An analysis has hundreds of choices: which weights, which range of scales, which model of the broad band. Each choice is defended by tests, and tests are run until they pass. If the analysts can see the cosmological answer while they work, there is a mechanism that biases the result without anyone intending it. The DESI paper describes it plainly: a result that does not match expectation is attributed to an unknown systematic, and the search for systematics stops when the result agrees [DESI2024VI §2.3.1].

Why that is a statistical bias. Think of the analysis as drawing an estimate $\hat{\theta}$ from its sampling distribution. Suppose an analyst looks at $\hat{\theta}$, keeps it if it lies within 2σ of the expected value, and otherwise searches for, finds, and corrects some small effect, which redraws part of the noise. The rule for when to stop depends on the value obtained. The published estimates are then a sample from the sampling distribution *conditioned on agreeing with expectation*: they are pulled towards the expected value, they scatter less than the quoted error says, and a genuine discrepancy is the one result the procedure is built to remove. It is the same mechanism as stopping an experiment when the p-value first drops below 0.05, run in the opposite direction.

What DESI did. Before analysis the redshifts in the galaxy and quasar catalogues were shifted, by an amount unknown to the analysts, in a way that moves the acoustic feature and changes the redshift-space distortion signal, and the weights were altered to hide the large-scale signal of primordial non-Gaussianity. All tests were run on the blinded catalogues; only after the pipeline was frozen and a prescribed list of validation tests had passed were the true redshifts used. Two changes made after unblinding were reported with their size: a correction to the BAO model, which moved the results by less than 0.3σ, and a completeness fix in the catalogues, by at most 0.6σ [DESI2024VI §2.3.1]. BOSS blinded the methods rather than the data: the four full-shape analyses of Section 11.5.7 were first run on high-fidelity mocks whose parameters the teams did not know, and their errors there entered the systematic budget [Alam2017 §7].

Reconstruction. The acoustic peak is blurred because galaxies have moved a few megaparsecs since the drag epoch, the damping $\Sigma_{\rm nl}$ of Section 11.4. The displacements are not random: on large scales they are the Zel'dovich field $\boldsymbol{\Psi}_{\boldsymbol{k}} = i\boldsymbol{k}\,\delta_{\boldsymbol{k}}/k^2$ of Section 11.5.6, and the observed galaxy field gives an estimate of δ. **Reconstruction** computes that $\boldsymbol{\Psi}$ from the smoothed galaxy field (divided by the bias, and with the redshift-space term $f\Psi_z$ removed), and moves every galaxy and random back by it [DESI2024VI §2.3]. The peak sharpens, roughly halving $\Sigma_{\rm nl}$ from 8 to $4\,h^{-1}$Mpc [Anderson2012 §5.2], and the small non-linear shift of the peak position, below 0.3%, is largely undone. In the Fisher forecast of Section 11.4 for DESI's bright galaxies this was the difference between 2.2% and 1.1% on α: halving the error is worth a factor four in survey volume, so reconstruction is worth about as much as quadrupling the survey, at the cost of one more step to validate on mocks.

Estimators and fits. The estimators are the ones we derived: Landy–Szalay pair counts against random catalogues many times larger than the data ([42] used sixteen times as many randoms), and FKP-weighted Fourier fields, now with fast-Fourier estimators for the multipoles P_0, P_2, P_4 [Alam2017 §7]. The BAO fit uses a template with two dilations, $\alpha_\parallel$ and $\alpha_\perp$, and nuisance parameters for the broad band that are marginalised over [DESI2024VI §2.3]; it is Section 11.4.2 with one more dilation and the Alcock–Paczyński information of Section 11.4.1 kept. The

theoretical systematic errors of this procedure are estimated at 0.1% for the isotropic and 0.2% for the anisotropic BAO parameters, with up to 0.2% more from how galaxies populate haloes; they are added in quadrature to the statistical error [DESI2024VI §2.3].

Mocks for the covariance. Every fit needs the covariance matrix of the measured ξ or P_ℓ. The Gaussian formula of Section 11.3 is not trusted for this: in Section 11.3.3 the scatter of the lognormal field already exceeded it by a factor 1.60 at $k = 0.10\,h\,\mathrm{Mpc}^{-1}$. Surveys instead run a thousand or more mock catalogues through the full pipeline, each with the survey's footprint, completeness, fibre collisions and $n(z)$. Full N-body simulations are too expensive at that number, so the mocks use approximate dynamics: Eisenstein et al. used 1278 PTHalos mocks, and BOSS DR12 used 1000 MultiDark-Patchy mocks (second-order Lagrangian perturbation theory plus a stochastic model for haloes) and 1000 QPM mocks (a low-resolution particle-mesh code), both populated with galaxies to match the observed clustering [Alam2017 §4]. DESI's first-year BAO analysis used semi-analytic covariance matrices validated against such mocks [DESI2024VI §2.3]. A few accurate N-body light-cone simulations serve as the reference that the approximate mocks and the models are checked against; one public set provides 108 full-sky realisations with haloes resolved down to the hosts of the BOSS CMASS and luminous red galaxies [51].

The cost of a finite number of mocks is the subject of Section 8.13.2. With N_S mocks and N_D data bins, the inverse of the sample covariance must be multiplied by the Hartlap factor $(N_S - N_D - 2)/(N_S - 1)$ (Eq. (3.43)), and even then the variance of each parameter is itself uncertain by a fraction $\varepsilon = \sqrt{2/(N_S - N_D - 4)}$ (Section 8.13.4). Since the error bar is the square root of the variance, its fractional scatter is $\varepsilon/2$ by the delta method (Section 3.6.3).

Example: our acoustic-peak mocks against the 2005 analysis. The covariance in Section 11.4.4 came from 1500 mocks for 16 bins: Hartlap factor 0.989 and $\varepsilon = 3.7\%$, so each error bar is uncertain by 1.8%. The 2005 analysis, 1278 mocks for 20 bins, had 0.984 and 2.0%. On this count our nerfed run is as good as the original, which is a statement about bins: a data vector of a few tens of bins is cheap.

Example: a joint fit of many multipoles. A modern full-shape fit uses P_0, P_2, P_4 in several redshift bins and both galactic caps, easily a hundred numbers. With $N_D = 100$ and 1000 mocks the Hartlap factor is 0.899: a 10% correction to every χ^2. For an error bar good to 5%, Eq. (8.79) asks for 304 mocks with $N_D = 100$ against 220 with $N_D = 16$. The mocks are cheap compared with their validation; what limits a survey is not their number but how well their clustering matches the real galaxies'.

Designing the survey. The same Fisher forecast that we checked against mocks in Section 11.4 is how surveys are planned: one writes the error on the BAO distances as a function of area, depth, redshift range and target density, and optimises it under the constraints of the instrument, the field of view, the number of fibres and the total observing time [HobsonBMC §5.3.3]. The effective volume V_{eff} of Section 11.3 is the quantity being traded: more area at fixed time means fewer galaxies per unit volume and a smaller $\bar{n}P$.

11.6.2 What our shortcuts cost

Every script ran in minutes on a laptop. Table 11.1 lists, step by step, what the research version does instead and what the simplification costs. Two kinds of cost appear. Some only widen error bars or shrink the range of scales and could be removed with more computing time; the

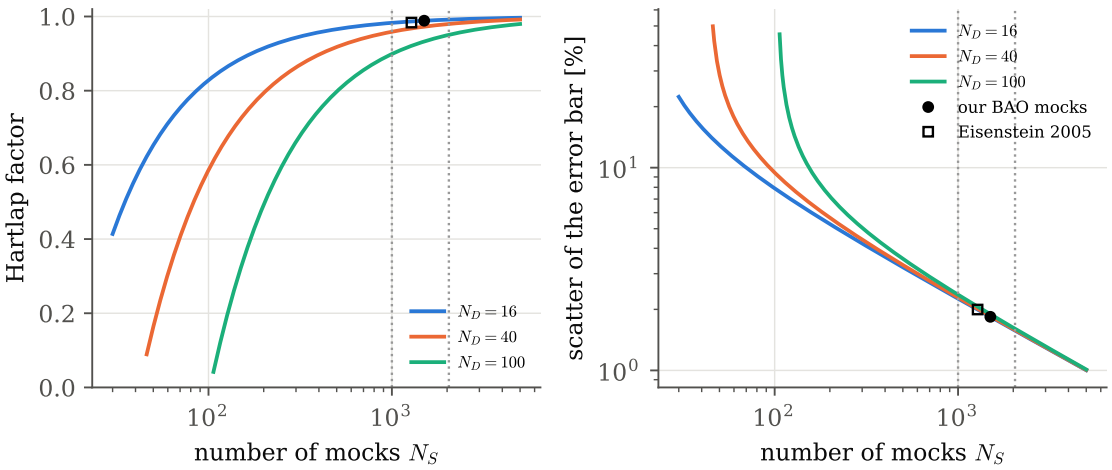


Figure 11.21: The price of a covariance matrix estimated from N_S mocks, for data vectors of $N_D = 16$, 40 and 100 bins. Left: the Hartlap factor that de-biases the inverse covariance. Right: the fractional scatter of a parameter's error bar, $\varepsilon/2$. Black points: our acoustic-peak mocks (filled) and the 2005 analysis (open). Dotted lines: the BOSS DR12 mock sets of 1000 and 2045.

cluster versions of the scripts would do so. Others leave a physical effect out altogether, and no amount of computing in the same model brings it back.

step	research	ours	what it costs
density field	N-body or 2LPT dynamics, galaxies placed in haloes	Gaussian and lognormal fields; Zel'dovich particles	covariance only roughly right (scatter 1.60× Gaussian in the FKP test); no realistic three-point function
geometry	light cone, true footprint and $n(z)$, curved sky	masked 2-D square for ξ; sphere with $\bar{n}(r)$ for FKP; periodic boxes for BAO and RSD	no window in the BAO fit; distant observer only; no evolution with redshift inside a sample
mocks	10^3–2×10^3 approximate mocks, a few N-body	1500 (BAO), 300 (ξ), 100 (FKP), 24 (RSD)	error bars uncertain by 1.8% for BAO; RSD only a verification
reconstruction	yes	no	σ_α forecast 2.2% instead of 1.1% for DESI's bright galaxies
BAO fit	$\alpha_\parallel$, $\alpha_\perp$, broad band marginalised	isotropic α	no separate D_M and D_H, no Alcock–Paczyński information
RSD model	perturbation theory with nuisance parameters, $k \lesssim 0.15$–0.2	linear Kaiser on $k < 0.08\,h\,\mathrm{Mpc}^{-1}$	$\hat{f} - f = -0.013$ (−4.6 errors), from the smearing; fewer modes
systematics	weights for stars, seeing, fibre collisions; blinding	none	nothing in simulations, which have no systematics; fatal on real data

Table 11.1: Each step of the analysis at research scale and in the scripts, with the cost of the simplification.

The first row is the one that matters most and the one we can do least about on a laptop. A lognormal field has the right one- and two-point statistics by construction and the wrong higher-order ones. Every statement that depends only on the two-point function, the shot

noise, the estimators' first-order behaviour, the FKP weight, the Kaiser factor, transferred correctly from our simulations to the theory. Every statement that depends on the four-point function, the absolute covariance and therefore the absolute error bars, transferred only roughly. That is precisely why surveys build mocks with realistic dynamics and why the cosmic web's higher-order structure, which the two-point function does not see, carries information of its own, which the halo model, voids and the topological summaries of Chapter 12 set out to extract.

The lesson. A survey measurement is the same short chain of estimators we derived, wrapped in a much longer chain of corrections, simulations and checks; the estimators fix what is measured, and the mocks and the blinding decide whether the quoted error can be believed.

Recap

- Galaxies are modelled as a Cox process: Poisson given a smooth field $\bar{n}(1+\delta_g)$. Counts in cells have variance $\langle N\rangle + \langle N\rangle^2\sigma_R^2$, pairs have density $\bar{n}^2(1+\xi)$, and the power of the points is $P(k) + 1/\bar{n}$ (Section 11.1).
- $\xi(r)$ is estimated by comparing data pairs with pairs of a random catalogue that has the survey's selection. Written to first order in the fluctuations, Landy–Szalay $(DD - 2DR + RR)/RR$ cancels the edge terms that the natural and Davis–Peebles estimators keep; every estimator carries the integral constraint (Section 11.2).
- The FKP estimator weights the data-minus-randoms field by $w = 1/(1+\bar{n}P)$, the weight that minimises the variance of $\hat{P}$; its mean is the window-convolved spectrum plus shot noise, and its information is measured by $V_{\rm eff}$ (Section 11.3).
- The acoustic peak is a ruler of length r_d; fitting a template with dilation α measures D_V/r_d, the detection significance is $\sqrt{\Delta\chi^2}$ between peaked and smooth templates, and the Fisher forecast on α is a lower bound reached only at high signal-to-noise (Section 11.4).
- Peculiar velocities move galaxies along the line of sight by $u_r/(aH)$. Number conservation and the linear continuity equation give $P_s(k,\mu) = (b + f\mu^2)^2 P(k)$, with multipoles P_0, P_2, P_4 only; clustering alone measures $b\sigma_8$ and $f\sigma_8$, and $f\sigma_8$ is the growth rate of the amplitude of structure (Section 11.5).
- Surveys add masks, weights, blinding, reconstruction and a thousand or more mocks for the covariance; our laptop versions keep the estimators and lose the realistic covariance (Section 11.6).

11.7 Two universes with one power spectrum

In Section 7.7 we took a map of thin rings, kept the amplitude $|\delta_k|$ of every Fourier mode, replaced every phase by a random number, and transformed back. The power spectrum did not change at all, and the rings became featureless noise. That map was a toy. The matter in the real Universe is not a toy: gravity has spent thirteen billion years pulling it into compact clumps, stretching it into filaments and emptying the space between them. If we play the same trick on an evolved matter field, how much of what gravity built does the power spectrum keep?

The answer decides how much of the rest of the chapter can be done with $\xi(r)$ and $P(k)$ alone. The BAO and RSD measurements of Sections 11.4 and 11.5 use the two-point function on large scales, where the field is nearly Gaussian and the power spectrum is, by Section 7.2, a complete description. On the scales of galaxies and clusters that is no longer true. The power

spectrum answers exactly one question, how strongly the fluctuation at one place is tied to the fluctuation at a place a distance r away, and whatever else the field holds is invisible to it. The question becomes concrete:

> *Two matter fields have exactly the same power spectrum. In what measurable ways can they still differ, and which physical structures carry the difference?*

11.7.1 The experiment

We need an evolved field we can make on a laptop. The Zel'dovich approximation of (11.46) is the simplest: put one particle at the centre of every cell of a regular grid (its Lagrangian position $\boldsymbol{q}$) and move it by the displacement $\boldsymbol{\Psi}(\boldsymbol{q})$ that linear theory predicts for today, $\boldsymbol{\Psi}_{\boldsymbol{k}} = i\boldsymbol{k}\,\delta_{\boldsymbol{k}}/k^2$, with $\delta_{\boldsymbol{k}}$ drawn from the CAMB linear spectrum at $z = 0$. The script `08_same_spectrum.py` does this for 192^3 particles in a periodic box of side $600\,h^{-1}$Mpc and deposits them on the grid with cloud-in-cell weights, which gives the density $1 + \delta(\boldsymbol{x})$ of the *Zel'dovich universe*. Particles that started in a slightly overdense patch converge, those in an underdense patch diverge, and the result already shows sheets, filaments and empty regions (Figure 11.22, left).

The *twin* is built from it exactly as in Section 7.7. Fourier transform δ, keep $|\delta_{\boldsymbol{k}}|$, take the phases from the transform of fresh white noise (which makes them uniformly random and keeps the reality condition), and transform back. The ratio of the two power spectra, mode by mode, differs from one by at most 2.519×10^{-12}, which is the rounding error of the FFT. Any analysis that measures only $P(k)$ would call the two fields the same universe. Figure 11.22 shows that they are not.

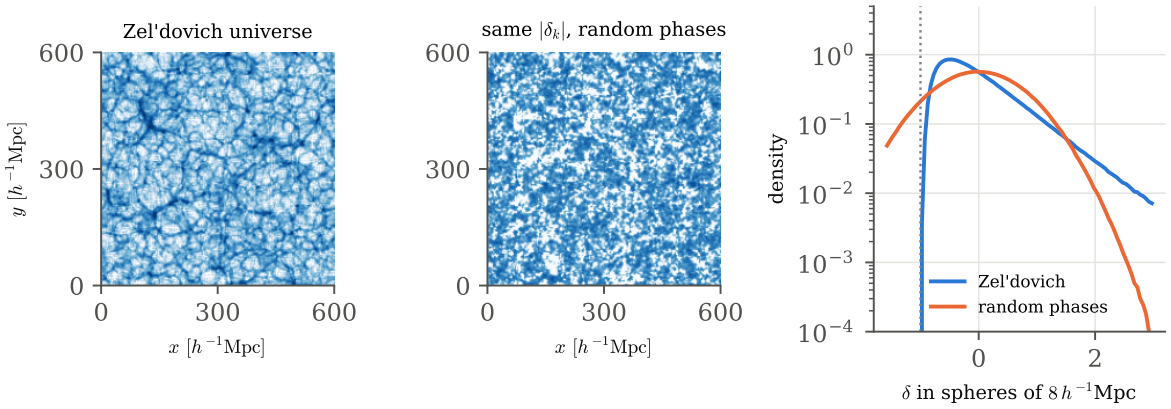


Figure 11.22: Left: a slice $9\,h^{-1}$Mpc thick through the Zel'dovich universe, darker where denser (logarithmic scale). Middle: the same slice of its twin, which has the same $|\delta_{\boldsymbol{k}}|$ for every mode and random phases. Right: the distribution of the density contrast averaged in spheres of radius $8\,h^{-1}$Mpc. The Zel'dovich field is bounded below by $\delta = -1$ (dotted) and has a long tail to high densities; the twin is symmetric and spills below -1.

11.7.2 What the two fields share, and what they do not

Example: the variance in spheres is the same. Average the density contrast over a sphere of radius R about every point; call the result δ_R. Its variance is the number σ_R^2 we met in the definition of σ_8 (Section T1.4). Write the sphere average as a convolution with the normalised top hat W_R, whose Fourier transform is $W(kR) = 3(\sin kR - kR\cos kR)/(kR)^3$. Then

$$\sigma_R^2 \overset{(1)}{=} \left\langle \left| \int \frac{\mathrm{d}^3k}{(2\pi)^3}\, \delta_{\boldsymbol{k}} W(kR)\, e^{i\boldsymbol{k}\cdot\boldsymbol{x}} \right|^2 \right\rangle \overset{(2)}{=} \int \frac{\mathrm{d}^3k}{(2\pi)^3}\, P(k)\, W(kR)^2 \overset{(3)}{=} \int \frac{\mathrm{d}k}{k}\, \frac{k^3 P(k)}{2\pi^2}\, W(kR)^2. \quad (11.47)$$

(1) A convolution in real space is a product in Fourier space, so δ_R is the inverse transform of $\delta_{\boldsymbol{k}} W(kR)$.
(2) Distinct modes are uncorrelated in a homogeneous field, $\langle \delta_{\boldsymbol{k}} \delta^*_{\boldsymbol{k}'} \rangle = (2\pi)^3 \delta_D(\boldsymbol{k} - \boldsymbol{k}') P(k)$ (Section 7.3.2), so the double integral collapses to one, and the phases $e^{i\boldsymbol{k}\cdot\boldsymbol{x}}$ cancel against their conjugates.
(3) P and W depend only on $k = |\boldsymbol{k}|$, so the angular integral gives $4\pi k^2$, and $4\pi k^2 \mathrm{d}k/(2\pi)^3 = (k^3/2\pi^2)\,\mathrm{d}k/k$.

Only $|\delta_{\boldsymbol{k}}|^2$ enters. The measured values agree: $\sigma_R = 0.7111$ in the Zel'dovich universe and 0.7111 in the twin, for $R = 8\,h^{-1}$Mpc. Every quantity built from $P(k)$, the correlation function, σ_8, the shot-noise-corrected FKP spectrum of Section 11.3, is identical in the two fields.

Example: the skewness is not. The third moment of δ_R involves three modes at once. Write $\delta_{\boldsymbol{k}} = |\delta_{\boldsymbol{k}}| e^{i\varphi_{\boldsymbol{k}}}$; then

$$\begin{aligned} \langle \delta_R^3 \rangle &\overset{(1)}{=} \int \frac{\mathrm{d}^3k_1\,\mathrm{d}^3k_2\,\mathrm{d}^3k_3}{(2\pi)^9} W_1 W_2 W_3 \left\langle \delta_{\boldsymbol{k}_1} \delta_{\boldsymbol{k}_2} \delta_{\boldsymbol{k}_3} \right\rangle e^{i(\boldsymbol{k}_1+\boldsymbol{k}_2+\boldsymbol{k}_3)\cdot\boldsymbol{x}} \\ &\overset{(2)}{=} \int \frac{\mathrm{d}^3k_1\,\mathrm{d}^3k_2}{(2\pi)^6} W_1 W_2 W_3 \left\langle |\delta_{\boldsymbol{k}_1}||\delta_{\boldsymbol{k}_2}||\delta_{\boldsymbol{k}_3}|\, e^{i(\varphi_{\boldsymbol{k}_1}+\varphi_{\boldsymbol{k}_2}+\varphi_{\boldsymbol{k}_3})} \right\rangle_{\boldsymbol{k}_3=-\boldsymbol{k}_1-\boldsymbol{k}_2}, \end{aligned} \tag{11.48}$$

(1) Cube the Fourier representation of δ_R, with the shorthand $W_i = W(k_i R)$.
(2) Homogeneity forces the three wavevectors to close into a triangle, $\boldsymbol{k}_1 + \boldsymbol{k}_2 + \boldsymbol{k}_3 = 0$ (the average of a product of modes is invariant under translations only if the total wavevector vanishes).

In the twin every phase is independent and uniform, so $\langle e^{i(\varphi_1+\varphi_2+\varphi_3)} \rangle = 0$ for every triangle and the third moment vanishes, up to the noise of one realisation. In the Zel'dovich universe the phases of the three sides of a triangle are correlated, because gravity couples modes: a dense clump is a place where many modes are in phase. The measured skewness $\langle \delta_R^3 \rangle / \sigma_R^3$ is 1.883 for the Zel'dovich universe and 0.01497 for the twin. The skewness is the simplest number that the power spectrum cannot predict; its Fourier-space counterpart is the bispectrum, the average over triangles in (11.48).

Example: the twin is not even a possible density. A density cannot be negative, so $\delta \geq -1$ everywhere. In the Zel'dovich universe the fraction of spheres with $\delta_R < -1$ is 0%. In the twin it is 7.992%. The twin's one-point distribution is the symmetric bell of a Gaussian, and a bell with $\sigma = 0.7111$ puts weight below -1. The real field has to fit its whole variance into the range $\delta > -1$, so it piles up just above -1, in the broad, nearly empty regions, and compensates with a long tail of rare dense regions (Figure 11.22, right). Where the mass cannot go below a floor, matter that leaves a region must end up concentrated somewhere else.

Example: the two fields have different voids. The spherical void finder of Section 11.11.2 grows a sphere about every local density minimum until the mean density inside reaches 20% of the cosmic mean, and keeps the non-overlapping spheres larger than $8\,h^{-1}$Mpc. It returns 3607 voids filling 8.482% of the Zel'dovich box, the largest of radius $23.16\,h^{-1}$Mpc, and only 2546 voids filling 4.569% of the twin (with the unphysical negative densities set to zero first), the largest of radius $15.9\,h^{-1}$Mpc. Same power spectrum, half the empty volume.

11.7.3 Where the extra information lives

The four examples point at two places. The positive tail of the distribution is made of compact, dense objects: the clumps in which the particles of the Zel'dovich universe have crossed one another. In the real Universe these are **dark matter haloes**, gravitationally bound and in rough virial equilibrium, and every galaxy lives inside one. The broad pile-up just above $\delta = -1$ is made of **voids**, large regions that matter has streamed out of, which fill most of the volume. Figure 11.23 draws the two along a line through a field.

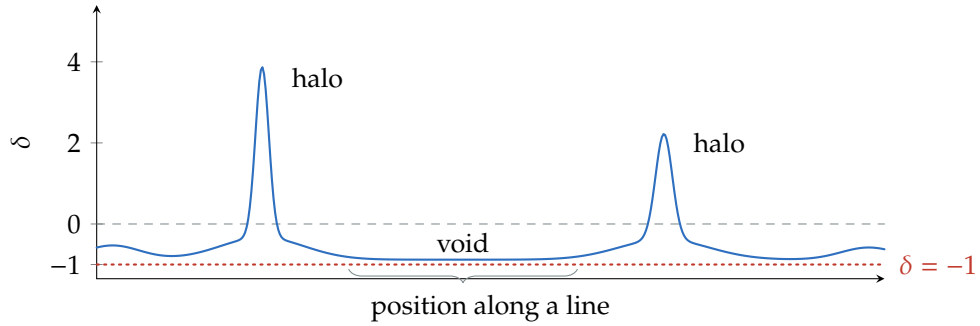


Figure 11.23: The two kinds of structure along a line through an evolved matter field (a drawn cartoon, not data). Haloes are narrow and tall, reaching hundreds of times the mean density; a void is a broad basin whose floor sits just above the hard limit $\delta = -1$ (dotted), and which fills most of the line. A Gaussian field with the same power spectrum would have neither.

This suggests two ways forward, and both pay.

(1) *Build the field out of its structures.* If all matter sits in haloes, the statistics of the field follow from three properties of haloes: how many there are of each mass, how matter is arranged inside each, and how haloes cluster. This is the **halo model**. Its first payoff is a prediction of the non-linear $P(k)$ from linear theory plus these ingredients; its second is a model of how galaxies sit in haloes, which explains the shape of the galaxy correlation function (Sections 11.8 to 11.10).

(2) *Count and measure the structures themselves.* The number of voids of each size, their average density profile and their average shape are summary statistics that, as the last example showed, respond to information the power spectrum discards. They need their own theory of what to expect, and their own estimators (Sections 11.11 and 11.12).

In the language of the pipeline that runs through the book (physical model, field, choice of information, summary, sampling distribution, inference), the halo model is a better *physical model* for the field on small scales, and the void statistics are new *summaries* whose distributions we have to derive. Both use the same mathematical engine, the statistics of random walks, which is where we start.

Key idea: The power spectrum keeps amplitudes and throws away phases

Two fields with identical $|\delta_k|$ have identical $P(k)$, $\xi(r)$ and σ_R, but can differ in their skewness, in their one-point distribution and in the number and size of their voids. For an evolved matter field the difference is carried by haloes, on the dense side, and by voids, on the empty side.

11.8 How many haloes of each mass

Galaxy clusters occupy some of the most massive haloes in the Universe, with masses of 10^{14}–10^{15} solar masses. Even a cube 1 h^{-1}Gpc on a side contains only a few hundred of the largest clusters. Their number is extremely sensitive to the amplitude of the initial fluctuations, because building a $10^{15}M_\odot$ object needs a large region to have been unusually dense, and unusual events in a Gaussian field are exponentially rare. Counting clusters has therefore been a cosmological probe since the 1970s, and the abundance of haloes of every mass is the first ingredient of the halo model. The question, in words:

> *Given the linear density field, Gaussian with a known power spectrum, what fraction of all the mass sits today in haloes with mass between M and $M + \mathrm{d}M$?*

This is a prediction, in the sense of Chapter 5: from a hypothesis about the initial field to the frequency of an outcome. Its answer becomes a likelihood for cluster counts, and an input to everything else in this part.

11.8.1 The inputs

The field. Let $\delta(\boldsymbol{x})$ be the linear density contrast extrapolated to today with the linear growth factor, as CAMB returns it. Smooth it with a sphere of radius R; the sphere holds, at the mean density $\bar\rho$, the mass $M = \frac{4\pi}{3}\bar\rho R^3$, so every smoothing scale is also a mass. The smoothed contrast $\delta_M(\boldsymbol{x})$ has variance $\sigma^2(M)$, given by (11.47) with $R = (3M/4\pi\bar\rho)^{1/3}$. It falls as M grows: in our fiducial cosmology $\sigma = 2.891$ for $M = 10^{11}\,h^{-1}M_\odot$ and $\sigma = 0.5348$ for $M = 10^{15}\,h^{-1}M_\odot$.

Collapse. A spherical region whose linear contrast exceeds a critical value δ_c has, according to the exact solution for a uniform spherical perturbation, stopped expanding, turned around and collapsed by today. The spherical-collapse solution belongs to cosmology and is taken from there (see the cosmology notes); for an Einstein–de Sitter background it gives

$$\delta_c = \frac{3}{5}\left(\frac{3\pi}{2}\right)^{2/3} = 0.6 \times 2.811 = 1.686, \tag{11.49}$$

with a very weak dependence on Ω_m that we ignore [CooraySheth2002 eq. (53)]. The collapsed object settles at a mean density a few hundred times $\bar\rho$; in an Einstein–de Sitter universe $\Delta_{\rm vir} = 18\pi^2 \approx 178$, and for ΛCDM today about 330 relative to the mean matter density, which is the value we use for halo edges in Section 11.9. The important property of (11.49) is that it does not depend on the mass: the collapse of a uniform sphere depends only on its density contrast, not on its size [CooraySheth2002 p. 20].

We now list what we assume, separating physics from mathematics, as for the Poisson process of Section 2.5.1.

Physical inputs.

(H1) The linear field is Gaussian, with the CAMB linear power spectrum.

(H2) A region of mass M around a point has collapsed into a halo by today if $\delta_M > \delta_c$ there. Collapse is spherical, so the threshold is the same constant for all masses.

(H3) A mass element belongs to the most massive collapsed region that contains it. A small clump inside a large collapsed region is part of the large halo, not a halo of its own.

Mathematical choice.

(M1) The smoothing is done with a sharp filter in Fourier space (keep all modes with $k < k_c$, drop the rest) rather than a sphere in real space.

(H1) is the standard inflationary initial condition, well tested by the CMB. (H2) is an idealisation: real collapse is ellipsoidal, and its correction is the subject of Section 11.8.5. (H3) is the rule that turns "collapsed regions" into a count of distinct objects. (M1) has no physics in it; it is chosen because it makes the problem exactly solvable, and Section 11.8.4 measures what it costs.

11.8.2 Press and Schechter's count, and the missing half

The simplest guess. Fix a mass M. By (H1), δ_M at a randomly chosen point is Gaussian with mean zero and variance $\sigma^2(M)$. By (H2), the point sits inside a collapsed region of mass at least M if $\delta_M > \delta_c$. Press and Schechter [225] identified the fraction of mass in haloes heavier than M with the fraction of points where this happens. Choosing a point at random is choosing a mass element at random, because the linear field is almost uniform, so

$$\mathrm{P}(\delta_M > \delta_c) = \int_{\delta_c}^{\infty} \frac{\mathrm{d}\delta}{\sqrt{2\pi}\,\sigma}\, e^{-\delta^2/2\sigma^2} = \frac{1}{2}\,\mathrm{erfc}\Big(\frac{\nu}{\sqrt{2}}\Big), \qquad \nu \equiv \frac{\delta_c}{\sigma(M)}. \tag{11.50}$$

The variable ν is the threshold in units of the standard deviation: a halo of mass M is a "ν-sigma" event. At $\nu = 1$ the fraction is 0.159, at $\nu = 2$ it is 0.023, at $\nu = 3$ it is 0.0013; this is the exponential sensitivity of cluster counts.

What goes wrong. Let $M \to 0$. Then $\sigma \to \infty$, $\nu \to 0$, and (11.50) tends to $\frac{1}{2}$. Half of the mass never belongs to any halo, however small. But every mass element in a Gaussian field is, on some scale, inside an overdense region. Where does the other half go? The points with $\delta_M < \delta_c$ on scale M include points that *are* inside a collapsed region on a larger scale $M' > M$. By (H3) they belong to a halo of mass M', but the count at scale M never sees them. Press and Schechter multiplied their result by 2 by hand. The reason the factor is exactly 2 became clear only when the problem was reformulated as a random walk [226].

11.8.3 The excursion set: a random walk in smoothing scale

Fix one point $\boldsymbol{x}$ and watch its smoothed contrast as the filter shrinks, from infinitely large (where $\delta = 0$, the mean) to infinitely small. With the sharp filter of (M1),

$$\delta(\boldsymbol{x}; k_c) = \int_{|\boldsymbol{k}|<k_c} \frac{\mathrm{d}^3k}{(2\pi)^3}\, \delta_{\boldsymbol{k}}\, e^{i\boldsymbol{k}\cdot\boldsymbol{x}}, \qquad S(k_c) \equiv \mathrm{Var}\big[\delta(\boldsymbol{x}; k_c)\big] = \int_{|\boldsymbol{k}|<k_c} \frac{\mathrm{d}^3k}{(2\pi)^3}\, P(k). \tag{11.51}$$

S grows from 0 as k_c grows, so we may use S itself as the "time" of the trajectory. Large S means a small smoothing scale and a small mass; we write $S = \sigma^2(M)$ for the mass associated with each scale.

How does δ change when the filter shrinks a little? Raising the cut from k_c to $k_c + \Delta k_c$ adds the modes in a thin shell,

$$\Delta\delta \overset{(1)}{=} \int_{k_c<|\boldsymbol{k}|<k_c+\Delta k_c} \frac{\mathrm{d}^3k}{(2\pi)^3}\, \delta_{\boldsymbol{k}} e^{i\boldsymbol{k}\cdot\boldsymbol{x}},$$

$$\big\langle \Delta\delta\, \delta(\boldsymbol{x}; k_c)\big\rangle \overset{(2)}{=} 0, \qquad \big\langle(\Delta\delta)^2\big\rangle \overset{(3)}{=} \int_{\text{shell}} \frac{\mathrm{d}^3k}{(2\pi)^3} P(k) = S(k_c + \Delta k_c) - S(k_c) \equiv \Delta S. \tag{11.52}$$

(1) The difference of two sharp-k integrals is the integral over the shell between them.

(2) The increment contains only modes in the shell, the old value only modes inside it. In a homogeneous field distinct modes are uncorrelated (Section 7.3.2).

(3) The same orthogonality applied to the shell alone.

For a Gaussian field, uncorrelated means independent (Section 7.3.3). So every step of the trajectory is a Gaussian of variance ΔS, independent of everything before it. That is the definition of Brownian motion, with S as time and $\delta = 0$ at $S = 0$. Taking two times $S_1 < S_2$ and writing $\delta(S_2) = \delta(S_1) + [\delta(S_2) - \delta(S_1)]$ gives, from the independence of the bracket, $\langle \delta(S_1)\delta(S_2)\rangle = S_1 = \min(S_1, S_2)$.

Working definition 11.5 (excursion set, first crossing)

At a fixed point, the **excursion-set walk** is the smoothed linear contrast $\delta(S)$ as a function of the variance $S = \sigma^2(M)$ of the filter. With a sharp-k filter it is a Brownian motion starting at $\delta(0) = 0$. The point belongs to a halo of mass M if the walk **first crosses** the threshold δ_c at $S = \sigma^2(M)$: no larger scale had collapsed, so by (H3) the largest collapsed region around the point has mass M. *Operational:* the fraction of mass in haloes heavier than M is the fraction of walks that have crossed δ_c at some $S' < \sigma^2(M)$.

The question in walk language. Out of all walks, what fraction has reached δ_c at least once before time S? Press and Schechter counted only the walks that are *above* δ_c at time S. A walk that went above earlier and came back down by time S is also inside a halo of mass larger than M, and it is missing from their count.

The reflection argument. Split the walks that have crossed by time S into those that end above the barrier and those that end below:

$$\mathrm{P}(\text{crossed by } S) = \mathrm{P}\big(\delta(S) > \delta_c\big) + \mathrm{P}\big(\text{crossed by } S \text{ and } \delta(S) < \delta_c\big). \tag{11.53}$$

Take any walk in the second group. It touched δ_c for the first time at some $S_1 < S$. Reflect the part of the walk after S_1 in the line $\delta = \delta_c$, replacing $\delta(S')$ by $2\delta_c - \delta(S')$ for $S' > S_1$ (Figure 11.24). The reflected walk ends above δ_c. Its steps after S_1 are the original steps with their signs flipped, and a Gaussian step of mean zero is as likely as its negative, so the reflected walk is exactly as probable as the original. The map is one to one, and every walk ending above the barrier is the image of exactly one walk in the second group (reflect it back after its own first crossing). The two groups have equal probability, and

$$\mathrm{P}(\text{crossed by } S) \overset{(1)}{=} 2\,\mathrm{P}\big(\delta(S) > \delta_c\big) \overset{(2)}{=} \operatorname{erfc}\!\Big(\frac{\delta_c}{\sqrt{2S}}\Big). \tag{11.54}$$

(1) The second term of (11.53) equals the first, by the reflection. A walk that ends above the barrier has necessarily crossed it.

(2) $\delta(S)$ is Gaussian with variance S, so the first term is (11.50) with $\sigma^2 = S$.

As $S \to \infty$ the right side tends to $\operatorname{erfc}(0) = 1$: every mass element ends up in some halo. The factor of 2 is not a fudge. It is the contribution of the walks that went up and came back down, the cloud-in-cloud mass that the single-scale count missed.

From the first-crossing distribution to the mass function. The fraction of walks that first cross between S and $S + \mathrm{d}S$ is the derivative of (11.54), $\mathcal{F}(S)\,\mathrm{d}S$. It is cleaner in terms of $\nu = \delta_c/\sqrt{S}$, so we change variable:

$$\mathcal{F}(S) \overset{(1)}{=} \frac{\mathrm{d}}{\mathrm{d}S}\operatorname{erfc}\!\Big(\frac{\delta_c}{\sqrt{2S}}\Big) \overset{(2)}{=} \frac{2}{\sqrt{\pi}}\,e^{-\delta_c^2/2S}\,\frac{\delta_c}{2\sqrt{2}\,S^{3/2}} = \frac{\delta_c}{\sqrt{2\pi}\,S^{3/2}}\,e^{-\delta_c^2/2S},$$

$$F(\nu) \overset{(3)}{=} \mathcal{F}(S)\left|\frac{\mathrm{d}S}{\mathrm{d}\nu}\right| \overset{(4)}{=} \frac{\delta_c\,\nu^3}{\sqrt{2\pi}\,\delta_c^3}\,e^{-\nu^2/2}\,\frac{2\delta_c^2}{\nu^3} = \sqrt{\frac{2}{\pi}}\,e^{-\nu^2/2}. \tag{11.55}$$

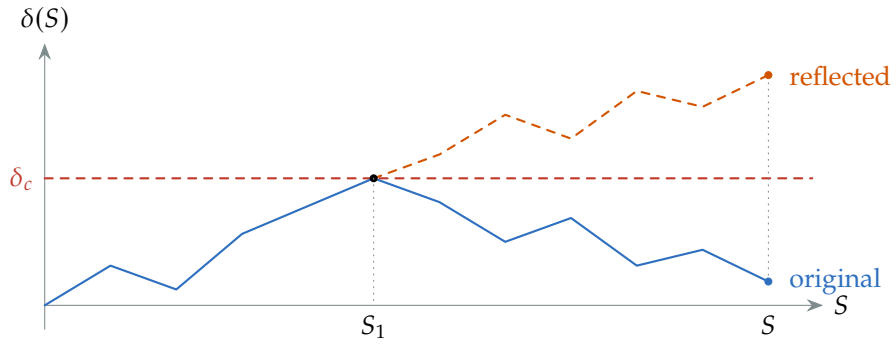


Figure 11.24: The reflection argument. A walk (blue) first reaches δ_c at S_1 and is below it again at S. Reflecting everything after S_1 in the barrier gives a walk (orange) that is equally probable and ends above the barrier. Every walk that crossed and came back is paired with one that ends above, so the walks that have crossed by S are twice those found above δ_c at S.

(1) The first-crossing density is the rate at which the crossed fraction grows.
(2) $\frac{\mathrm{d}}{\mathrm{d}y}\operatorname{erfc}(y) = -\frac{2}{\sqrt{\pi}}e^{-y^2}$ with $y = \delta_c/\sqrt{2S}$ and $\frac{\mathrm{d}y}{\mathrm{d}S} = -\frac{\delta_c}{2\sqrt{2}}S^{-3/2}$; the two minus signs cancel.
(3) The same fraction of walks expressed per unit ν instead of per unit S: a density changes by the Jacobian (Chapter 2).
(4) $S = \delta_c^2/\nu^2$, so $S^{-3/2} = \nu^3/\delta_c^3$ and $|\mathrm{d}S/\mathrm{d}\nu| = 2\delta_c^2/\nu^3$.

$F(\nu)\,\mathrm{d}\nu$ is the fraction of all mass in haloes whose ν lies in $[\nu, \nu + \mathrm{d}\nu]$, and $\int_0^\infty F\,\mathrm{d}\nu = 1$, as it must be now that every mass element is in a halo. It is called the **multiplicity function**, and it is *universal*: the power spectrum, the redshift and the mass have all disappeared into ν.

To turn mass fractions into numbers of haloes, note that a halo of mass M holds mass M. The mass per unit volume in haloes with masses in $[M, M + \mathrm{d}M]$ is $\bar\rho\, F(\nu)|\mathrm{d}\nu/\mathrm{d}M|\,\mathrm{d}M$; dividing by M counts them:

$$\frac{\mathrm{d}n}{\mathrm{d}\ln M} \overset{(1)}{=} \frac{\bar\rho}{M}F(\nu)\left|\frac{\mathrm{d}\nu}{\mathrm{d}\ln M}\right| \overset{(2)}{=} \frac{\bar\rho}{M}\nu F(\nu)\left|\frac{\mathrm{d}\ln\sigma}{\mathrm{d}\ln M}\right|. \tag{11.56}$$

(1) Haloes per unit volume per unit $\ln M$: the mass density in that range divided by the mass of one halo, with $\mathrm{d}M = M\,\mathrm{d}\ln M$.
(2) $\nu = \delta_c/\sigma$, so $\mathrm{d}\ln\nu = -\mathrm{d}\ln\sigma$ and $\mathrm{d}\nu/\mathrm{d}\ln M = -\nu\,\mathrm{d}\ln\sigma/\mathrm{d}\ln M$.

This is the Press–Schechter mass function. In the notation of [CooraySheth2002 eqs. (56)–(57)], who write ν for our ν^2, it reads $m^2 n/\bar\rho\;\mathrm{d}m/m = \nu f(\nu)\,\mathrm{d}\nu/\nu$ with $\nu f(\nu) = \sqrt{\nu/2\pi}\,e^{-\nu/2}$; substituting $\nu \to \nu^2$ gives $\nu F(\nu)/2$, the same thing.

Watch out: Two meanings of ν

Papers on the mass function use $\nu = \delta_c/\sigma$ (Press–Schechter's and ours) or $\nu = (\delta_c/\sigma)^2$ (Sheth and Tormen's, and Cooray and Sheth's). The multiplicity "$\nu f(\nu)$" then differs by a factor of 2, because $\mathrm{d}\ln\nu^2 = 2\,\mathrm{d}\ln\nu$. Check the definition before comparing a formula or a plot.

Example: the characteristic mass and the clusters. The mass M_* at which $\sigma(M_*) = \delta_c$, so $\nu = 1$, separates common haloes from rare ones. For our cosmology at $z = 0$, $M_* = 4.413\times10^{12}\,h^{-1}M_\odot$. A Milky-Way-sized halo of $10^{12}\,h^{-1}M_\odot$ has $\nu = 0.7941$, and the fraction of all mass in haloes heavier than it is $\int_\nu^\infty F\,\mathrm{d}\nu' = \operatorname{erfc}(\nu/\sqrt{2}) = 0.4269$: a bit under half of all matter sits in haloes at least as massive as ours. For clusters the number density follows from integrating (11.56) over $\ln M$: $n(>10^{14}) = 3.185\times10^{-5}\,h^3\mathrm{Mpc}^{-3}$ and $n(>10^{15}) = 1.083\times10^{-7}\,h^3\mathrm{Mpc}^{-3}$ (masses

in $h^{-1}M_\odot$), so a box $1\,h^{-1}$Gpc on a side holds about $10^9 \times 1.083 \times 10^{-7} \approx 100$ haloes above $10^{15}\,h^{-1}M_\odot$, where $\sigma = 0.5348$ and $\nu \approx 3$.

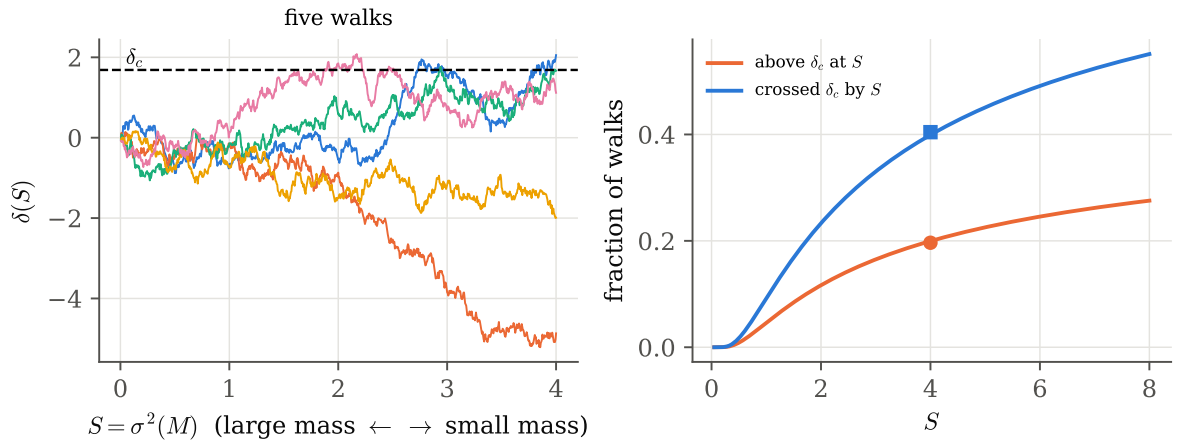


Figure 11.25: Left: five excursion-set walks with independent Gaussian steps, against the threshold δ_c; moving right means smoothing on smaller scales. Right: the fraction of walks above δ_c at S (orange) and the fraction that have crossed it at least once by S (blue), with the simulated values at $S = 4$ as points. The second curve is twice the first at every S.

11.8.4 Random walks on a computer

The reflection argument is short enough to doubt, so we simulate it. `09_mass_function.py` draws 50000 walks on a grid of 4000 values of S between 10^{-3} and 60, each step a Gaussian of variance equal to the grid spacing, and records the first grid point at which each walk exceeds δ_c. A walk sampled only at grid points can cross the barrier and come back between two of them unseen. For Brownian motion the probability of such a hidden excursion is known exactly, again by reflection: between two neighbouring points at heights $x_0, x_1 < \delta_c$, separated by ΔS, the walk crosses with probability

$$\frac{p_{\Delta S}(x_1 - x_0 \text{ via } \delta_c)}{p_{\Delta S}(x_1 - x_0)} = \frac{e^{-(2\delta_c - x_0 - x_1)^2/2\Delta S}}{e^{-(x_1 - x_0)^2/2\Delta S}} = \exp\!\left[-\frac{2(\delta_c - x_0)(\delta_c - x_1)}{\Delta S}\right], \tag{11.57}$$

because the paths from x_0 to x_1 that touch δ_c are, after reflection of their end, the paths from x_0 to $2\delta_c - x_1$, and the difference of the two squares in the exponent is $4(\delta_c - x_0)(\delta_c - x_1)$. The function `first_crossings` in `lib_web.py` draws this event too, which makes the simulated crossings those of the continuous walk.

Listing 11.8: The excursion-set walks and their first crossings. (`code/ch11/lib_web.py`)

```
def first_crossings(S_grid, nwalk, rng, barrier=DELTA_C, lower=None, chunk=2000):
    """Excursion-set walks delta(S) with independent N(0, dS) steps on the grid S_grid.

    Returns, per walk, the S of the first up-crossing of 'barrier' (np.inf if none) and, if
    'lower' is given, the S of the first down-crossing of 'lower' that happens before any
    up-crossing (np.inf otherwise). Between two grid points the walk is a Brownian bridge,
    which
    crosses a level b that both end points lie below with probability
    exp(-2 (b - x0)(b - x1)/dS); drawing that event too makes the result exact for
    continuous
    Brownian motion instead of only for the discrete walk.
    """
    dS = np.diff(np.concatenate([[0.0], S_grid]))
```

```
    up = np.full(nwalk, np.inf)
    down = np.full(nwalk, np.inf)
    done = 0
    while done < nwalk:
        n = min(chunk, nwalk - done)
        walk = np.cumsum(rng.standard_normal((n, len(S_grid))) * np.sqrt(dS), axis=1)
        prev = np.concatenate([np.zeros((n, 1)), walk[:, :-1]], axis=1)
        u = rng.random((n, len(S_grid)))
        bridge = np.exp(-2 * np.clip(barrier - prev, 0, None) * np.clip(barrier - walk, 0,
    None) / dS)
        hit = (walk >= barrier) | (u < bridge)
        first = np.where(hit.any(1), hit.argmax(1), -1)
        up[done:done + n] = np.where(first >= 0, S_grid[np.maximum(first, 0)], np.inf)
        if lower is not None:
            bridge_v = np.exp(-2 * np.clip(prev - lower, 0, None) * np.clip(walk - lower,
    0, None) / dS)
            hitv = (walk <= lower) | (rng.random((n, len(S_grid))) < bridge_v)
            fv = np.where(hitv.any(1), hitv.argmax(1), -1)
            ok = (fv >= 0) & ((first < 0) | (fv < first))
            down[done:done + n] = np.where(ok, S_grid[np.maximum(fv, 0)], np.inf)
        done += n
    return up, down
```

The loop works in chunks of walks to keep memory small. `walk` is the cumulative sum of the Gaussian steps; `prev` is the same walk shifted by one point, so each pair `(prev, walk)` is one step. `hit` marks a step where the walk is above the barrier at its end or crossed it in between (the uniform number `u` decides the hidden crossing with probability (11.57)), and `argmax` finds the first such step. The optional lower barrier is used for voids in Section 11.11.

The factor of 2, measured. At $S = 4$ ($\nu = 0.843$) the fraction of walks above δ_c is 0.1965, against $\frac{1}{2}\operatorname{erfc}(\nu/\sqrt{2}) = 0.1996$, and the fraction that have crossed by then is 0.4043, against $\operatorname{erfc}(\nu/\sqrt{2}) = 0.3992$ (Figure 11.25, right). This check uses 20000 walks, so the binomial standard error of each fraction, $\sqrt{p(1-p)/N}$, is 0.0028 and 0.0035. Each measured fraction should be judged against its own error bar. The ratio of the two measured fractions is 2.058 ± 0.034, against 2.

The multiplicity function, measured. Histogramming $\nu = \delta_c/\sqrt{S_{\text{first}}}$ over all walks gives the blue points of Figure 11.26 (left), on top of $\sqrt{2/\pi}\, e^{-\nu^2/2}$. Over the 26 bins with more than fifty walks the χ^2 of the points about (11.55), with Poisson errors, is 23.1; the expected value for a correct formula is the number of bins, with a standard deviation of about $\sqrt{2 \times 26} \approx 7$. The grey dotted curve, without the factor 2, misses every point.

The same field, two filters. A real field is not smoothed with a sharp-k filter: a halo is a region in real space. The second experiment in the script makes 4 Gaussian fields on a 160^3 grid and reads the walk at every cell by smoothing each field at 120 scales, once with the sharp-k filter and once with a top hat. The error bars come from the scatter between the fields. The cells of one field are not independent: neighbours share the same smoothed modes, so counting cells as independent would give error bars far too small. With the sharp-k filter (green squares) the points follow Press–Schechter; at $\nu \approx 2.475$ they lie a factor 0.72 ± 0.05 below it. Part of the gap is that these walks are read at only 120 scales, without the correction (11.57), so crossings between two scales are missed; part is the luck of the fields, since one field alone scatters by 0.10 in this ratio. With the top hat (orange triangles) the points fall clearly below Press–Schechter at $\nu \lesssim 1.5$ and reach it, even slightly exceed it, at large ν (a factor 1.07 ± 0.02 at $\nu \approx 2.475$). A top hat in real space has a broad Fourier transform, so neighbouring scales share most of their

modes. The steps of the walk are then strongly correlated, the walk is smooth, and a smooth walk rarely makes the brief excursions above the barrier that the reflection argument pairs with the walks ending above it. The factor of 2 belongs to the sharp-k filter; with a realistic filter the exact first-crossing distribution has no closed form, which is one reason the theory is calibrated on simulations [226].

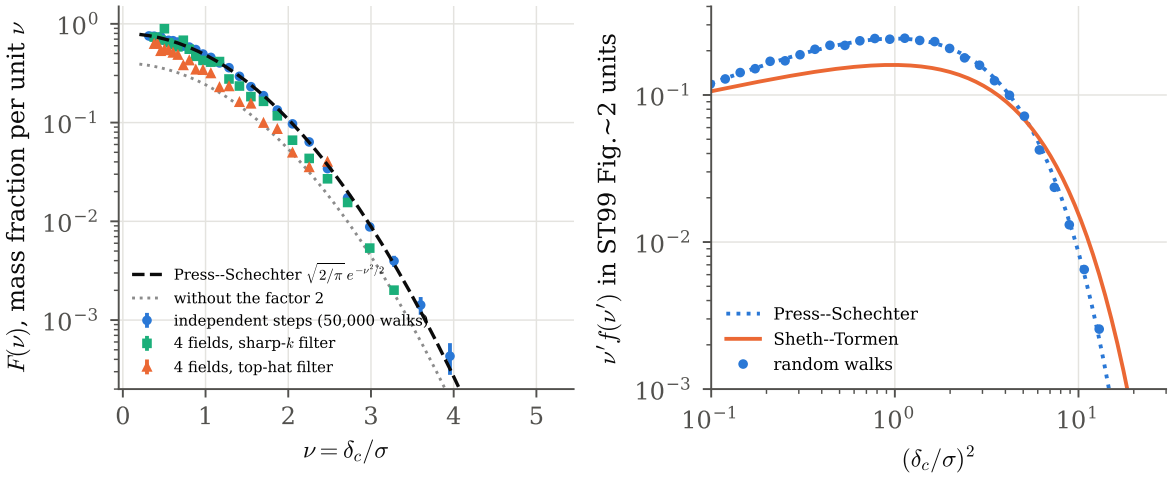


Figure 11.26: Left: the fraction of mass per unit ν from 50000 walks with independent steps (blue), from walks read off 4 Gaussian fields with a sharp-k filter (green) and a top-hat filter (orange), against Press–Schechter (dashed) and the same formula without the factor 2 (dotted). Right: Press–Schechter (dotted) and Sheth–Tormen (solid) on the axes of [ShethTormen1999 Fig. 2], with the random walks again as points.

11.8.5 Reproducing Sheth and Tormen

Their question. By the late 1990s N-body simulations were large enough to count haloes over four decades of mass. Press–Schechter got the shape roughly right and the details wrong: too many small haloes and too few large ones [CooraySheth2002 Fig. 3]. Sheth and Tormen [227] asked for a mass function that fits the simulated haloes at all masses and redshifts, and for the halo bias that goes with it (the bias is the subject of Section 11.9.3).

Their setup. They used the GIF simulations of three cold-dark-matter cosmologies, identified haloes at several output times, and plotted the abundance against the scaled variable $\nu_{\rm ST} = (\delta_c/\sigma)^2$, the square of ours. If the excursion-set picture is right in spirit, the points from all times should lie on one curve, and they did [ShethTormen1999 Fig. 2].

Their key equation. The curve they fitted is [ShethTormen1999 eq. (10)]; translated to our $\nu = \delta_c/\sigma$ by $\nu_{\rm ST} = \nu^2$ and $F(\nu)\,\mathrm{d}\nu = f(\nu_{\rm ST})\,\mathrm{d}\nu_{\rm ST}$, it reads

$$F_{\rm ST}(\nu) = A\sqrt{\frac{2a}{\pi}}\left[1 + (a\nu^2)^{-p}\right] e^{-a\nu^2/2}, \qquad a = 0.707, \quad p = 0.3. \tag{11.58}$$

The two shape parameters have direct meanings. $a < 1$ stretches the exponential cut-off to higher ν, which gives more massive haloes. The bracket $1 + (a\nu^2)^{-p}$ with $p > 0$ rises toward small ν as a weak power, where Press–Schechter has a constant. Together with the smaller normalisation A below (Press–Schechter's is $\frac{1}{2}$), the net effect over the masses that matter for

galaxies and groups is fewer haloes than Press–Schechter. The constant A is not free: every mass element is in some halo, so $\int_0^\infty F_{\rm ST}\,\mathrm{d}\nu = 1$. With $t = a\nu^2/2$,

$$\int_0^\infty F_{\rm ST}\,\mathrm{d}\nu \overset{(1)}{=} A\sqrt{\frac{2a}{\pi}}\int_0^\infty \left[1+(2t)^{-p}\right]e^{-t}\,\frac{\mathrm{d}t}{\sqrt{2at}} \overset{(2)}{=} \frac{A}{\sqrt{\pi}}\left[\Gamma\left(\tfrac12\right)+2^{-p}\,\Gamma\left(\tfrac12-p\right)\right]$$
$$\overset{(3)}{=} A\left[1+\frac{2^{-p}\,\Gamma(\frac12-p)}{\sqrt{\pi}}\right]=1 \quad\Longrightarrow\quad A=\left[1+\frac{2^{-p}\,\Gamma(\frac12-p)}{\sqrt{\pi}}\right]^{-1}=0.3222. \tag{11.59}$$

(1) $\nu=\sqrt{2t/a}$, so $\mathrm{d}\nu=\mathrm{d}t/\sqrt{2at}$, and $a\nu^2=2t$.
(2) The prefactors combine to $A/\sqrt{\pi}$; the two terms are Gamma integrals, $\int_0^\infty t^{s-1}e^{-t}\mathrm{d}t=\Gamma(s)$ with $s=\frac12$ and $s=\frac12-p$.
(3) $\Gamma(\frac12)=\sqrt{\pi}$.

For $p=0.3$, $\Gamma(0.2)=4.591$ and $2^{-0.3}=0.812$, giving the value quoted by [CooraySheth2002 eq. (59)]. With $a=1$ and $p=0$ the bracket is 2, so $A=\frac12$ and (11.58) becomes $\frac12\sqrt{2/\pi}\,(1+1)\,e^{-\nu^2/2}$, which is Press–Schechter: the fit contains the excursion-set result as a special case.

What we compute. `lib_web.py` implements both multiplicities and (11.56), with $\sigma(M)$ from the CAMB linear spectrum; `09_mass_function.py` plots them on Sheth and Tormen's axes, $\nu_{\rm ST}f(\nu_{\rm ST})=\nu F(\nu)/2$ against $\nu_{\rm ST}$ (Figure 11.26, right), and converts both to numbers of haloes (Figure 11.27).

Side by side. The right panel of Figure 11.26 reproduces the two theory curves of their Fig. 2: Sheth–Tormen lies below Press–Schechter for $0.05 \lesssim \nu_{\rm ST} \lesssim 5$ and above it beyond, by a growing factor (it overtakes Press–Schechter again only at masses far below the range plotted). In numbers of haloes at $z=0$ (Figure 11.27), Sheth–Tormen predicts 0.7214 times as many haloes as Press–Schechter at $10^{11}\,h^{-1}M_\odot$, 0.6643 times as many at 10^{13}, and 1.809 times as many at 10^{15}. Above $10^{15}\,h^{-1}M_\odot$ it gives $2.435\times10^{-7}\,h^3\mathrm{Mpc}^{-3}$ against 1.083×10^{-7}, about 240 instead of 100 such clusters per $(h^{-1}\mathrm{Gpc})^3$. The mass in haloes above $10^{12}\,h^{-1}M_\odot$ drops from 0.4269 to 0.3125 of the total.

Why they differ. Each difference has one cause: input (H2). Real perturbations are triaxial, and a triaxial region feels the tidal pull of its surroundings, which resists collapse along the longest axis. Small haloes come from regions with large σ, where tides are relatively strong; they need a higher linear contrast than δ_c to collapse, so there are fewer of them. Massive haloes come from rare, high peaks, which are nearly spherical and collapse almost as the spherical model says. In walk language, ellipsoidal collapse replaces the flat barrier δ_c by one that rises with S, a *moving barrier*. Sheth and Tormen pointed to this explanation [ShethTormen1999 §4], and the first crossings of such a barrier were later shown to reproduce the shape of (11.58) [228]. The random walks themselves cannot tell us this: they verify the mathematics of (H1)–(H3), and the simulations test (H2).

The lesson. The Press–Schechter factor of 2 is the cloud-in-cloud mass, recovered exactly by the first-crossing problem; what remains between Press–Schechter and the simulations, a factor of two at the extremes, is the physics of non-spherical collapse.

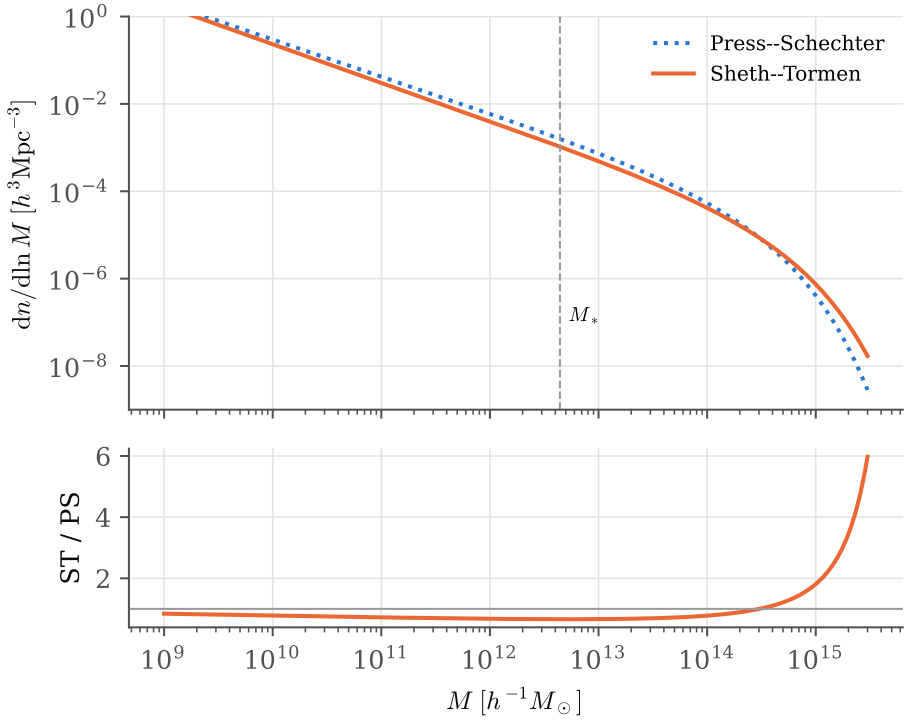


Figure 11.27: Haloes per unit volume per unit $\ln M$ at $z = 0$ in the fiducial cosmology, from (11.56) with the Press–Schechter (dotted) and Sheth–Tormen (solid) multiplicities; the dashed line marks M_*. Lower panel: their ratio. The abundance falls exponentially beyond M_*, and there the two formulas differ most.

Key idea: The halo mass function from first crossings

With a sharp-k filter the smoothed linear contrast at a point is a Brownian motion in $S = \sigma^2(M)$. The fraction of mass in haloes of mass M is the fraction of walks that first cross δ_c at $S = \sigma^2(M)$; by reflection it is twice the naive Press–Schechter fraction, and $\mathrm{d}n/\mathrm{d}\ln M = (\bar\rho/M)\,\nu F(\nu)\,|\mathrm{d}\ln\sigma/\mathrm{d}\ln M|$ with $F = \sqrt{2/\pi}\,e^{-\nu^2/2}$. Simulations prefer the Sheth–Tormen form (11.58), the signature of ellipsoidal collapse.

11.9 Inside and around a halo

The mass function of Section 11.8 tells us how many haloes there are of each mass. To build a density field out of haloes we need two more facts, one for each end of the range of scales. On small scales, inside one halo, the field is the halo's own density profile: how is the mass arranged between its centre and its edge? On large scales, the field is made by how haloes are placed relative to one another: do they follow the matter, or do they cluster more strongly? In words, the two questions are:

(1) *How is the mass of a halo of mass M spread around its centre, and what does that look like to a Fourier mode of wavenumber k?*

(2) *If a large region is slightly denser than average, by how much does its number of haloes of mass M go up?*

11.9.1 The density profile of a halo

N-body haloes of all masses, in all the cold-dark-matter cosmologies that have been simulated, have nearly the same spherically averaged profile [94],

$$\rho(r\,|\,M) = \frac{\rho_s}{(r/r_s)\,(1 + r/r_s)^2}, \tag{11.60}$$

the **NFW profile**. It falls as r^{-1} well inside the scale radius r_s and as r^{-3} well outside it. There is no complete derivation of this shape from first principles; it is a description of what simulations produce [CooraySheth2002 p. 34]. We take it as an input.

Where does a halo end? The r^{-3} tail would hold an infinite mass, so the profile must be cut. We cut it at the radius $r_{\rm vir}$ inside which the mean density is $\Delta_{\rm vir} = 330$ times the mean matter density, the virialised density of the spherical-collapse model for ΛCDM today (Section 11.8.1): $M = \frac{4\pi}{3}\,330\,\bar\rho\, r_{\rm vir}^3$. The ratio $c = r_{\rm vir}/r_s$ is the **concentration**. Simulated haloes of lower mass are more concentrated, because they formed earlier, when the Universe was denser, and on average

$$c(M) = 9\left(\frac{M}{M_*}\right)^{-0.13} \tag{11.61}$$

at $z = 0$, with a log-normal scatter of about 0.25 in $\ln c$ that we ignore [CooraySheth2002 eqs. (77)–(78)].

How much mass is inside a radius? Integrate the profile in shells, with $x = r/r_s$:

$$\begin{aligned} M(<r) &\overset{(1)}{=} \int_0^r 4\pi r'^2 \rho(r')\,\mathrm{d}r' \overset{(2)}{=} 4\pi\rho_s r_s^3 \int_0^{r/r_s} \frac{x\,\mathrm{d}x}{(1+x)^2} \overset{(3)}{=} 4\pi\rho_s r_s^3 \Big[\ln(1+x) + \frac{1}{1+x}\Big]_0^{r/r_s} \\ &\overset{(4)}{=} 4\pi\rho_s r_s^3\, m\!\left(\frac{r}{r_s}\right), \qquad m(y) \equiv \ln(1+y) - \frac{y}{1+y}. \end{aligned} \tag{11.62}$$

(1) The mass of a thin shell is its area times its thickness times the density.
(2) Substitute $r' = r_s x$; then $r'^2\rho\,\mathrm{d}r' = r_s^3\rho_s\, x^2\mathrm{d}x/[x(1+x)^2]$.
(3) Write $x/(1+x)^2 = 1/(1+x) - 1/(1+x)^2$ and integrate each term.
(4) At $x = 0$ the bracket is 1; subtracting it gives $\ln(1+y) + 1/(1+y) - 1 = \ln(1+y) - y/(1+y)$.

The total mass is $M = 4\pi\rho_s r_s^3\, m(c)$ [CooraySheth2002 eq. (76)], so a halo is fixed completely by M: M gives $r_{\rm vir}$, (11.61) gives c and hence r_s, and $M = 4\pi\rho_s r_s^3 m(c)$ gives ρ_s.

Example: a group of $10^{14}\,h^{-1}M_\odot$. Here $c = 5.999$, $r_{\rm vir} = 0.9387\,h^{-1}$Mpc and $r_s = 0.1565\,h^{-1}$Mpc. The fraction of the mass inside r_s is $m(1)/m(c) = (\ln 2 - \frac{1}{2})/(\ln 7 - \frac{6}{7}) = 0.193/1.089 = 0.18$: less than a fifth of the mass, in less than $1/c^3 = 0.5\%$ of the volume. The characteristic density follows from equating the two expressions for M, $\frac{4\pi}{3}330\bar\rho\, c^3 r_s^3 = 4\pi\rho_s r_s^3 m(c)$, so $\rho_s/\bar\rho = (330/3)\,c^3/m(c) \approx 2.2\times10^4$.

11.9.2 A halo as seen by a Fourier mode

Power spectra are built from Fourier modes, so we need the transform of the profile. Normalise it by the mass, so that it describes the shape alone:

$$u(\boldsymbol{k}\,|\,M) = \frac{1}{M}\int \mathrm{d}^3x\;\rho(\boldsymbol{x}\,|\,M)\,e^{-i\boldsymbol{k}\cdot\boldsymbol{x}}. \tag{11.63}$$

For a spherical profile only $k = |\boldsymbol{k}|$ matters. Put $\boldsymbol{k}$ along the polar axis, so $\boldsymbol{k}\cdot\boldsymbol{x} = kr\cos\theta$:

$$u(k\,|\,M) \overset{(1)}{=} \frac{1}{M}\int_0^{r_{\rm vir}} r^2\rho(r)\,\mathrm{d}r \int_0^{2\pi}\mathrm{d}\phi \int_{-1}^{1} e^{-ikr\mu}\,\mathrm{d}\mu \overset{(2)}{=} \frac{1}{M}\int_0^{r_{\rm vir}} 4\pi r^2\rho(r)\,\frac{\sin kr}{kr}\,\mathrm{d}r. \tag{11.64}$$

(1) Spherical coordinates, with $\mu = \cos\theta$ and $\mathrm{d}^3x = r^2 \mathrm{d}r\, \mathrm{d}\phi\, \mathrm{d}\mu$.
(2) $\int_{-1}^{1} e^{-ikr\mu} \mathrm{d}\mu = (e^{ikr} - e^{-ikr})/(ikr) = 2\sin(kr)/(kr)$, and the ϕ integral gives 2π.

Two limits are visible at once. For $k\, r_{\rm vir} \ll 1$, $\sin(kr)/(kr) \approx 1 - k^2r^2/6$, so $u \approx 1 - k^2\langle r^2\rangle/6$: a mode much longer than the halo sees a point mass, $u \to 1$. For $k\, r_s \gg 1$ the factor $\sin(kr)/(kr)$ oscillates rapidly and u falls: the mode resolves the halo and averages over its interior.

The NFW transform in closed form. With $x = r/r_s$, $\kappa = kr_s$ and $4\pi\rho_s r_s^3/M = 1/m(c)$ from (11.62),

$$
\begin{aligned}
u(k\,|\,M) &\overset{(1)}{=} \frac{1}{m(c)\,\kappa} \int_0^c \frac{\sin\kappa x}{(1+x)^2}\,\mathrm{d}x \overset{(2)}{=} \frac{1}{m(c)\,\kappa}\left\{\left[-\frac{\sin\kappa x}{1+x}\right]_0^c + \kappa\int_0^c \frac{\cos\kappa x}{1+x}\,\mathrm{d}x\right\} \\
&\overset{(3)}{=} \frac{1}{m(c)}\left\{-\frac{\sin c\kappa}{(1+c)\kappa} + \int_\kappa^{(1+c)\kappa} \frac{\cos(t-\kappa)}{t}\,\mathrm{d}t\right\} \\
&\overset{(4)}{=} \frac{1}{m(c)}\left\{\sin\kappa\,\big[\mathrm{Si}((1+c)\kappa) - \mathrm{Si}(\kappa)\big] + \cos\kappa\,\big[\mathrm{Ci}((1+c)\kappa) - \mathrm{Ci}(\kappa)\big] - \frac{\sin c\kappa}{(1+c)\kappa}\right\}. \qquad (11.65)
\end{aligned}
$$

(1) Insert (11.60) into (11.64): $4\pi r^2\rho\,\mathrm{d}r/M = x^2\,\mathrm{d}x/[m(c)\,x(1+x)^2]$, and $\sin(kr)/(kr) = \sin(\kappa x)/(\kappa x)$; one x cancels.
(2) Integrate by parts, with $(1+x)^{-2} = \frac{\mathrm{d}}{\mathrm{d}x}[-(1+x)^{-1}]$.
(3) The boundary term at $x = 0$ vanishes. In the integral substitute $t = \kappa(1+x)$, so $\kappa x = t - \kappa$ and $\mathrm{d}x/(1+x) = \mathrm{d}t/t$.
(4) $\cos(t-\kappa) = \cos t\cos\kappa + \sin t\sin\kappa$, and $\int \sin t/t\,\mathrm{d}t$ and $\int \cos t/t\,\mathrm{d}t$ are the sine and cosine integrals, $\mathrm{Si}(y) = \int_0^y \sin t/t\,\mathrm{d}t$ and $\mathrm{Ci}(y) = -\int_y^\infty \cos t/t\,\mathrm{d}t$.

This is [CooraySheth2002 eq. (81)]. `lib_web.py` evaluates it with `scipy.special.sici`:

Listing 11.9: The halo edge, the concentration and the NFW transform. (`code/ch11/lib_web.py`)

```
def concentration(M, mstar):
    """Mean NFW concentration c(M) = 9 (M/M*)^-0.13 (Cooray & Sheth 2002, eq. 78, z = 0)."""
    return 9.0 * (np.asarray(M, float) / mstar) ** (-0.13)


def r_halo(M, delta=DELTA_HALO):
    """Halo radius: the sphere of mean density DELTA_HALO * rho_m (the virial radius) that
    holds M."""
    return (3 * np.asarray(M, float) / (4 * np.pi * delta * RHO_M)) ** (1 / 3)


def u_nfw(k, M, c, delta=DELTA_HALO):
    """Normalised Fourier transform of an NFW profile truncated at r_halo (CS02 eq. 81).

    k may be 1-D and M, c 1-D; returns shape (len(M), len(k)).
    """
    k = np.atleast_1d(k)[None, :]
    M = np.atleast_1d(M)[:, None]
    c = np.atleast_1d(c)[:, None]
    rs = r_halo(M, delta) / c
    x = k * rs
    si1, ci1 = sici((1 + c) * x)
    si0, ci0 = sici(x)
    mnorm = np.log(1 + c) - c / (1 + c)
    return (np.sin(x) * (si1 - si0) - np.sin(c * x) / ((1 + c) * x) + np.cos(x) * (ci1 -
    ci0)) / mnorm
```

`r_halo` inverts $M = \frac{4\pi}{3}\Delta\bar\rho r^3$; `u_nfw` builds $r_s = r_{\rm vir}/c$, $\kappa = kr_s$, and returns the bracket of (11.65) divided by $m(c)$, for a whole grid of masses and wavenumbers at once. `10_halo_model.py`

checks it against the integral (11.64) done numerically for $M = 10^{14}\,h^{-1}M_\odot$ at three wavenumbers; the largest relative difference is 8.295×10^{-10}.

Example: when does a mode see inside a halo? At $k = 1\,h\,\mathrm{Mpc}^{-1}$, $u = 0.9598$ for the group of the previous example and $u = 0.9984$ for a $10^{12}\,h^{-1}M_\odot$ halo. A mode of wavelength $2\pi/k \approx 6\,h^{-1}$Mpc is still long compared with either halo, and both look nearly like points. Figure 11.28 (left) shows that u drops below $\frac{1}{2}$ at $k \approx 2\,h\,\mathrm{Mpc}^{-1}$ for $10^{15}\,h^{-1}M_\odot$, at $k \approx 4$ for 10^{14} and only near $k \approx 50$ for 10^{11}: the smallest scales are resolved inside the smallest haloes, which therefore control the power at the largest k [CooraySheth2002 Fig. 9].

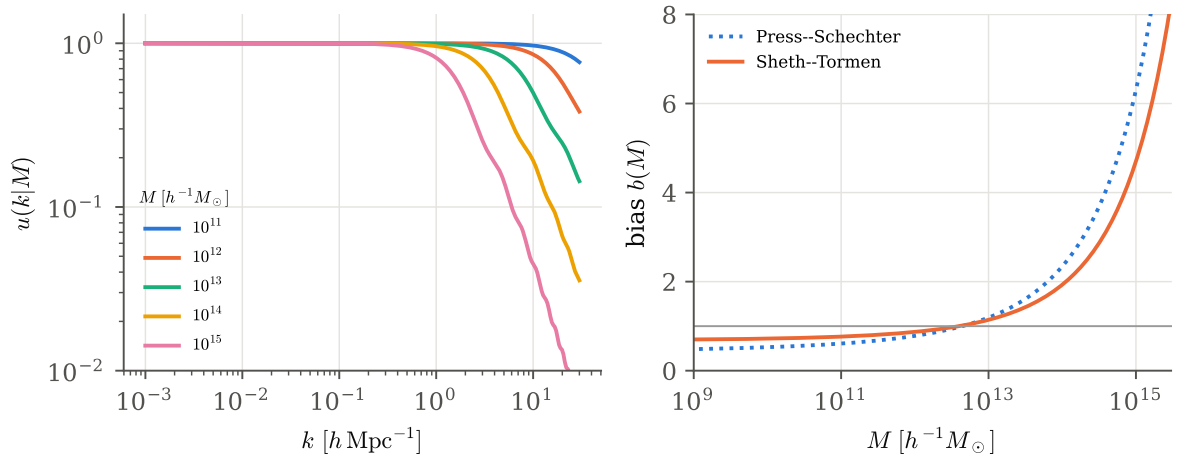


Figure 11.28: Left: the normalised Fourier transform $u(k\,|\,M)$ of NFW haloes with the concentrations (11.61), for five masses. Each curve is 1 for modes longer than the halo and falls once the mode resolves its interior; heavier haloes are larger and fall at smaller k. Right: the large-scale bias of haloes from the peak-background split with the Press–Schechter (dotted) and Sheth–Tormen (solid) mass functions. Haloes much heavier than $M_* = 4.413 \times 10^{12}\,h^{-1}M_\odot$ are several times more clustered than the matter.

11.9.3 Where haloes sit: the peak-background split

Clusters of galaxies are far more strongly clustered than galaxies, and galaxies more than the dark matter. The excursion set explains why, with no new input. Picture the linear density as a sum of a long wave, the *background*, and short-wavelength ripples on top of it (Figure 11.29). Haloes form where the total crosses δ_c. Where the background is high, the ripples need to climb less far, and the threshold is crossed more often; where it is low, hardly ever. Haloes therefore end up concentrated where the long wave is high: their number density is modulated by the large-scale density, and modulated more strongly than the matter is. That is **halo bias**.

The question in walk language. Consider a large region, of Lagrangian volume V, whose linear contrast smoothed on its own scale is δ_ℓ. Its variance on that scale is $S_\ell = \sigma^2(\bar\rho V)$, small because the region is large. Every walk starting from a point inside it passes through (S_ℓ, δ_ℓ). How many of these walks first cross δ_c at the much larger $S = \sigma^2(M)$ of a small halo?

Move the origin. From (S_ℓ, δ_ℓ) onwards the walk is again a Brownian motion, independent of its past. Measured from that point, it has to climb $\delta_c - \delta_\ell$ in time $S - S_\ell$. For $S_\ell \ll S$ the time is $\approx S$, and the problem is the one we solved with δ_c replaced by $\delta_c - \delta_\ell$ [CooraySheth2002 eq. (63)], [Bernardeau2002 eqs. (550)–(551)]. The number of haloes per unit Lagrangian volume in the region is therefore the mass function with a lowered barrier, $n(M\,|\,\delta_\ell) = n(M)|_{\delta_c \to \delta_c - \delta_\ell}$.

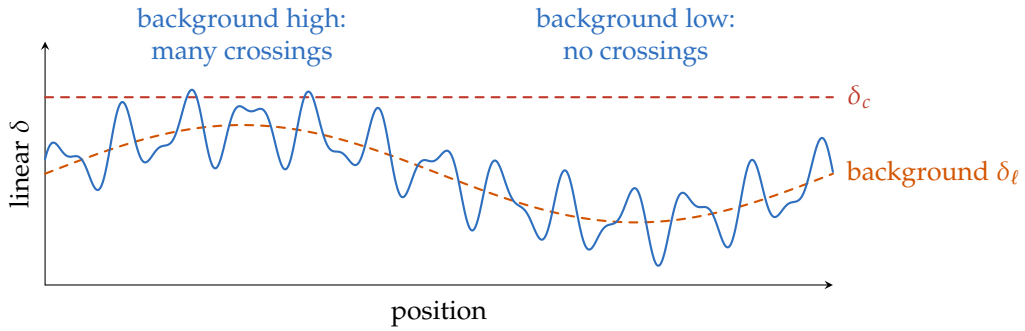


Figure 11.29: The peak-background split. The linear field (blue) is a long wave, the background δ_ℓ (orange), plus short ripples. The ripples are statistically the same everywhere, but they reach the collapse threshold δ_c only where the background lifts them; haloes form there, so their number follows the long wave with an amplified amplitude.

How large is the change? Expand to first order in δ_ℓ, which is small for a large region. In (11.56) only $\nu = \delta_c/\sigma$ depends on δ_c, through the factor $\nu F(\nu)$, and $\partial\nu/\partial\delta_c = \nu/\delta_c$. Hence

$$\delta_h^L \equiv \frac{n(M\,|\,\delta_\ell)}{n(M)} - 1 \overset{(1)}{=} -\delta_\ell \frac{\partial \ln n}{\partial \delta_c} \overset{(2)}{=} -\frac{\delta_\ell}{\delta_c}\frac{\mathrm{d}\ln[\nu F(\nu)]}{\mathrm{d}\ln\nu} \equiv b_L(M)\,\delta_\ell. \tag{11.66}$$

(1) Taylor expansion of n in its δ_c argument, with step $-\delta_\ell$.
(2) Chain rule: $\partial \ln n/\partial\delta_c = (\partial \ln \nu/\partial\delta_c)\,\mathrm{d}\ln(\nu F)/\mathrm{d}\ln\nu$ and $\partial\ln\nu/\partial\delta_c = 1/\delta_c$.

δ_h^L is the excess of haloes per unit *initial* volume, so b_L is the **Lagrangian bias**. For the two multiplicities:

$$\begin{aligned}
\text{Press–Schechter:}\quad & \frac{\mathrm{d}\ln(\nu F)}{\mathrm{d}\ln\nu} \overset{(1)}{=} \frac{\mathrm{d}}{\mathrm{d}\ln\nu}\left[\ln\nu - \frac{\nu^2}{2}\right] = 1-\nu^2, \qquad b_L = \frac{\nu^2-1}{\delta_c},\\
\text{Sheth–Tormen:}\quad & \frac{\mathrm{d}\ln(\nu F)}{\mathrm{d}\ln\nu} \overset{(2)}{=} 1 - a\nu^2 + \frac{\mathrm{d}\ln[1+(a\nu^2)^{-p}]}{\mathrm{d}\ln\nu} \overset{(3)}{=} 1 - a\nu^2 - \frac{2p}{1+(a\nu^2)^p},\\
& b_L = \frac{a\nu^2-1}{\delta_c} + \frac{2p}{\delta_c\,[1+(a\nu^2)^p]}.
\end{aligned} \tag{11.67}$$

(1) $\nu F \propto \nu\, e^{-\nu^2/2}$ by (11.55), and $\mathrm{d}(\nu^2/2)/\mathrm{d}\ln\nu = \nu^2$.
(2) $\nu F_{\rm ST} \propto \nu\,[1+(a\nu^2)^{-p}]\,e^{-a\nu^2/2}$ by (11.58); the first and last factors give $1-a\nu^2$ as before.
(3) $\frac{\mathrm{d}}{\mathrm{d}\ln\nu}(a\nu^2)^{-p} = -2p\,(a\nu^2)^{-p}$, so the derivative of the logarithm is $-2p(a\nu^2)^{-p}/[1+(a\nu^2)^{-p}]$; multiply top and bottom by $(a\nu^2)^p$.

From the initial volume to today's. We observe haloes per unit volume today, not per unit initial volume. The region has contracted: by mass conservation its volume today is $V/(1+\delta)$, where δ is its actual density contrast, which to first order is the linear δ_ℓ. The haloes it contains are the same, so the halo density contrast today, the *Eulerian* one, is

$$1+\delta_h \overset{(1)}{=} (1+\delta)\,(1+\delta_h^L) \overset{(2)}{=} 1 + \delta_\ell + b_L\delta_\ell + O(\delta_\ell^2) \quad\Longrightarrow\quad b(M) = 1 + b_L(M). \tag{11.68}$$

(1) Same number of haloes in a volume smaller by a factor $1+\delta$.
(2) To first order $\delta = \delta_\ell$; drop the product of two small terms.

With (11.67) this is the Mo–White bias for Press–Schechter, $b = 1 + (\nu^2-1)/\delta_c$ [229], and [ShethTormen1999 eq. (12)] for Sheth–Tormen; [CooraySheth2002 eq. (66)] and [Bernardeau2002 eq. (554)] write the same results with ν^2 in place of our ν.

Working definition 11.6 (large-scale halo bias)

On scales much larger than a halo, the number density of haloes of mass M follows the matter, $\delta_h(\boldsymbol{x} \mid M) = b(M)\,\delta(\boldsymbol{x})$, so their power spectrum is $b^2(M)P(k)$. In the peak-background split $b(M) = 1 - \delta_c^{-1}\,\mathrm{d}\ln[\nu F(\nu)]/\mathrm{d}\ln\nu$ at $\nu = \delta_c/\sigma(M)$. *Operational:* measure $b(M)$ in a simulation as $\sqrt{P_{hh}(k)/P(k)}$ at small k, or as P_{hm}/P from the halo–matter cross-spectrum.

Example: bias across the mass range. At M_*, $\nu = 1$ and the Press–Schechter bias is exactly 1: typical haloes cluster like the matter. For $\nu \gg 1$, $b \approx \nu^2/\delta_c$, which grows fast: $b(10^{15}\,h^{-1}M_\odot) = 6.301$ (Press–Schechter) or 4.701 (Sheth–Tormen). For $\nu \to 0$ the Press–Schechter bias approaches $1 - 1/\delta_c = 0.41$: small haloes are anti-biased, because they avoid the dense regions, where their mass has been swallowed by bigger haloes [CooraySheth2002 p. 25]. On our mass grid the smallest values are 0.4485 and 0.6894. A Milky-Way halo of $10^{12}\,h^{-1}M_\odot$ has $b = 0.8707$ (Sheth–Tormen), a group of 10^{14} has $b = 1.93$ (Figure 11.28, right). Sheth–Tormen gives smaller biases at the top end because its exponential cut-off sits at $\sqrt{a}\,\nu$, so a given mass is a less extreme event; that agrees better with the simulations [CooraySheth2002 Fig. 4].

A consistency relation. Average the bias over all the mass. In the region of the previous paragraphs, all the mass is in haloes before and after the background is added. Per unit Lagrangian volume the mass is $\bar\rho$ in both cases, because a Lagrangian volume is defined by the mass it holds:

$$\bar\rho \overset{(1)}{=} \int \mathrm{d}M\, M\, n(M \mid \delta_\ell) \overset{(2)}{=} \int \mathrm{d}M\, M\, n(M)\,[1 + b_L(M)\delta_\ell] \overset{(3)}{=} \bar\rho + \delta_\ell \int \mathrm{d}M\, M\, n(M)\, b_L(M),$$
$$\Longrightarrow \quad \int \mathrm{d}M\, \frac{M\,n(M)}{\bar\rho}\, b(M) \overset{(4)}{=} \int \mathrm{d}M\, \frac{M\,n(M)}{\bar\rho}\,[1 + b_L(M)] = 1. \tag{11.69}$$

(1) All mass in haloes, counted per unit initial volume in the region.
(2) (11.66).
(3) The unperturbed mass function also puts all the mass in haloes, $\int M\,n\,\mathrm{d}M = \bar\rho$; this needs $\int F\,\mathrm{d}\nu = 1$.
(4) The first line forces the δ_ℓ term to vanish for every δ_ℓ, so $\int M n\, b_L = 0$; then use $b = 1 + b_L$.

The mass-weighted mean bias of all haloes is 1 [CooraySheth2002 eq. (71)]: the matter, which is all in haloes, is unbiased with respect to itself. The relation becomes the large-scale limit of the halo model in Section 11.10.

Key idea: Peak-background split

A large-scale overdensity δ_ℓ lowers the collapse barrier of the small-scale walks inside it to $\delta_c - \delta_\ell$. To first order this raises the halo density by $b_L\delta_\ell$ with $b_L = -\delta_c^{-1}\mathrm{d}\ln(\nu F)/\mathrm{d}\ln\nu$, and the contraction of the region adds δ_ℓ, so $b = 1 + b_L$. The bias is fixed by the shape of the mass function alone, and its mass-weighted mean is 1.

11.10 The power spectrum built from haloes

Linear theory predicts $P(k)$ on large scales; on small scales gravity makes the real spectrum climb far above it. In Section T1.4 we took that non-linear spectrum from halofit, a formula fitted to N-body simulations [41]. A fit reproduces a simulation; it does not explain it. The

three facts about haloes from Sections 11.8 and 11.9 let us *build* the non-linear spectrum and see where each part of it comes from:

> *If every particle of matter belongs to a halo, haloes come in the numbers of the mass function, have NFW profiles and cluster with the peak-background-split bias, what is the power spectrum of the matter?*

The construction mirrors the one in Section 11.1, where galaxies were points scattered by a smooth field. Here the matter is made of haloes scattered by a smooth field, and every halo carries a profile instead of being a point.

11.10.1 The inputs

Physical inputs.

(K1) All matter is in haloes. Halo i, of mass M_i, centred at $\boldsymbol{x}_i$, contributes the density $M_i\, u(\boldsymbol{x}-\boldsymbol{x}_i \,|\, M_i)$, where $u(\cdot\,|\,M)$ is the NFW profile divided by its mass, so $\int u\,\mathrm{d}^3x = 1$ (Section 11.9.1). Haloes of equal mass have equal profiles.

(K2) Haloes of mass M have mean number density $n(M)\,\mathrm{d}M$, the Sheth–Tormen mass function (Section 11.8.5).

(K3) The probability of finding the centres of two haloes, of masses M_1 and M_2, in volumes $\mathrm{d}V_1$ and $\mathrm{d}V_2$ a distance r apart is $n(M_1)n(M_2)\,[1+\xi_{hh}(r)]\,\mathrm{d}V_1\mathrm{d}V_2$, with $\xi_{hh}(r) = b(M_1)b(M_2)\,\xi_{\mathrm{lin}}(r)$: halo centres trace the linear field with the bias of Section 11.9.3.

Mathematical choice. We work in a periodic box of volume V, where $\delta_{\boldsymbol{k}} = \int_V \delta(\boldsymbol{x})e^{-i\boldsymbol{k}\cdot\boldsymbol{x}}\mathrm{d}^3x$ and $\langle|\delta_{\boldsymbol{k}}|^2\rangle = V\,P(k)$ for $\boldsymbol{k}\neq 0$ (Section 7.4.1).

(K3) is the weakest input. It is right on large scales, where Section 11.9.3 derived it, and wrong on small ones: two haloes cannot overlap, so ξ_{hh} must fall to -1 below the sum of their radii, while $b_1b_2\xi$ does not. Where this matters will show up in the comparison with simulations.

11.10.2 One halo and two haloes

The density is a sum over haloes, $\rho(\boldsymbol{x}) = \sum_i M_i u(\boldsymbol{x}-\boldsymbol{x}_i\,|\,M_i)$, and a shift in real space is a phase in Fourier space, so $\rho_{\boldsymbol{k}} = \sum_i M_i\,u(k\,|\,M_i)\,e^{-i\boldsymbol{k}\cdot\boldsymbol{x}_i}$, with $u(k\,|\,M)$ the transform (11.65). For $\boldsymbol{k}\neq 0$, $\delta_{\boldsymbol{k}} = \rho_{\boldsymbol{k}}/\bar\rho$. The power is the average of $|\delta_{\boldsymbol{k}}|^2 = \delta_{\boldsymbol{k}}\delta^*_{\boldsymbol{k}}$, a double sum over haloes, and the terms fall into two groups: both factors from the same halo, or from two different haloes.

$$
\begin{aligned}
\bar\rho^2\,\langle|\delta_{\boldsymbol{k}}|^2\rangle &\overset{(1)}{=} \Big\langle \sum_i M_i^2\,u_i^2 \Big\rangle + \Big\langle \sum_{i\neq j} M_iM_j\,u_iu_j\,e^{-i\boldsymbol{k}\cdot(\boldsymbol{x}_i-\boldsymbol{x}_j)} \Big\rangle \\
&\overset{(2)}{=} V\int \mathrm{d}M\,n(M)\,M^2u^2(k\,|\,M) + \int \mathrm{d}M_1\mathrm{d}M_2\,n_1n_2\,M_1M_2\,u_1u_2 \\
&\qquad \times \int_V \mathrm{d}^3x_1 \int_V \mathrm{d}^3x_2\,[1+\xi_{hh}(\boldsymbol{x}_1-\boldsymbol{x}_2)]\,e^{-i\boldsymbol{k}\cdot(\boldsymbol{x}_1-\boldsymbol{x}_2)} \\
&\overset{(3)}{=} V\int \mathrm{d}M\,n(M)\,M^2u^2(k\,|\,M) + V\int \mathrm{d}M_1\mathrm{d}M_2\,n_1n_2\,M_1M_2\,u_1u_2\,b_1b_2\,P_{\mathrm{lin}}(k). \qquad (11.70)
\end{aligned}
$$

(1) Expand the square. We write $u_i = u(k\,|\,M_i)$, which is real for a spherical profile; the phase cancels against its conjugate when $i=j$.

(2) First sum: the expected number of haloes of mass M in the box is $V\,n(M)\,\mathrm{d}M$ (K2). Second sum: the expected number of pairs is the pair density of (K3) integrated over both positions.
(3) The "1" in the bracket gives $\int \mathrm{d}^3r\, e^{-i\boldsymbol{k}\cdot\boldsymbol{r}} = 0$ for $\boldsymbol{k} \neq 0$ in a periodic box. With $\boldsymbol{r} = \boldsymbol{x}_1 - \boldsymbol{x}_2$, the ξ_{hh} term gives V times the Fourier transform of $b_1 b_2 \xi_{\rm lin}$, which is $b_1 b_2 P_{\rm lin}(k)$.

Dividing by $V\bar\rho^2$ gives the **halo model** power spectrum [CooraySheth2002 eqs. (88)–(89)], [230]:

$$P(k) = P_{1h}(k) + P_{2h}(k),$$
$$P_{1h}(k) = \int \mathrm{d}M\, n(M) \Big(\frac{M}{\bar\rho}\Big)^2 u^2(k\,|\,M), \qquad P_{2h}(k) = P_{\rm lin}(k)\Big[\int \mathrm{d}M\, n(M)\,\frac{M}{\bar\rho}\,b(M)\,u(k\,|\,M)\Big]^2. \tag{11.71}$$

The **one-halo term** P_{1h} counts pairs of mass elements in the same halo; the **two-halo term** P_{2h} counts pairs in different haloes.

Example: the large-scale limit. For $k \to 0$ every $u \to 1$, and the bracket in P_{2h} is $\int \mathrm{d}M\, n\,(M/\bar\rho)\,b = 1$ by the consistency relation (11.69). So $P_{2h} \to P_{\rm lin}$: on large scales the model returns linear theory, as it must. This is where the consistency relation earns its place. If it failed, the halo model would predict a large-scale amplitude different from the one it was built on.

Example: the one-halo term is a shot noise. For $k \to 0$, $P_{1h} \to \int \mathrm{d}M\, n\, M^2/\bar\rho^2$, a constant. To see what it is, suppose all haloes had one mass M_0. Then $n\,M_0 = \bar\rho$, and $P_{1h} = M_0^2 n/\bar\rho^2 = 1/n$: exactly the Poisson shot noise $1/\bar n$ of a catalogue of points, (11.8), with the haloes as the points. Matter made of discrete lumps has the shot noise of the lumps. For the Sheth–Tormen mass function the constant is $342.9\,(h^{-1}\mathrm{Mpc})^3$, the shot noise of a population of haloes all of mass $\bar\rho \times 342.9 \approx 3\times10^{13}\,h^{-1}M_\odot$, the mass-weighted mean halo mass.

11.10.3 Computing it, and comparing with simulations

`10_halo_model.py` evaluates (11.71) at $z = 0$ on a grid of 500 masses from 1.000×10^7 to $3\times10^{15}\,h^{-1}M_\odot$ and 300 wavenumbers, with the CAMB linear spectrum, Sheth–Tormen abundances and biases, and the NFW transform with the concentrations (11.61).

Listing 11.10: The one- and two-halo terms. (`code/ch11/10_halo_model.py`)

```
w1 = n_st * (M / RHO_M) ** 2                        # weight per dlnM of the one-halo term
P1h = np.trapezoid(w1[:, None] * u ** 2, lnM, axis=0)
I2 = np.trapezoid((n_st * M / RHO_M * b_st)[:, None] * u, lnM, axis=0) + (1 - bias_int) *
    u[0]
P2h = P_lin * I2 ** 2
```

`w1` is the weight $n(M)(M/\bar\rho)^2$ per unit $\ln M$ (the array `n_st` already holds $\mathrm{d}n/\mathrm{d}\ln M$); `P1h` integrates it against u^2 over $\ln M$ for all k at once. `I2` is the bracket of P_{2h}, with one correction. The mass grid ends at $1.000\times10^7\,h^{-1}M_\odot$, but Sheth–Tormen keeps putting mass into ever smaller haloes: on our grid $\int n\,(M/\bar\rho)\,\mathrm{d}M = 0.592$ and $\int n\,(M/\bar\rho)\,b\,\mathrm{d}M = 0.7056$ instead of 1. The missing haloes are so small that $u = 1$ for them at every k we use, so their only role is to carry their share of the large-scale bias; the term `(1 - bias_int) * u[0]` adds that share to the smallest bin, which makes $P_{2h} \to P_{\rm lin}$ exactly at small k. Their one-halo contribution, weighted by M^2, is negligible.

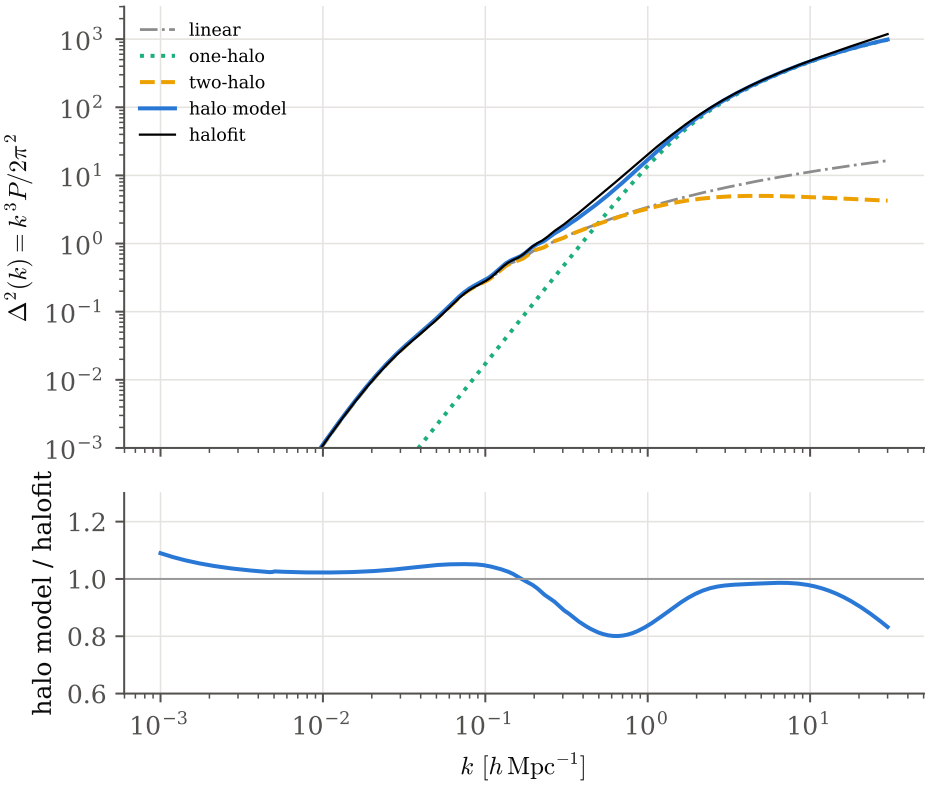


Figure 11.30: The matter power spectrum at $z = 0$ as $\Delta^2(k) = k^3P/2\pi^2$, the contribution of each logarithmic interval of k to the variance. The two-halo term (dashed) follows linear theory (dash-dotted) and is cut off when k resolves the haloes; the one-halo term (dotted) rises as k^3 and takes over beyond $k \approx 0.4957\,h\,\mathrm{Mpc}^{-1}$. Their sum (blue) is compared with halofit (black). Lower panel: the ratio of the two.

Side by side with halofit. Figure 11.30 shows the result. The two terms are equal at $k = 0.4957\,h\,\mathrm{Mpc}^{-1}$. The ratio of the halo model to halofit is 1.022 at $k = 0.01$, 1.047 at 0.1, 0.8937 at 0.3, 0.837 at 1, 0.9773 at 3 and 0.9775 at $10\,h\,\mathrm{Mpc}^{-1}$. Between $k = 0.05$ and 10 the worst point is 0.8007 at $k = 0.6311\,h\,\mathrm{Mpc}^{-1}$. A handful of numbers about haloes (a and p of the mass function, the amplitude and slope of the concentration relation, the halo edge and δ_c), each fixed by simulated or spherical-collapse haloes and none adjusted to the power spectrum, reproduce a fit to simulations to within 20% over three decades in k, and to about 2% between $k = 3$ and $10\,h\,\mathrm{Mpc}^{-1}$.

Which haloes make the small-scale power? At $k = 1\,h\,\mathrm{Mpc}^{-1}$ the one-halo term dominates, and the script measures how it is shared among masses: 10% of it comes from haloes below $4.381 \times 10^{13}\,h^{-1}M_\odot$ and 10% from haloes above 116×10^{13}. The non-linear power on the scale of a few megaparsecs is the internal structure of groups and clusters. That is why it is sensitive to σ_8 and to everything that changes the cluster abundance, and why baryonic processes inside clusters (gas blown out by active nuclei) change it measurably.

Where the halo model fails, and why. Each of the three disagreements traces back to one of the inputs.

- *The largest scales.* The one-halo term tends to a constant. Real matter has no such term: halo formation moves mass around locally and conserves it, and a local rearrangement of

mass cannot create power that stays constant as $k \to 0$ (it must vanish as k^4). The model's white noise adds 2% at $k = 0.01\,h\,\mathrm{Mpc}^{-1}$, more at smaller k (Figure 11.30, lower panel). The cause is (K1): the matter that went into a halo came from a surrounding region that is now underdense, and the model forgets the hole [CooraySheth2002 §4.4].

- *The transition,* $k \approx 0.3$–$1\,h\,\mathrm{Mpc}^{-1}$. Here the model is up to 20% low. The two-halo term uses linear P and linear bias (K3), with no halo exclusion and no quasi-linear growth of the field between haloes, and the one-halo term only counts the matter inside r_{vir}, not the infalling matter around it. Neither term is accurate where both are comparable. Research versions repair this region with fitted extra parameters [231].
- *The smallest scales.* At $k \gtrsim 3$ the answer depends on the concentration and on where the halo is cut. Keeping (11.61) but cutting haloes at $200\bar{\rho}$ instead of $330\bar{\rho}$ changes the ratio at $k = 10$ from 0.9775 to 0.6821. The concentration relation is calibrated together with a halo definition, and only the pair is meaningful [CooraySheth2002 §5.4].

The halo model is therefore a physical interpretation of the non-linear spectrum accurate at the 10–20% level, and a precise tool only after calibration. Its value is that every piece has a meaning that can be changed on purpose: massive neutrinos change $\sigma(M)$, baryons change $u(k\,|\,M)$, modified gravity changes δ_c, and the model says which scales respond.

11.10.4 Galaxies in haloes

Galaxies are not mass. Their correlation function is close to a power law, $\xi_{gg} \propto r^{-\gamma}$ with $\gamma \approx 1.8$, over two decades in separation, while the matter correlation function is not a power law at all [CooraySheth2002 §6]. Why should galaxies obey so simple a law when the matter does not? In the halo model the answer comes from one more ingredient. Every galaxy lives in a halo, so the galaxy field is the halo field reweighted by the number of galaxies each halo hosts:

> *If a halo of mass M hosts on average $\langle N\,|\,M\rangle$ galaxies, arranged in a known way, what are the galaxy power spectrum and correlation function?*

The occupation. A halo either has a galaxy at its centre or not; a central galaxy appears once the halo is massive enough to cool gas and form stars. Further galaxies, the satellites, orbit inside large haloes, distributed like the dark matter. A standard form [232] is

$$\langle N_c\,|\,M\rangle = \frac{1}{2}\Big[1 + \mathrm{erf}\Big(\frac{\log_{10} M - \log_{10} M_{\mathrm{min}}}{\sigma_{\log M}}\Big)\Big], \qquad \langle N_s\,|\,M\rangle = \langle N_c\,|\,M\rangle \Big(\frac{M - M_0}{M_1}\Big)^{\alpha}, \quad (11.72)$$

the **halo occupation distribution** (HOD). The central count is a Bernoulli variable (zero or one) with a smooth step in $\log M$; given a central, the number of satellites is Poisson with mean $\lambda(M) = [(M - M_0)/M_1]^{\alpha}$, and there are no satellites without a central. We take illustrative values, $\log_{10} M_{\mathrm{min}} = 12.8$, $\sigma_{\log M} = 0.5$, $\log_{10} M_0 = 12.6$, $\log_{10} M_1 = 14$ and $\alpha = 1.1$ (masses in $h^{-1}M_\odot$), roughly those of luminous galaxies (Figure 11.31, left).

The pairs inside one halo. The one-halo term of the matter came from pairs of mass elements in one halo, weighted by M^2. For galaxies it comes from pairs of distinct galaxies in one halo, so we need the mean number of ordered pairs, $\langle N(N-1)\,|\,M\rangle$. Write $N = C + S$ with $C \in \{0, 1\}$ the central and S the satellites:

$$\langle N(N-1)\,|\,M\rangle \overset{(1)}{=} \langle N_c\rangle\ \mathbb{E}\big[(1+S)S\,|\,C=1\big] \overset{(2)}{=} \langle N_c\rangle\,\big(\lambda + \lambda + \lambda^2\big) \overset{(3)}{=} 2\langle N_s\rangle + \langle N_c\rangle\lambda^2. \quad (11.73)$$

(1) Without a central there are no galaxies and no pairs; with one, $N = 1 + S$, and this happens with probability $\langle N_c \rangle$.
(2) $\mathbb{E}[S] = \lambda$ and $\mathbb{E}[S^2] = \lambda + \lambda^2$ for a Poisson variable (Chapter 2).
(3) $\langle N_s \rangle = \langle N_c \rangle \lambda$.

The two terms are different kinds of pair. $2\langle N_s \rangle$ counts the ordered pairs of the central with each satellite: one member sits at the centre and the other is displaced by the profile, so in Fourier space such a pair carries one factor $u(k \mid M)$. $\langle N_c \rangle \lambda^2$ counts the ordered pairs of two satellites, both displaced independently, which carry u^2 [CooraySheth2002 pp. 65–66]. A Monte Carlo of 2×10^6 haloes of $10^{14}\, h^{-1} M_\odot$ ($\langle N_c \rangle = 0.9997$, $\lambda = 0.9541$) gives $\langle N(N-1) \rangle = 2.818 \pm 0.002864$ against 2.818 from (11.73).

The galaxy power spectrum. Replacing the mass weights $M/\bar{\rho}$ of (11.70) by galaxy counts divided by the mean galaxy density $\bar{n}_g = \int \mathrm{d}M\, n(M) \langle N \mid M \rangle$, and the self-pairs by the pair counts just derived, gives

$$P_{gg}^{1h}(k) = \frac{1}{\bar{n}_g^2} \int \mathrm{d}M\, n(M) \left[2\langle N_s \mid M \rangle\, u(k \mid M) + \langle N_c \mid M \rangle \lambda^2(M)\, u^2(k \mid M) \right],$$
$$P_{gg}^{2h}(k) = P_{\mathrm{lin}}(k) \left[\frac{1}{\bar{n}_g} \int \mathrm{d}M\, n(M)\, b(M) \left(\langle N_c \mid M \rangle + \langle N_s \mid M \rangle\, u(k \mid M) \right) \right]^2, \tag{11.74}$$

[CooraySheth2002 eqs. (128)–(129)]. The pair of a galaxy with itself, which we excluded, is the Poisson shot noise $1/\bar{n}_g$ of (11.8), added on top when a catalogue is analysed. At small k the two-halo bracket becomes the **galaxy bias** $b_g = \bar{n}_g^{-1} \int \mathrm{d}M\, n\, b\, \langle N \mid M \rangle$, the number-weighted mean bias of the host haloes, so $P_{gg} \to b_g^2 P_{\mathrm{lin}}$ [CooraySheth2002 eqs. (130)–(131)].

Example: what the illustrative HOD predicts. The script gives $\bar{n}_g = 0.00116\, h^3 \mathrm{Mpc}^{-3}$, a satellite fraction of 0.1287, and $b_g = 1.298$. The mean host mass, weighted by galaxy number, is $4.112 \times 10^{13}\, h^{-1} M_\odot$. The one-halo term tends to $462.2\, (h^{-1}\mathrm{Mpc})^3$ at small k, on top of the Poisson shot noise $1/\bar{n}_g = 861.8$ of the catalogue itself: galaxies sharing a halo add pairs that a Poisson sample of a smooth field would not have, and this extra white noise is about half the size of the Poisson term. This is the excess over the input (G2) of Section 11.1.1, "no memory between places", which holds between haloes but not inside one. The correlation function, the Fourier transform of P_{gg}, is close to a power law with $\gamma = 1.917$ and correlation length $r_0 = 7.161\, h^{-1}\mathrm{Mpc}$ between 0.2 and $10\, h^{-1}\mathrm{Mpc}$ (Figure 11.31, right), while the matter curve bends.

Why a power law? Below about $1\, h^{-1}\mathrm{Mpc}$ almost all pairs of galaxies sit in the same halo, and the one-halo term, steep because it comes from the inner profiles of massive haloes, sets ξ_{gg}. Above a few megaparsecs, pairs sit in different haloes and $\xi_{gg} \approx b_g^2 \xi_{\mathrm{lin}}$. Where the two hand over, the one-halo term is falling and the two-halo term is flattening, and for galaxies whose occupation is concentrated in massive haloes they happen to join smoothly. Change the occupation and the slope changes. The power law is not a law; it is a coincidence of two terms, and the HOD parameters are what a galaxy clustering measurement on small scales actually constrains [CooraySheth2002 p. 66].

11.11 Voids

A redshift survey of the nearby Universe shows galaxies arranged in walls and filaments around nearly empty regions tens of megaparsecs across. Bond, Kofman and Pogosyan called the

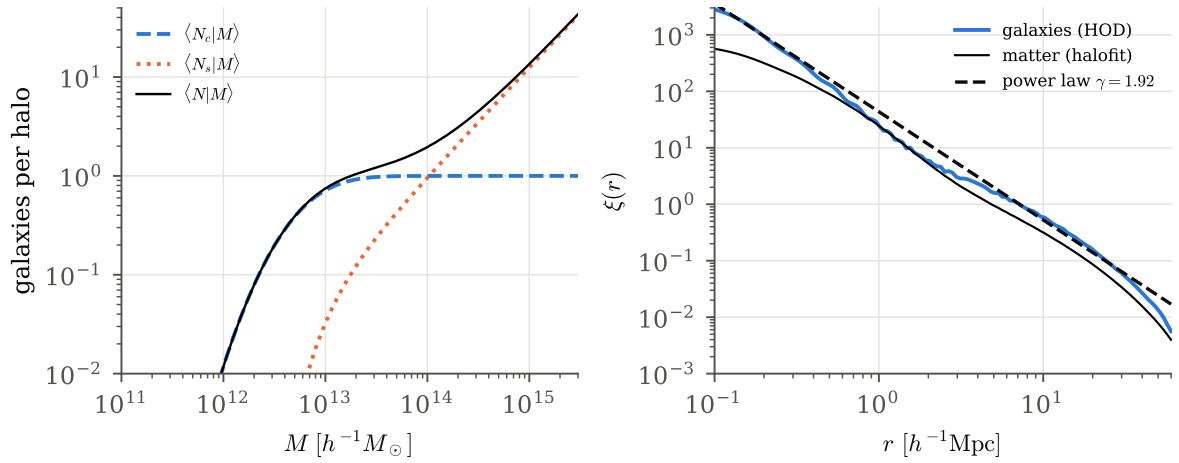


Figure 11.31: Left: the illustrative occupation (11.72): centrals (dashed) switch on near $10^{13}\,h^{-1}M_\odot$, satellites (dotted) grow almost in proportion to the halo mass. Right: the galaxy correlation function from (11.74) (blue) against the matter correlation function from halofit (black) and a power law (dashed). The galaxy curve is steeper than the matter on small scales, where it is all one-halo pairs, and joins the scaled matter curve on large scales.

arrangement the **cosmic web** and showed that it is already present, in embryonic form, in the initial Gaussian field: gravity only sharpens the image [233]. Clusters sit at the nodes, the filaments join them, and the space in between, most of the volume, belongs to the voids [4]. In Section 11.7 the voids were where the Zel'dovich universe and its phase-scrambled twin differed most. To use them as data we need three things, which we take in turn:

(1) *What exactly counts as a void?* A rule that a computer can apply to a catalogue or a simulated field.

(2) *How many voids of each size should a universe contain?* The prediction that turns a void count into a likelihood.

(3) *What do voids look like on average, and what can their shape measure?* This is the subject of Section 11.12.

11.11.1 How an empty region evolves

Start with the simplest void, a uniform spherical region of linear contrast $\delta_{\rm lin} < 0$ (extrapolated to today) in an Einstein–de Sitter background. It is the spherical-collapse problem of Section 11.8.1 run backwards: the region expands faster than its surroundings and empties. The exact solution is again taken from cosmology. For an underdense sphere of initial comoving radius R_0 and linear contrast δ_0 today, its comoving radius R and the redshift are given, in terms of a parameter θ, by [CooraySheth2002 eq. (50)]

$$\frac{R}{R_0} = \frac{(1+z)}{(5/3)|\delta_0|}\,\frac{\cosh\theta - 1}{2}, \qquad \frac{1}{1+z} = \left(\frac{3}{4}\right)^{2/3}\frac{(\sinh\theta - \theta)^{2/3}}{(5/3)|\delta_0|}. \tag{11.75}$$

What are the actual and the linear contrast at a given θ? The mass inside the sphere is fixed, so its actual mean contrast is $1 + \Delta = (R_0/R)^3$; its linear contrast at that time is δ_0 times the growth

factor, which in Einstein–de Sitter is $1/(1+z)$. Then

$$\delta_{\rm lin} \overset{(1)}{=} \frac{\delta_0}{1+z} \overset{(2)}{=} -\frac{3}{5}\Big(\frac{3}{4}\Big)^{2/3}(\sinh\theta-\theta)^{2/3},$$
$$1+\Delta \overset{(3)}{=} \Big[\frac{2\,(5/3)|\delta_0|}{(1+z)(\cosh\theta-1)}\Big]^3 \overset{(4)}{=} \Big[2\Big(\frac{3}{4}\Big)^{2/3}\frac{(\sinh\theta-\theta)^{2/3}}{\cosh\theta-1}\Big]^3 = \frac{9}{2}\,\frac{(\sinh\theta-\theta)^2}{(\cosh\theta-1)^3}. \tag{11.76}$$

(1) Linear growth in Einstein–de Sitter is $D \propto a = 1/(1+z)$, normalised so that $D = 1$ today.
(2) Insert $1/(1+z)$ from (11.75), with $\delta_0 = -|\delta_0|$; the $|\delta_0|$ cancels.
(3) $1+\Delta = (R_0/R)^3$, with R/R_0 from (11.75).
(4) Again insert $1/(1+z)$; the $|\delta_0|$ cancels, and $8\,(3/4)^2 = 9/2$.

Figure 11.32 plots the curve. Two numbers on it define voids in practice. First, the conventional threshold $\Delta = -0.8$ is reached at $\delta_{\rm lin} = -2.781$. Second, a void whose density decreases outwards does not stay smooth: its inner shells, more underdense, expand faster than the outer ones and catch up with them, piling matter into a ridge at the edge. For a top hat this *shell crossing* happens at $\theta \approx 3.53$, where $\delta_{\rm lin} = -2.811$ and $\Delta = -0.8021$ [ShethvdWeygaert2004 §3.1]. The void has then become a distinct object with a wall, the counterpart of collapse for a halo, and the threshold

$$\delta_v = -2.81 \tag{11.77}$$

plays the role of δ_c. That is why void finders look for regions with $\Delta \approx -0.8$. Since the mass inside is fixed, a void with $\Delta = -0.8$ is larger than the Lagrangian region it came from by

$$\frac{R}{R_L} = (1+\Delta)^{-1/3} = 0.2^{-1/3} = 1.71, \tag{11.78}$$

[ShethvdWeygaert2004 eq. (7)].

Example: the same void in the Zel'dovich approximation. Our simulated voids live in a Zel'dovich box, whose dynamics differ from the exact spherical solution. For a uniform sphere of linear contrast $\delta_{\rm lin}$, the Zel'dovich displacement is radial and, from $\nabla\cdot\boldsymbol{\Psi} = -\delta_{\rm lin}$ (Section 11.5.6), equal to $\boldsymbol{\Psi} = -\frac{1}{3}\delta_{\rm lin}\,\boldsymbol{q}$, because $\nabla\cdot\boldsymbol{q} = 3$. A shell starting at q moves to $r = q(1-\delta_{\rm lin}/3)$, so

$$1+\Delta_{\rm ZA} \overset{(1)}{=} \Big(\frac{q}{r}\Big)^3 = \Big(1-\frac{\delta_{\rm lin}}{3}\Big)^{-3}, \qquad \Delta_{\rm ZA} = -0.8 \overset{(2)}{=} \delta_{\rm lin} = -3\,(5^{1/3}-1) = -2.13. \tag{11.79}$$

(1) Mass conservation inside the shell, as for the exact solution.
(2) Solve $(1-\delta/3)^{-3} = 0.2$: $1-\delta/3 = 5^{1/3}$.

The same formula with $\delta_{\rm lin} > 0$ makes an overdense sphere collapse to a point at $\delta_{\rm lin} = 3$, in place of $\delta_c = 1.686$. A Zel'dovich void empties too fast (it keeps its initial velocities, while real gravity decelerates the outflow), so it reaches $\Delta = -0.8$ at a smaller $|\delta_{\rm lin}|$. These thresholds will matter when we compare the counted voids with a prediction.

11.11.2 Finding voids

There is no unique definition of a void, only rules, and different rules find different objects. Two families are in use.

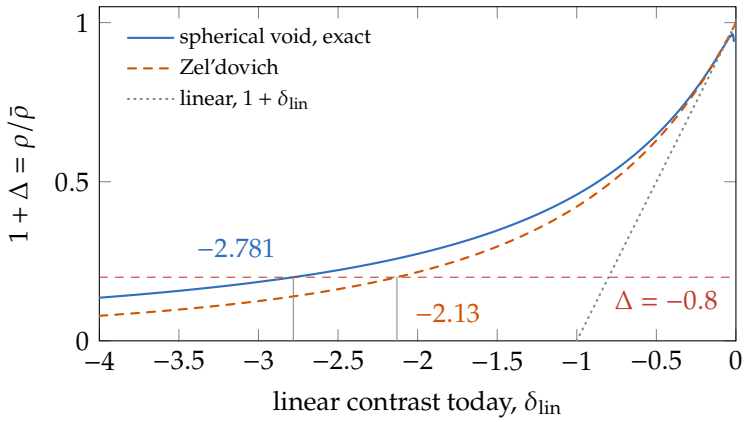


Figure 11.32: The mean density inside a spherical void against its linear contrast. Linear theory (dotted) reaches zero density at $\delta_{\rm lin} = -1$ and then goes negative; the exact solution (blue) empties smoothly and reaches 20% of the mean density at $\delta_{\rm lin} = -2.781$; the Zel'dovich approximation (orange, dashed) gets there earlier, at -2.13.

Working definition 11.7 (void (two operational definitions))

Spherical underdensity: a sphere about a local density minimum, grown until the mean contrast inside it, $\Delta(<R)$, first rises to a threshold, usually -0.8; overlapping spheres are pruned, largest first. The void radius is R. *Watershed:* treat the density as a landscape and flood it from every local minimum; a void is the basin of one minimum, ending on the density ridges that separate it from its neighbours [234, 235]. Its effective radius is that of the sphere of equal volume, $R = (3V/4\pi)^{1/3}$.

The spherical rule ties directly to the theory of Section 11.11.1: it finds exactly the regions that have reached $\Delta = -0.8$. The watershed rule needs no threshold and follows the real shapes of voids; it is what most research catalogues use, including the ZOBOV code behind the profiles of Section 11.12.2.

Listing 11.11: The spherical void finder. (`code/ch11/lib_web.py`)

```
def find_voids(rho, L, radii, smooth=1.0, seed_max=-0.5, threshold=-0.8, rmin=0.0):
    """Spherical underdensity void finder.

    rho      : 1 + delta on an N^3 periodic grid
    radii    : increasing trial radii [Mpc/h]
    Seeds are local minima of rho smoothed by a Gaussian of 'smooth' cells with delta <
    seed_max.
    About each seed the mean enclosed contrast Delta(<R) is read off sphere-averaged copies
    of
    rho; the void radius is where Delta(<R) first rises through 'threshold' (linear
    interpolation).
    Overlapping spheres are removed, largest first. Returns centres (n, 3) and radii (n,).
    """
    from scipy import ndimage
    N = rho.shape[0]
    cell = L / N
    rs = ndimage.gaussian_filter(rho, smooth, mode="wrap")
    mins = (rs == ndimage.minimum_filter(rs, size=3, mode="wrap")) & (rs - 1 < seed_max)
    idx = np.argwhere(mins)
    Dl = np.empty((len(idx), len(radii)))
    for j, R in enumerate(radii):
        Dl[:, j] = sphere_average(rho, R, L)[tuple(idx.T)] - 1
    rv = np.full(len(idx), np.nan)
    for i in range(len(idx)):
```

```
        above = np.nonzero(Dl[i] >= threshold)[0]
        if len(above) == 0 or above[0] == 0:
            continue
        j = above[0]
        d0, d1 = Dl[i, j - 1], Dl[i, j]
        rv[i] = radii[j - 1] + (threshold - d0) / (d1 - d0) * (radii[j] - radii[j - 1])
    ok = np.isfinite(rv) & (rv >= rmin)
    cen = (idx[ok] + 0.5) * cell
    rv = rv[ok]
    order = np.argsort(rv)[::-1]
    keep_c, keep_r = [], []
    for i in order:
        c = cen[i]
        if keep_c:
            d = np.array(keep_c) - c
            d -= L * np.round(d / L)
            if np.any(np.sqrt((d ** 2).sum(1)) < np.array(keep_r) + rv[i]):
                continue
        keep_c.append(c)
        keep_r.append(rv[i])
    return np.array(keep_c), np.array(keep_r)
```

The finder smooths the gridded density with a Gaussian of one cell and takes as seeds the cells that are lower than all 26 neighbours, with $\delta < -0.5$. For every trial radius it computes the mean density in a sphere about *every* cell at once, by a Fourier-space convolution (`sphere_average`); reading those maps at the seeds gives $\Delta(<R)$ for every seed and every radius. The void radius is where $\Delta(<R)$ first rises through the threshold, interpolated between trial radii. Finally the voids are sorted by size and each is kept only if it does not overlap a larger one already kept, with distances measured across the periodic boundary.

Example: the voids of the Zel'dovich box. On the Zel'dovich universe of Section 11.7 ($600\,h^{-1}$Mpc, 192^3 particles, cells of $3.1\,h^{-1}$Mpc) the finder returns 3607 voids larger than $8\,h^{-1}$Mpc, with median radius $9.732\,h^{-1}$Mpc and largest $23.16\,h^{-1}$Mpc. Together they fill only 8.482% of the box. That is not a contradiction of "voids fill most of the volume": the spherical rule keeps only the cores of the deepest regions, and the non-overlap condition discards everything between them. The watershed rule, run on the same field with the requirement that the basin floor be below $\delta = -0.8$, gives 7801 basins of median effective radius $14.68\,h^{-1}$Mpc, and the basins tile the box. The two catalogues describe the same field and agree on neither number nor size, which is why every void statistic must be predicted with the same rule that measures it.

11.11.3 How many voids: a walk between two barriers

The obvious guess, and why it is wrong. By symmetry with haloes, a point should belong to a void of mass M if its excursion-set walk first crosses δ_v, now from above, at $S = \sigma^2(M)$. That counts the *void-in-void* process correctly: a small void inside a larger one is part of the larger one, exactly as a small halo inside a larger one is. But voids have a second process with no counterpart for haloes [ShethvdWeygaert2004 §3.2]. A small underdense region inside a large region that is collapsing is crushed along with it. This is the *void-in-cloud* process. In walk language: a walk that crossed δ_c on some larger scale lies inside a halo, and any later crossing of δ_v is not a void. So

> *a point belongs to a void of mass M if its walk first crosses δ_v at $S = \sigma^2(M)$ without having crossed δ_c at any smaller S.*

This is a random walk between two absorbing barriers, δ_v below and δ_c above (Figure 11.33, right).

What fraction of the mass ends up in voids? Let $p(x)$ be the probability that a walk now at height x, between the barriers, reaches δ_v before δ_c. Over a small step ΔS the walk moves to $x+\epsilon$ with ϵ Gaussian of mean 0 and variance ΔS, and from there the same question starts afresh, so

$$p(x) \overset{(1)}{=} \mathbb{E}\big[p(x+\epsilon)\big] \overset{(2)}{=} p(x) + p'(x)\,\mathbb{E}[\epsilon] + \tfrac{1}{2}p''(x)\,\mathbb{E}[\epsilon^2] + \ldots \overset{(3)}{=} p(x) + \tfrac{1}{2}p''(x)\,\Delta S + \ldots$$
$$\Longrightarrow \quad p''(x) = 0 \quad \overset{(4)}{=} \quad p(x) = \frac{\delta_c - x}{\delta_c - \delta_v} \quad \Longrightarrow \quad p(0) = \frac{\delta_c}{\delta_c + |\delta_v|} \equiv 1 - D. \tag{11.80}$$

(1) Independent increments: after one step, the probability is that of a walk started from the new position, averaged over the step.
(2) Taylor expansion of p about x.
(3) $\mathbb{E}[\epsilon] = 0$, $\mathbb{E}[\epsilon^2] = \Delta S$; higher terms are of higher order in ΔS.
(4) A function with zero second derivative is a straight line; the boundary conditions $p(\delta_v) = 1$ and $p(\delta_c) = 0$ fix it.

This is the gambler's ruin of probability theory, and the result is [ShethvdWeygaert2004 eq. (3)], with the void-and-cloud parameter $D = |\delta_v|/(\delta_c + |\delta_v|)$. For $\delta_v = -2.781$ and $\delta_c = 1.686$, $1 - D = 0.3774$: a bit over a third of the mass ends up in voids, the rest in haloes. With $\delta_c = 1.06$, the threshold for turnaround rather than full collapse, it is 0.276. The script `11_voids.py` runs 40000 walks with both barriers, using `first_crossings` of Section 11.8.4 with its lower barrier switched on, and finds 0.3765 and 0.2752; the binomial standard error is 0.0024.

When do they cross? Now the distribution of the crossing scale. Let $p(\delta, S)$ be the density of walks that are at height δ at time S and have touched neither barrier. A walk at δ after $S + \Delta S$ was at $\delta - \epsilon$ at S; the same Taylor expansion as in (11.80), now in δ, gives the diffusion equation $\partial p/\partial S = \frac{1}{2}\partial^2 p/\partial\delta^2$. Walks that touch a barrier are removed, so $p = 0$ at $\delta = \delta_v$ and at $\delta = \delta_c$; all walks start at 0, so $p(\delta, 0) = \delta_D(\delta)$. Write $L = \delta_c - \delta_v$ for the gap between the barriers. Sine waves that vanish at both ends solve the equation, each decaying at its own rate:

$$p(\delta, S) \overset{(1)}{=} \sum_{j\ge 1} A_j \sin\!\Big(\frac{j\pi(\delta - \delta_v)}{L}\Big) e^{-j^2\pi^2 S/2L^2}, \qquad A_j \overset{(2)}{=} \frac{2}{L}\sin\!\Big(\frac{j\pi|\delta_v|}{L}\Big) = \frac{2}{L}\sin(j\pi D),$$
$$\mathcal{F}_v(S) \overset{(3)}{=} \frac{1}{2}\frac{\partial p}{\partial \delta}\Big|_{\delta_v} \overset{(4)}{=} \sum_{j\ge 1} \frac{j\pi}{L^2}\,\sin(j\pi D)\, e^{-j^2\pi^2 S/2L^2}. \tag{11.81}$$

(1) Separation of variables: $\sin(j\pi(\delta - \delta_v)/L)$ vanishes at both barriers, and its second derivative is $-(j\pi/L)^2$ times itself, so the diffusion equation makes it decay as $e^{-j^2\pi^2 S/2L^2}$.
(2) The coefficients of a delta function at $\delta = 0$ in this sine series (Fourier sine series on an interval of length L); $0 - \delta_v = |\delta_v|$ and $|\delta_v|/L = D$.
(3) The rate at which walks leave through the lower barrier is the probability current there. For the diffusion equation with coefficient $\frac{1}{2}$ the current is $-\frac{1}{2}\partial p/\partial\delta$, which at the lower wall points downward with magnitude $\frac{1}{2}\partial p/\partial\delta$.
(4) Differentiate the series term by term: $\partial_\delta \sin(j\pi(\delta - \delta_v)/L) = j\pi/L$ at $\delta = \delta_v$.

Multiplying by S and writing $1/L = D/|\delta_v|$ turns this into

$$S\,\mathcal{F}_v(S) = \sum_{j\ge 1} \frac{j^2\pi^2 D^2}{\delta_v^2/S}\,\frac{\sin(j\pi D)}{j\pi}\,\exp\!\Big(-\frac{j^2\pi^2 D^2}{2\,\delta_v^2/S}\Big), \tag{11.82}$$

which is [ShethvdWeygaert2004 eq. (1)]. The left panel of Figure 11.33 compares it with the simulated walks, for both values of δ_c, and with the single-barrier answer ($\delta_c \to \infty$, $D \to 0$, the void-in-void process alone). The two-barrier curves are cut off at small δ_v^2/S, that is at small voids: those are the walks most likely to have been inside a collapsing region first. At large voids the three agree, because a large underdense region is almost never embedded in a larger collapsing one [ShethvdWeygaert2004 p. 11].

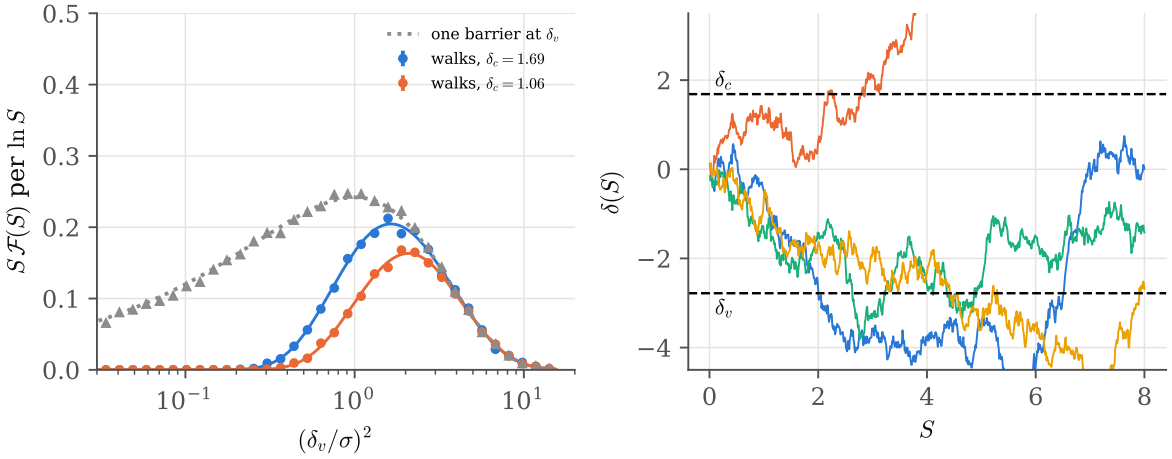


Figure 11.33: Left: the distribution of first down-crossings of δ_v that happen before any crossing of δ_c, per unit ln S, from 40000 simulated walks (points) and the series (11.82) (lines), for $\delta_c = 1.686$ and 1.06; grey: one barrier only. Right: four walks between the two barriers. A walk that hits δ_c first belongs to a halo, and its later dips are voids that have been crushed.

From walks to a size function. $S\mathcal{F}_v \,\mathrm{d}\ln S$ is the fraction of mass in voids whose Lagrangian mass has ln S in that interval. As for haloes, dividing by the mass of one void counts them; a void of Lagrangian radius R_L holds mass $\bar\rho V_L$, with $V_L = \frac{4}{3}\pi R_L^3$, and today has radius $R = 1.71\,R_L$. Hence the number of voids per unit volume and unit ln R,

$$\frac{\mathrm{d}n}{\mathrm{d}\ln R} = \frac{S\mathcal{F}_v(S)}{V_L}\left|\frac{\mathrm{d}\ln S}{\mathrm{d}\ln R_L}\right|, \qquad S = \sigma^2(R_L), \quad R_L = R/1.71, \tag{11.83}$$

[ShethvdWeygaert2004 eq. (6)]. This model has a flaw that Sheth and van de Weygaert point out themselves: every void is counted at the density of its Lagrangian volume, then inflated by $1.71^3 = 5$, so the total void volume can exceed the volume of the Universe [ShethvdWeygaert2004 p. 10]. Jennings, Li and Hu [236] proposed instead that voids, as they expand, merge rather than overlap, conserving the total volume in voids. That replaces V_L by the Eulerian volume $V = \frac{4}{3}\pi R^3$ in (11.83), the "volume-conserving" size function, a factor 1.71^3 lower.

Side by side with the Zel'dovich box. Figure 11.34 compares the voids found in the box with both models, computed with the CAMB $\sigma(R)$. With the thresholds of the dynamics that actually made the box, $\delta_v = -2.13$ and $\delta_c = 3$ from (11.79), the counts lie between the two models: over radii 9–19 h^{-1}Mpc Sheth–van de Weygaert is a factor 2.751 too high (median ratio) and the volume-conserving model a factor 0.5502 too low. With the exact spherical thresholds, $\delta_v = -2.781$ and $\delta_c = 1.686$, both predictions drop steeply at large R; at $R = 20\,h^{-1}$Mpc the Sheth–van de Weygaert curve is 51.83 times lower than with the Zel'dovich thresholds. The void abundance is exponentially sensitive to the threshold, for the same reason as the cluster abundance: large voids are rare events in a Gaussian field. That sensitivity is what makes the size function a probe of cosmology, and what makes it fragile: a threshold that does not

match the finder and the dynamics shifts the prediction more than any change of cosmological parameters. This is why research analyses calibrate the size function on simulations processed with the same finder [237].

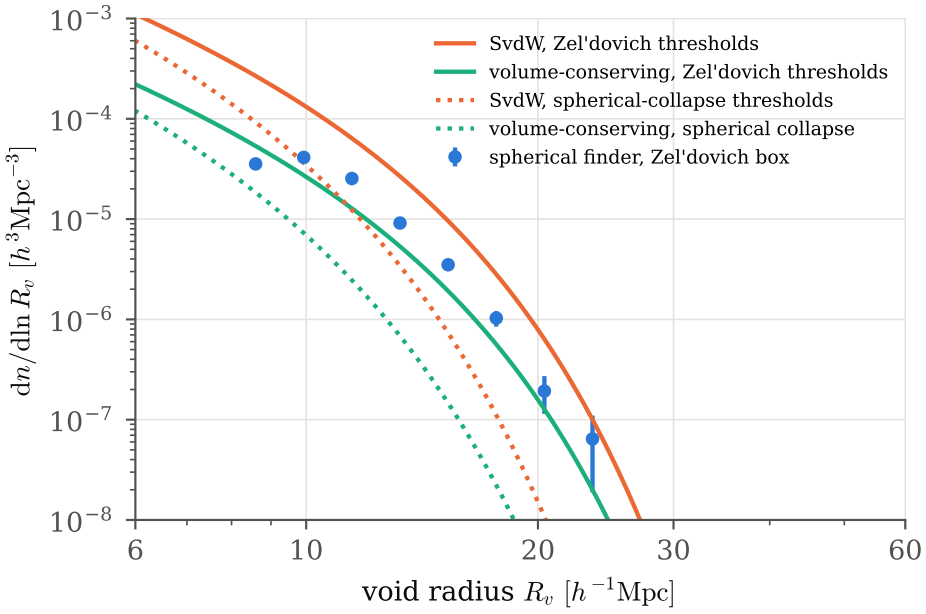


Figure 11.34: The void size function of the Zel'dovich box (points, Poisson errors) against the two-barrier excursion-set models: Sheth–van de Weygaert (orange) and volume-conserving (green), with the thresholds of the Zel'dovich dynamics (solid) and of exact spherical evolution (dotted). Over 9–19 h^{-1}Mpc the measured counts lie between the two solid curves; the smallest bin, next to the finder's minimum radius, falls below both.

Key idea: Voids from two barriers

A point is in a void of Lagrangian mass M if its excursion-set walk first crosses $\delta_v \approx -2.8$ at $S = \sigma^2(M)$ without first crossing δ_c. The mass fraction in voids is $\delta_c/(\delta_c + |\delta_v|)$, and the crossing distribution is the sine series (11.82). Converting it into numbers of voids of each radius needs an assumption about how voids expand, and the result is exponentially sensitive to the thresholds.

11.12 The average void: profile, outflow and shape

A single void is an irregular hole with a lumpy wall. Average a few thousand of them, each centred on its own centre and scaled by its own radius, and the irregularities wash out. Nothing in a homogeneous, isotropic universe picks out a direction, so the stacked void must be spherical, and its density and velocity depend only on the distance from the centre in units of the void radius. That average object is useful twice. Its radial profile records how matter has streamed out of underdense regions, which depends on gravity and on the growth rate. Its *shape*, if we observe it with the wrong distances, is not spherical, which turns the stack into a geometric ruler. We need the profile first.

11.12.1 Matter streams out of voids

The question. At distance r from the centre of a spherical underdense region, how fast does matter move away from it?

Linear theory. The linear continuity equation, $\dot\delta + \nabla\cdot\boldsymbol{v}/a = 0$ ((11.38)), relates the growth of the density contrast to the divergence of the peculiar velocity $\boldsymbol{v}$. Integrate it over a sphere of comoving radius r about the centre, and let $\Delta(r) = \frac{3}{r^3}\int_0^r \delta(r')\,r'^2\mathrm{d}r'$ be the mean contrast inside:

$$\frac{4\pi r^3}{3}\,\dot\Delta \overset{(1)}{=} \int_{|\boldsymbol{x}|<r} \dot\delta\,\mathrm{d}^3x \overset{(2)}{=} -\frac{1}{a}\oint_{|\boldsymbol{x}|=r} \boldsymbol{v}\cdot\mathrm{d}\boldsymbol{A} \overset{(3)}{=} -\frac{4\pi r^2}{a}\,v_r(r),$$
$$v_r(r) \overset{(4)}{=} -\frac{a}{3}\,r\,\dot\Delta \overset{(5)}{=} -\frac{1}{3}\,f\,aH\,r\,\Delta(r). \tag{11.84}$$

(1) The integral of δ over the sphere is its volume times Δ; the comoving sphere has fixed radius, so the time derivative acts on Δ alone.
(2) The continuity equation, and the divergence theorem.
(3) By spherical symmetry the radial velocity is the same all over the sphere.
(4) Solve for v_r.
(5) In linear theory every contrast grows as $D(t)$, so $\dot\Delta = (\dot D/D)\Delta = fH\Delta$ with $f = \mathrm{d}\ln D/\mathrm{d}\ln a$.

At $z = 0$, $a = 1$ and this is [Hamaus2014 eq. (4)] with $f \approx \Omega_m^{0.55}$. Inside a void $\Delta < 0$, so $v_r > 0$: matter flows out, faster where the void is deeper and larger. With r in h^{-1}Mpc and $H_0 = 100\,h\,\mathrm{km\,s^{-1}Mpc^{-1}}$, the factor $fH_0 r$ is in $\mathrm{km\,s^{-1}}$; for our cosmology $f = 0.5273$ at $z = 0$.

Zel'dovich dynamics. Linear theory is not trustworthy where $\Delta \approx -0.8$. In the Zel'dovich approximation a particle moves as $\boldsymbol{x} = \boldsymbol{q} + D\boldsymbol{\Psi}_0$, so its peculiar velocity is $a\dot{\boldsymbol{x}} = a\dot D\boldsymbol{\Psi}_0 = f\,aH\,\boldsymbol{\Psi}$. For a spherical void the displacement is radial, and mass conservation fixes its size: the shell that is now at r started at q with $q^3 = (1+\Delta)r^3$, so $\Psi_r = r - q = r[1-(1+\Delta)^{1/3}]$, and

$$v_r^{\mathrm{ZA}}(r) = f\,aH\,r\left[1-(1+\Delta(r))^{1/3}\right], \tag{11.85}$$

which reduces to (11.84) for small Δ, since $(1+\Delta)^{1/3} \approx 1 + \Delta/3$. For $\Delta = -0.8$ the bracket is $1 - 0.2^{1/3} = 0.42$, against 0.27 in linear theory: the linear formula underestimates the outflow from deep voids by a third.

11.12.2 Reproducing Hamaus, Sutter and Wandelt

Their question. Halo profiles have a universal shape, NFW; do void profiles? In 2014 Hamaus, Sutter and Wandelt [238] asked for a simple formula that describes the stacked density profile of voids of every size, out to several void radii, including the overdense wall around the void, so that void statistics could be modelled analytically rather than read off simulations.

Their setup. They used an N-body simulation of 2048^3 dark matter particles in a box $1\,h^{-1}$Gpc on a side with Planck cosmology, subsampled to $\bar n = 2\times10^{-2}\,h^3\mathrm{Mpc}^{-3}$, about the density of a galaxy survey. Voids were found by the watershed code ZOBOV, keeping zones below $\Delta = -0.8$; each void has an effective radius r_v (the radius of the sphere of the same volume), only voids larger than twice the mean particle separation were used, and the profiles were measured in shells out to $3r_v$ and averaged in eight bins of void radius [Hamaus2014 p. 1].

Their key equation. The formula they proposed is [Hamaus2014 eq. (2)]:

$$\frac{\rho_v(r)}{\bar\rho} - 1 = \delta_{\mathrm{cen}}\,\frac{1-(r/r_s)^\alpha}{1+(r/r_v)^\beta}. \tag{11.86}$$

It is an empirical fit, not a derived law, but every parameter can be read off from the limits of the formula, and that is the derivation available. At $r \to 0$ the fraction is 1, so δ_{cen} is the

central contrast. At $r = r_s$ the numerator vanishes: r_s is where the density crosses the mean, the inner edge of the wall. For $r_s < r \lesssim r_v$ the numerator is negative, so the profile is positive, the wall, with α setting how steeply it rises. For $r \gg r_v$ the profile behaves as $-\delta_{\rm cen}(r/r_s)^\alpha(r_v/r)^\beta$, which goes to zero only if $\beta > \alpha$; β sets how fast the wall fades into the mean. The total missing mass, $\int_0^\infty 4\pi r^2 \bar\rho\, \delta(r)\, {\rm d}r$, converges only if the integrand, $\propto r^{2+\alpha-\beta}$ at large r, falls faster than $1/r$, that is if $\beta > \alpha + 3$ [Hamaus2014 eqs. (7)–(8)]. Its sign splits voids into *overcompensated* ones, whose wall holds more than the missing mass (small voids, embedded in collapsing regions), and *undercompensated* ones (large voids, embedded in larger underdense regions). Hamaus et al. also found that α and β follow r_s/r_v [Hamaus2014 eqs. (10)–(11)]: $\alpha \simeq -2(r_s/r_v - 2)$, and $\beta \simeq 17.5\, r_s/r_v - 6.5$ for $r_s/r_v < 0.91$ and $\beta \simeq -9.8\, r_s/r_v + 18.4$ above. The velocity profile they compare with is (11.84), which we derived.

Watch out: Two different "δ_c"

Hamaus et al. call the central void contrast δ_c. It has nothing to do with the collapse threshold $\delta_c = 1.686$ of Section 11.8. We write it $\delta_{\rm cen}$.

What we compute. `11_voids.py` runs the watershed finder of Section 11.11.2 on the Zel'dovich box, which gives 7801 voids, and for each void counts the particles in 30 shells out to $3R_v$ (using a k-d tree with periodic boundaries) and divides by the expected count. It stacks the profiles in four bins of radius, 10–13, 13–16, 16–20 and 20–30 h^{-1}Mpc, with the standard error of the mean as the error bar (never smaller than the Poisson error of the stacked count). It fits (11.86) to each stack with four free parameters, by least squares, ignoring the innermost shells $r < 0.15R_v$. For the velocities, each particle's radial velocity is $fH_0\Psi_r$, and the stacked mean is compared with (11.84) and (11.85), both evaluated with the measured $\Delta(r)$ of the stack.

Side by side. Figure 11.35 shows the stacks and the fits, and Table 11.2 lists the parameters next to the values that the empirical relations for $\alpha(r_s/r_v)$ and $\beta(r_s/r_v)$ quoted above would assign at our r_s/R_v.

$\bar R_v$	N	$\delta_{\rm cen}$	r_s/R_v	α	β	χ^2/dof	α (H14)	β (H14)
11.81	1780	−0.9515	1.065	2.088	25 (bound)	1.302e + 04/25	1.87	7.961
14.48	3165	−0.9162	0.9493	2.35	25 (bound)	3451/25	2.101	9.097
17.5	2245	−0.91	0.9238	2.381	11.08	352/25	2.152	9.347
21.31	313	−0.9216	0.9442	2.618	9.128	38.54/25	2.112	9.147

Table 11.2: Fits of (11.86) to the stacked watershed voids of the Zel'dovich box ($\bar R_v$ in h^{-1}Mpc). The last two columns are the values of α and β that the relations of [Hamaus2014 eqs. (10)–(11)] give for the fitted r_s/R_v. "Bound": the fit ran into the upper limit $\beta = 25$ allowed in the fit.

For the two stacks of large voids the formula gets the shape roughly right. The largest stack is fitted with $\chi^2 = 38.54$ for 25 degrees of freedom, with $r_s/R_v = 0.9442$ close to the compensation value 0.91 of [Hamaus2014 p. 3], and α and β near the values Hamaus's relations assign. The next stack, with more voids and so smaller errors, already has $\chi^2 = 352$: close is not the same as right. The central contrasts, −0.91 to −0.95, are like theirs (−0.95 for their voids of 11.7 h^{-1}Mpc, [Hamaus2014 Fig. 3]). For the two stacks of small voids it fails: β runs to the edge of its allowed range and the χ^2 is thousands. The reason is visible in the left panel. Our small voids have almost no wall: the highest density outside $0.8R_v$ is −0.08312 for the smallest stack, against 0.1142 for the largest, and their density approaches the mean slowly from below out to $3R_v$. Hamaus's small voids have the highest walls of all.

Why they differ. Two causes, both in our simulation rather than in the formula. First, the Zel'dovich approximation has no gravity after the initial kick: particles that meet in a wall do not stop there but stream through it, so walls, which in a real universe are built by matter piling up and staying, are thin and short-lived. The small overcompensated voids of an N-body simulation are surrounded by just such walls, collapsing around them (the void-in-cloud process of Section 11.11.3). Second, the watershed of a smoothed Zel'dovich field splits large underdense regions into several basins, so many of our small "voids" are sub-basins of a larger void, undercompensated by construction. Only large voids, whose walls formed from large-scale linear modes that Zel'dovich treats well, look like the N-body ones. The errors of the stacks are also small (thousands of voids, millions of particles), so the χ^2 detects any mismatch of shape.

The outflow. The right panel of Figure 11.35 shows the stacked radial velocities. For the largest stack the peak outflow is 251.7 km s^{-1}, against 267 from the spherical Zel'dovich formula (11.85) and 181.7 from linear theory (11.84). The linear formula is too low by about a third inside the void, close to the factor estimated after (11.85). The Zel'dovich formula comes within a few per cent, about as close as the few hundred voids of one stack allow. That agreement is partly a consistency check, since the particle velocities in our box are Zel'dovich velocities; it confirms that the stacked void behaves like a spherical one.

The lesson. A fitting formula is universal only within the dynamics that produced it: our large voids follow the Hamaus profile, our small ones do not, because a Zel'dovich universe cannot build the walls that make small voids overcompensated.

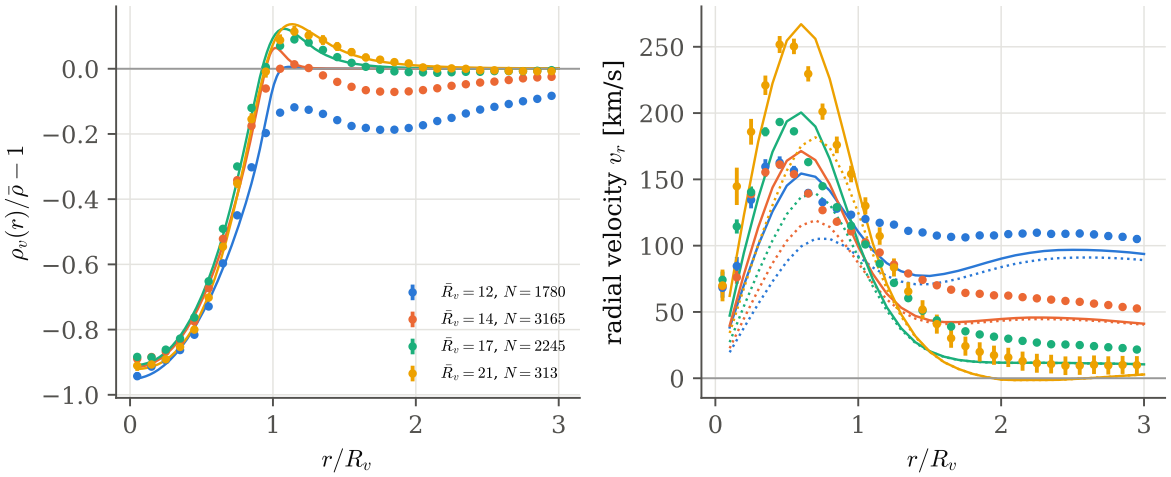


Figure 11.35: Stacked watershed voids of the Zel'dovich box in four bins of radius. Left: density profiles (points) and fits of (11.86) (lines); only the larger voids have an overdense wall near R_v. Right: radial velocity profiles (points) with the spherical Zel'dovich prediction (solid) and linear theory (dotted), both from the measured enclosed contrast.

11.12.3 Voids as standard spheres

The question. In a homogeneous, isotropic universe the stacked void is a sphere. A survey does not measure distances; it measures angles and redshifts, and converts them with a fiducial cosmology (Section 11.4.1). If the fiducial cosmology is wrong, does the stacked void still look round? Lavaux and Wandelt [239] proposed using exactly this, a test of sphericity first suggested by Alcock and Paczyński for any population of round objects.

The geometry. Write $\alpha_\perp = D_M/D_M^{\rm fid}$ and $\alpha_\parallel = D_H/D_H^{\rm fid}$ for the ratios of the true to the fiducial transverse and radial distances, as in (11.26). A sphere of true radius R_0 appears with transverse radius $R_0/\alpha_\perp$ and radial radius $R_0/\alpha_\parallel$, so its apparent axis ratio is

$$q \equiv \frac{R_\parallel^{\rm fid}}{R_\perp^{\rm fid}} = \frac{\alpha_\perp}{\alpha_\parallel} = \frac{D_M/D_M^{\rm fid}}{D_H/D_H^{\rm fid}} = \frac{D_M(z)H(z)}{[D_M(z)H(z)]^{\rm fid}}, \tag{11.87}$$

using $D_H = c/H$. The size of the void depends on the unknown true radius, but the ratio does not: a population of spheres measures the product $D_M H$ without any ruler of known length [LavauxWandelt2012 eqs. (5)–(9)]. This is the **Alcock–Paczyński test**.

What curve do we fit? A sphere of radius R_0 stretched by q along the line of sight has points $(r_\perp, r_\parallel)$ with $r_\perp^2 + (r_\parallel/q)^2 = R_0^2$. In polar form, $r_\perp = r\sqrt{1-\mu^2}$ and $r_\parallel = r\mu$, with μ the cosine of the angle to the line of sight:

$$r^2\left(1 - \mu^2 + \frac{\mu^2}{q^2}\right) = R_0^2 \quad\Longrightarrow\quad r(\mu) = \frac{R_0}{\sqrt{1 - \mu^2\,(1 - 1/q^2)}}. \tag{11.88}$$

The experiment. `12_void_ap.py` takes the 1530 spherical voids with $10 \le R_v < 25\,h^{-1}$Mpc from Section 11.11.2, and stacks the particles around them in bins of r/R_v and $|\mu|$, with the box's z axis as the line of sight. In each of ten μ bins it finds the radius where the stacked density rises through -0.5, and fits (11.88) to those ten radii. The error of q comes from a **jackknife**, a resampling estimate in the spirit of the bootstrap of Section 3.15: cut the box into $G = 27$ sub-cubes and put each void in the group of the sub-cube that holds its centre, repeat the whole measurement G times leaving out one group at a time to get $q_1, \dots, q_G$, and take $\sigma_q^2 = \frac{G-1}{G}\sum_g (q_g - \bar q)^2$. The factor $G-1$ compensates for the leave-one-out values being much closer to each other than independent measurements would be, since any two of them share $G-2$ of the G groups. The groups are regions of space, not voids picked at random, because neighbouring voids sit in the same large-scale waves and are not independent; leaving out a whole region removes those waves too. Even so, every region lies in the same box, and the jackknife cannot see waves longer than the box. The error is a rough one.

The script measures q four times: in real space; with the line of sight stretched by a true $q = 1.05$, as a wrong fiducial cosmology would do; in redshift space, moving each particle by its line-of-sight velocity, $s_z = z + f\Psi_z$ (Section 11.5.1); and with both.

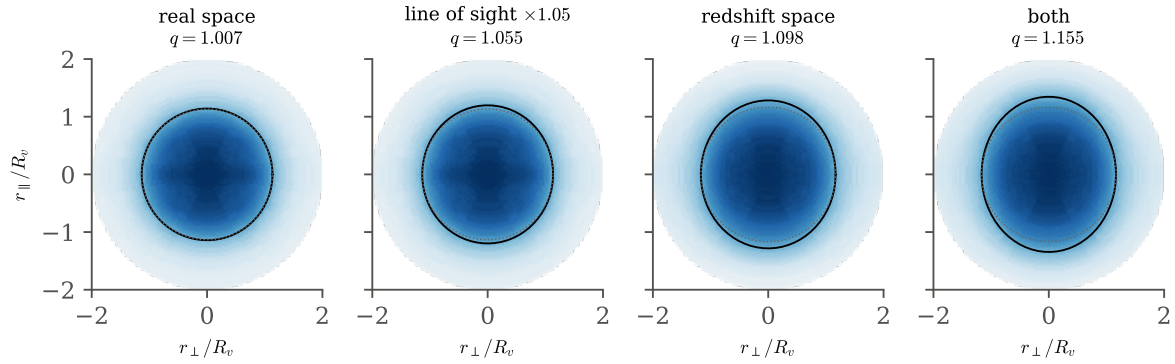


Figure 11.36: The stacked void in the plane of the line of sight (vertical) and a transverse direction, coloured by density (blue empty, red dense), with the fitted ellipse (black) and a circle (dotted). From left to right: real space; real space with distances along the line of sight stretched by 1.05; redshift space; both. Stretching and redshift-space outflow both elongate the stack.

What it finds. In real space $q = 1.007 \pm 0.009846$: the stack is round, as isotropy demands. With the stretched line of sight $q = 1.055 \pm 0.01172$, which recovers the input 1.05. About fifteen hundred voids in $0.2\,(h^{-1}\mathrm{Gpc})^3$ measure the axis ratio to about 1.1%. But in redshift space, with the correct geometry, $q = 1.098 \pm 0.01035$, and with both effects $q = 1.155 \pm 0.01141$, close to the product $1.055 \times 1.098 = 1.159$. The outflow of matter from voids elongates them along the line of sight by more than the geometric signal we inserted.

Why the outflow elongates the void. Redshift space moves every particle along the line of sight by $f\Psi_z$. Near a spherical void the displacement is radial, $\mathbf{\Psi} = \Psi_r(r)\,\hat{\mathbf{r}}$, so $\Psi_z = \Psi_r(r)\,z/r$. The density in redshift space is the real density divided by the Jacobian of the map, $1 + f\,\partial\Psi_z/\partial z$ (Section 11.5.2), and

$$\frac{\partial \Psi_z}{\partial z} \overset{(1)}{=} \Psi_r'(r)\,\frac{z}{r}\,\frac{\partial r}{\partial z} + \Psi_r(r)\left(\frac{1}{r} - \frac{z}{r^2}\frac{\partial r}{\partial z}\right) \overset{(2)}{=} \mu^2\,\Psi_r'(r) + (1-\mu^2)\,\frac{\Psi_r(r)}{r}. \tag{11.89}$$

(1) Product rule on $\Psi_r(r) \cdot z/r$.
(2) $\partial r/\partial z = z/r = \mu$.

Along the line of sight ($\mu = 1$) the shell at r moves out by $f\Psi_r$ and its density is divided by $1 + f\Psi_r'$. Across it ($\mu = 0$) nothing moves, but the density is still divided by $1 + f\Psi_r/r$, because the outflow stretches the line of sight around every point. Both make the void look emptier, and so larger, but by different amounts, so the void becomes elongated. Using the measured real-space stack and the spherical Zel'dovich Ψ_r, the script predicts that the -0.5 contour moves out by a factor 1.158 along the line of sight and 1.066 across it (linear theory: 1.131 and 1.049), an axis ratio of 1.086 (linear: 1.079), against the measured 1.098 ± 0.01035. The ratio is right. The two radii separately are both overpredicted by a few per cent (measured: 1.116 in the μ bin nearest the line of sight and 1.043 in the bin nearest the transverse direction), a common offset that the spherical model does not capture and that cancels in the ratio.

Example: what the test would measure. At $z = 0.51$, the mean redshift of the BOSS void sample, our fiducial flat ΛCDM has $D_M H/c = 0.5937$, and $\mathrm{d}\ln(D_M H)/\mathrm{d}\Omega_m = 0.3431$. An error of 0.01172 on $q \approx 1.05$ therefore corresponds to $\sigma(\Omega_m) \approx 0.01172/(1.05 \times 0.3431) = 0.03254$, for a stack of the size of ours (as rough as the jackknife error it comes from), if the redshift-space elongation were known exactly. The BOSS analysis of Hamaus et al. [240] fitted the redshift-space distortion (f/b, the growth rate over the galaxy bias) and the geometry at the same time, and found $D_M H/c = 0.588 \pm 0.004$ at $z = 0.51$ (they write D_A for our comoving D_M), with $f/b = 0.540 \pm 0.091$ and, in flat ΛCDM, $\Omega_m = 0.312 \pm 0.020$.

Key idea: Stacked voids as standard spheres

The stacked void is spherical in true coordinates, so its axis ratio in survey coordinates measures $D_M(z)H(z)$ relative to the fiducial value, with no ruler of known length. The outflow from voids elongates the stack along the line of sight by a comparable amount, so the geometry and the growth rate must be fitted together.

11.12.4 What voids are sensitive to

Voids are the regions where the matter is thinnest, and that makes them respond to physics that the dense regions hide [Pisani2019 §2].

- *Dark energy.* A void is the part of the Universe where dark energy dominates first. Its expansion is set by gravity, which empties it, and by the cosmic expansion, which

stretches it, so the size function depends on the dark-energy equation of state; and the Alcock–Paczyński test above measures $D_M H$ directly. The BOSS void size function alone gives $w = -1.1 \pm 0.2$ [237].

- *Modified gravity.* Many alternatives to general relativity add a force that is screened in dense environments, so that solar-system tests are passed. Voids are where screening fails: the extra force makes them expand faster and empty further, which shows up in their size function and in the outflow measured by redshift-space distortions around them.

- *Massive neutrinos.* Neutrinos stream freely over distances comparable to void sizes and do not follow the dark matter out of voids. Their share of the mass is therefore higher in voids than in dense regions, which changes how many voids form, how empty they become, and how they cluster.

Each of these is a change of the inputs of Section 11.11.3 ($\sigma(R)$, the thresholds, the expansion factor) or of the outflow (11.84) (through f), which is why the excursion-set and velocity models are the starting point of every void forecast.

11.12.5 Voids and the topology of the field

Does the way the regions of a field are connected carry information that the power spectrum does not expose? Topology is the tool made for that question, and the cosmic web has an exact topological description, which Chapter 12 builds from the beginning. Take the region where the density exceeds a threshold t, the superlevel set $\{\delta > t\}$. Its Betti numbers count, in three dimensions, its connected pieces (β_0), its independent tunnels (β_1) and the cavities it encloses (β_2) (see the topology notes). A void, seen from the matter, is a cavity: an empty region enclosed by walls. Voids are the β_2 features of the cosmic web [4, 5]. Equivalently, a cavity of the superlevel set is a separate connected piece of the complementary sublevel set $\{\delta < t\}$, and connected pieces are what a computer counts most easily.

`11_voids.py` counts them in the Zel'dovich box, after smoothing by one cell as for the watershed: the number of connected regions of $\{\delta_s < t\}$ (with the periodic faces glued, by a union–find over the labels on opposite faces), and the number of local minima below t, as t rises from -0.95 to 0 (Figure 11.37). Every region is born at a local minimum, so the minima are the births. At $t = -0.8$ there are 6414 regions and 9318 minima: most minima are still separate voids. The number of regions peaks at 6807 near $t = -0.75$ and then falls while minima keep appearing, because regions merge at saddle points of the density; by $t = -0.5$ there are 2257 regions for 18867 minima, and at $t = 0$ only 155. The underdense regions have joined into a network that runs through the box. That the separate voids exist only below $\delta \approx -0.8$ is the same threshold that came out of spherical evolution, and the one Wilding et al. find in N-body simulations, below which voids are the only topological features of the field [Wilding2021 p. 10].

The picture is the one that persistence formalises. Each underdense region is born at the value of its minimum and dies when it merges with an older one at a saddle; the pair (birth, death) is a point of a persistence diagram, and the watershed basins of Section 11.11.2 are the regions between those saddles. A void finder is therefore a particular, hand-tuned reading of the H_0 persistence of the sublevel sets. Chapter 12 turns that reading into a statistic with a sampling distribution and feeds it to the same likelihood machinery as $P(k)$.

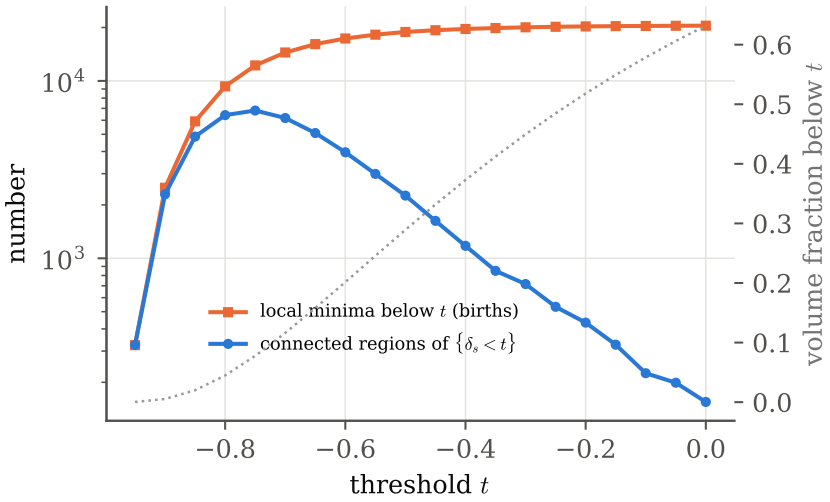


Figure 11.37: Underdense regions of the smoothed Zel'dovich field as the threshold t rises: the number of separate connected regions of $\{\delta_s < t\}$ (blue), the number of local minima below t (orange), and the volume fraction below t (dotted, right axis). Each region is born at a minimum; the gap between the curves counts the mergers at saddles.

11.13 Haloes and voids in research

Every measurement of haloes and voids we have made came from a script that runs in seconds to minutes on a laptop. The research versions keep the same ideas (first crossings, profiles, one- and two-halo terms, two barriers, stacked spheres) and replace almost every ingredient by something measured in large simulations.

Haloes. Research groups do not take the mass function from a formula. They run N-body simulations of 10^9–10^{11} particles in boxes of a few h^{-1}Gpc, identify haloes with a halo finder (linking particles closer than a fraction of the mean separation, or growing spheres to a fixed overdensity, the same choice of edge that mattered in Section 11.10.3), and fit or emulate $\mathrm{d}n/\mathrm{d}\ln M$ and $b(M)$ as functions of cosmology. The excursion set and the peak-background split survive as the reason the fits are nearly universal in ν and as a check on them. For cluster counts the hard part is not the mass function but the mass: clusters are found by their galaxies, hot gas or lensing signal, and the relation between that observable and M is calibrated with weak lensing, which is now the dominant uncertainty.

The non-linear power spectrum. The version of the halo model used in practice is HMcode [231], available in CAMB: the halo model of Section 11.10 with a handful of extra parameters fitted to simulations, which smooth the transition region where our version was 20% low and remove the large-scale one-halo excess. Because each parameter has a physical meaning, the same code models baryonic feedback by changing the concentration and the inner profiles of haloes, which is how weak lensing surveys marginalise over baryons. Halofit [41], which we compared with, plays the same role without the physical parameters, and emulators trained on suites of simulations are replacing both at the per-cent level.

Galaxies. The halo occupation distribution of Section 11.10.4 is fitted to the measured small-scale clustering of each galaxy sample, usually the projected correlation function, with the HOD parameters as nuisance parameters. The fitted HOD is then used to place galaxies in

the haloes of N-body or approximate simulations, which produces the thousands of mock catalogues from which the covariance matrices of Section 11.6 are estimated. So the halo model enters every survey analysis, if only through its mocks.

Voids. Void catalogues are built from the galaxies of spectroscopic surveys (BOSS, eBOSS, and now DESI, with Euclid to follow) with watershed codes descended from ZOBOV [235], run on the survey's light cone with its mask and selection. Galaxies are sparse and biased, so the voids are defined by the tracers, not by the matter, and their sizes depend on the tracer density; every prediction is therefore calibrated on mock catalogues processed by the same finder. The two leading analyses are the ones met above: the joint fit of the redshift-space distortions and the Alcock–Paczyński distortion of the stacked void–galaxy cross-correlation [240], and the void size function with a volume-conserving model extended by calibrated nuisance parameters [237].

How this is used in real research: Verification and research

Our simulations played two roles. The random walks of Sections 11.8.4 and 11.11.3 *verified* results we had derived exactly: the factor of 2, the Press–Schechter multiplicity, the gambler's-ruin void fraction and the two-barrier series. The Zel'dovich box was used as a stand-in for the *research* instrument, the place where no closed form exists: the void size function measured with a particular finder, the stacked profiles, the redshift-space elongation of voids, the topology of the underdense regions. In research that instrument is an N-body simulation with the survey's tracers, mask and finder, and the analytic models are calibrated against it, never the other way round.

11.13.1 What our shortcuts cost

Table 11.3 lists each simplification, what research does instead, and what it costs. As in Section 11.6.2, some costs are only wider error bars or a smaller range of scales; others leave a physical effect out, and no amount of laptop time in the same model brings it back.

The first row matters most. Almost everything that went wrong in our comparisons, the missing walls of small voids, the need for Zel'dovich thresholds in the size function, the close but partly circular agreement of the void outflow, comes from the Zel'dovich dynamics. A particle-mesh N-body code in a 256^3 box would run on the laptop in tens of minutes and would remove most of it; the full research version runs N-body simulations with 10^{10} particles on thousands of cores and identifies haloes and voids in hundreds of snapshots. What the laptop versions do transfer is the mechanism: why the mass function has the shape it has, where each part of the non-linear power spectrum comes from, why the galaxy correlation function is close to a power law, and why voids measure both geometry and growth.

Recap

- Two fields with the same $|\delta_{\boldsymbol{k}}|$ share $P(k)$, $\xi(r)$ and σ_R but can differ in skewness, one-point distribution and voids; the difference lives in the phases and is carried by haloes and voids (Section 11.7).
- With a sharp-k filter the smoothed linear contrast at a point is a Brownian motion in $S = \sigma^2(M)$; by reflection the fraction of walks that have crossed δ_c is $\mathrm{erfc}(\nu/\sqrt{2})$, twice the naive count, giving $F(\nu) = \sqrt{2/\pi}\, e^{-\nu^2/2}$ and $\mathrm{d}n/\mathrm{d}\ln M = (\bar{\rho}/M)\,\nu F|\mathrm{d}\ln\sigma/\mathrm{d}\ln M|$; simulations prefer Sheth–Tormen, the mark of ellipsoidal collapse (Section 11.8).

step	research	ours	what it costs
dynamics	N-body, 10^9–10^{11} particles, Gpc boxes	Zel'dovich particles, 192^3 in 600 h^{-1}Mpc	no bound haloes, walls dissolve after shell crossing; small voids lack their walls (Hamaus fit fails for $R_v \lesssim 15\,h^{-1}$Mpc); void thresholds are the Zel'dovich ones
haloes	found in simulations; mass function and bias fitted or emulated	analytic only (Press–Schechter, Sheth–Tormen)	abundance uncertain by tens of per cent at the extremes; no test of the peak-background split on simulated haloes
first crossings	correlated steps (real-space filters), moving barriers	sharp-k walks; 4 fields of 160^3 cells for the top hat	the factor of 2 is exact only for the sharp-k filter; the top-hat result has no closed form
non-linear $P(k)$	HMcode or emulators, with baryons	plain halo model, one concentration relation	0.8007 at $k = 0.6311\,h\,\mathrm{Mpc}^{-1}$; white-noise excess on large scales; no baryons
galaxies	HOD fitted to measured clustering; mocks	illustrative HOD, no fit	the power-law ξ_{gg} is a demonstration, not a measurement
void tracers	sparse, biased galaxies on the light cone, survey mask	all particles, periodic box, $z = 0$	voids smaller and emptier than in a survey; no mask or edge effects
void finder	ZOBOV/VIDE on tracers; calibrated on mocks	grid-based spherical and watershed finders, one-cell smoothing	size function differs from the two analytic models by factors of about 2–3
geometry	joint fit of AP and RSD on the void–galaxy cross-correlation	ellipse fitted to one contour; AP and RSD separately	$\sigma_q \approx 0.01172$ only for known RSD; no joint posterior
covariance	thousands of mocks	jackknife over 27 subcubes of one box, Poisson errors	error bars approximate; stacked-profile χ^2 not a calibrated test

Table 11.3: The steps of the halo and void analyses at research scale and in our scripts, with the cost of each simplification.

- Haloes have NFW profiles with $c(M)$ falling with mass; their transform $u(k\,|\,M)$ is 1 for long modes and falls once k resolves the halo. A long-wavelength δ_ℓ lowers the barrier to $\delta_c - \delta_\ell$, giving $b = 1 - \delta_c^{-1}\mathrm{d}\ln(\nu F)/\mathrm{d}\ln\nu$, with mass-weighted mean 1 (Section 11.9).
- The halo model $P = P_{1h} + P_{2h}$ follows from summing halo profiles: same-halo pairs give $\int n(M/\bar\rho)^2u^2$, different-halo pairs give $P_{\rm lin}[\int n(M/\bar\rho)bu]^2$. It reproduces halofit to 20%, fails in the transition region and on the largest scales, and with an HOD explains the near power-law ξ_{gg} (Section 11.10).
- A void is a region that reached $\Delta \approx -0.8$, the shell-crossing density of a spherical void ($\delta_v = -2.81$, radius 1.71 times its Lagrangian one). Voids are counted by walks crossing δ_v before δ_c: mass fraction $\delta_c/(\delta_c + |\delta_v|)$, crossing distribution a sine series; the size function needs an expansion model and is exponentially sensitive to the thresholds (Section 11.11).
- Stacked voids have a universal-looking profile where walls form, flow outwards as

$v_r = -\frac{1}{3} f a H r \Delta$ in linear theory, and are spheres whose apparent axis ratio measures $D_M H$; the redshift-space outflow elongates them by a comparable amount, so geometry and growth are fitted together. Voids are the β_2 features of the cosmic web, separate only below $\delta \approx -0.8$ (Section 11.12).

11.14 Hubble's diagram, refitted

In 1929 Edwin Hubble published a table of twenty-four galaxies, then called extra-galactic nebulae. For each he had a distance and a radial velocity, the second measured from the Doppler shift of its spectral lines. Plotted against each other, the points rose to the right: the farther the galaxy, the faster it receded. Hubble drew a straight line through the origin and called its slope K. He found $K \approx 500\,\mathrm{km/s}$ per megaparsec [241]. Today the slope, the Hubble constant H_0, is about 70 in the same units. The data were real, the line was real, and the number was wrong by a factor of seven.

Two different things went into that factor, and they belong to two different parts of statistics. The scatter of the points about the line is the business of the estimator: how the slope is computed from noisy points, and how large its error bar is. The factor of seven is the business of calibration: every one of Hubble's distances was too small by a common factor, and no amount of fitting can see a common factor. Refitting Hubble's own numbers lets us watch both at work. The peculiar velocities that scatter his points are the next subject: where they come from, how large they are, and what they do to every local measurement of the expansion.

Where we have met H_0 before. In Section 3.17 we fitted mock surveys of 100 supernovae made with $H_0 = 70$ and an absolute magnitude M that we simply declared known, and the weighted mean of $5 \log_{10} H_0$ recovered the 70 at the Cramér–Rao bound. Section 3.17.4 then freed M and found that the supernovae fix only $M - 5 \log_{10} H_0$; the real distance ladder, which supplies M from parallaxes and Cepheids, gives 73.04 ± 1.04, 4.8σ above the CMB's 67.36 ± 0.54 (Sections T2.12 and T2.12.3). The model here is the same kind: a linear law, noise that is Gaussian and independent, and a slope found in closed form. What differs is the data and where the noise lives. The numbers are real, so no script supplies the truth; the velocities are measured directly rather than as redshifts of supernovae; and the noise that matters is not the 0.15 mag of a supernova magnitude but the peculiar velocities, half the recession velocity at 1 Mpc on Hubble's scale, together with the Sun's own motion, which must be fitted alongside the slope. What sets the scale is the question of the degeneracy again: Hubble's distances play the role of M, and his calibration of them was wrong by a factor that no fit can see. That is why 465 lies nowhere near 70, 73 or 67: the gap is a calibration error of about 4 mag, more than twenty times the 0.18 mag that separates the ladder from the CMB.

The question. Given Hubble's twenty-four pairs (r_i, v_i),

1. where does the noise live, and which estimator of the slope follows from that;
2. does our fit reproduce his K, including his correction for the motion of the Sun;
3. how much does the answer depend on whether the noise sits in the velocities or in the distances;
4. why does no error bar computed from the scatter warn that K is seven times too large?

11.14.1 Hubble's question and his data

Hubble wanted to know whether the radial velocities of the nebulae, already measured for dozens of them, depend on their distances. Many velocities were large and positive, which hinted at a systematic recession, but a velocity alone says nothing about how it grows with distance. What was new in 1929 was a set of distances. For the nearest galaxies they came from stars that can be recognised individually, Cepheid variables among them; for the more distant ones, from the brightest stars that could be resolved, assumed to have the same luminosity in every galaxy; and the four galaxies of the Virgo cluster were all placed at the distance assigned to the cluster, 2 Mpc. Table 11.4 reproduces the table.

object	r	v	object	r	v	object	r	v
SMC	0.032	170	NGC 4736	0.5	290	NGC 1068	1.0	920
LMC	0.034	290	NGC 5194	0.5	270	NGC 5055	1.1	450
NGC 6822	0.214	−130	NGC 4449	0.63	200	NGC 7331	1.1	500
NGC 598	0.263	−70	NGC 4214	0.8	300	NGC 4258	1.4	500
NGC 221	0.275	−185	NGC 3031	0.9	−30	NGC 4151	1.7	960
NGC 224	0.275	−220	NGC 3627	0.9	650	NGC 4382	2.0	500
NGC 5457	0.45	200	NGC 4826	0.9	150	NGC 4472	2.0	850
			NGC 5236	0.9	500	NGC 4486	2.0	800
						NGC 4649	2.0	1090

Table 11.4: Hubble's twenty-four nebulae with individual distances [241, Table 1]: distance r in Mpc on his scale, radial velocity v in km/s relative to the Sun. NGC 224 is the Andromeda galaxy, NGC 221 its companion, NGC 598 the Triangulum galaxy; the last four are in the Virgo cluster.

Two features of the table matter for what follows. Several nearby galaxies have *negative* velocities: Andromeda approaches us at 220 km/s. So the velocities are not a clean function of distance; something of order a few hundred km/s rides on top of any trend. And the velocities are measured from the Sun, which itself moves. A galaxy in the direction the Sun is heading appears to recede more slowly than it would from a resting observer, and one behind appears to recede faster. Hubble modelled both effects, and we do the same.

11.14.2 The model, and where the noise lives

We state the inputs first, separating physics from bookkeeping, as in Section 2.5.1.

Physical inputs.

(H1) *A linear law.* A galaxy at distance r that partakes only in the general recession moves away at Kr.

(H2) *Peculiar velocities.* On top of that, each galaxy has its own velocity through space, its **peculiar velocity** [its motion relative to the smooth expansion]. Along the line of sight it adds ϵ_i, which we treat as random, with mean zero and spread σ_v, independent from galaxy to galaxy, and Gaussian.

(H3) *The Sun moves* with a fixed velocity $\boldsymbol{v}_\odot$ relative to the mean of the galaxies.

Bookkeeping input. The distances r_i are exact; we relax this in Section 11.14.4.

What the Sun's motion does. Let $\hat{\boldsymbol{n}}_i$ be the unit vector from the Sun towards galaxy i. A velocity measured from a moving platform is the velocity relative to the platform, so the Sun's motion subtracts $\boldsymbol{v}_\odot \cdot \hat{\boldsymbol{n}}_i$ from every radial velocity (Figure 11.38). In equatorial coordinates, with right ascension α_i and declination δ_i, the unit vector has components

$\hat{\boldsymbol{n}}_i = (\cos\alpha_i\cos\delta_i,\ \sin\alpha_i\cos\delta_i,\ \sin\delta_i)$. Writing $(X, Y, Z) \equiv -\boldsymbol{v}_\odot$ for the reflex of the solar motion,

$$v_i = K\, r_i + X\cos\alpha_i\cos\delta_i + Y\sin\alpha_i\cos\delta_i + Z\sin\delta_i + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma_v^2). \tag{11.90}$$

This is Hubble's model. It is linear in the four unknowns (K, X, Y, Z).

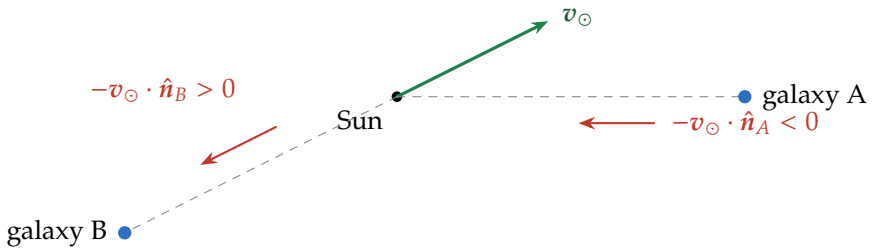


Figure 11.38: The reflex of the Sun's motion. Galaxy A lies ahead of the Sun's motion, so its measured velocity of recession is reduced by the component of $\boldsymbol{v}_\odot$ along its line of sight; galaxy B lies behind, and its measured recession is increased. The term is a dipole on the sky: $-v_\odot\cos\theta$, with θ the angle between the line of sight and the direction of motion.

The event, and what could have produced it. What we observed is a list of twenty-four velocities. For a candidate set of parameters (K, X, Y, Z) the model predicts each velocity, and the only thing that can turn the prediction into the observed number is the peculiar velocity ϵ_i. How plausible that required peculiar velocity is, under the Gaussian of (H2), is the likelihood (Section 3.8). Since the ϵ_i are independent, the likelihood is a product of Gaussians and $-2\ln\mathcal{L} = \chi^2 + \text{const}$ with $\chi^2 = \sum_i (v_i - \text{prediction}_i)^2/\sigma_v^2$, so the maximum-likelihood estimate is the least-squares fit (Section 3.10.1).

The slope alone first. Drop the solar motion for a moment, so the prediction is Kr_i. Then

$$\begin{aligned}
\chi^2(K) &\overset{(1)}{=} \frac{1}{\sigma_v^2}\sum_i \left(v_i - Kr_i\right)^2 = \frac{1}{\sigma_v^2}\Big[\sum_i v_i^2 - 2K\sum_i v_i r_i + K^2\sum_i r_i^2\Big], \\
\frac{\mathrm{d}\chi^2}{\mathrm{d}K} &\overset{(2)}{=} \frac{1}{\sigma_v^2}\Big[-2\sum_i v_i r_i + 2K\sum_i r_i^2\Big] = 0 \quad\Longrightarrow\quad \hat{K} = \frac{\sum_i v_i r_i}{\sum_i r_i^2}, \\
\operatorname{Var}\hat{K} &\overset{(3)}{=} \frac{\sum_i r_i^2\,\operatorname{Var} v_i}{\left(\sum_i r_i^2\right)^2} \overset{(4)}{=} \frac{\sigma_v^2}{\sum_i r_i^2}.
\end{aligned} \tag{11.91}$$

(1) Expand the square; K does not depend on i.
(2) The minimum of a parabola in K; the coefficient of K^2 is positive.
(3) $\hat{K}$ is a fixed linear combination $\sum_i c_i v_i$ with $c_i = r_i/\sum_j r_j^2$; for independent v_i the variance is $\sum_i c_i^2 \operatorname{Var} v_i$.
(4) $\operatorname{Var} v_i = \sigma_v^2$ by (H2), and $\sum_i r_i^2$ cancels once.

Each galaxy enters with weight r_i: distant galaxies, whose recession is large compared with the fixed σ_v, count for more. This is the same weighted ratio as the chronometer estimator of Section 3.17.6, with r_i in the place of E_i: a parameter that multiplies a known regressor gives a closed-form estimator.

Example: Hubble's numbers by hand. From Table 11.4, $\sum_i r_i^2 = 29.518\,\text{Mpc}^2$ and $\sum_i v_i r_i = 12514\,\text{km s}^{-1}\,\text{Mpc}$, so $\hat{K} = 12514/29.518 = 424\,\text{km/s/Mpc}$. The residuals $v_i - \hat{K}r_i$ scatter by 229 km/s; using that as σ_v, $\sigma_{\hat{K}} = 229/\sqrt{29.518} = 42$. The peculiar velocities are not a small nuisance: at 1 Mpc on Hubble's scale they are half the recession velocity.

With the solar motion. Now restore the four parameters. Stack them as $\boldsymbol{p} = (K, X, Y, Z)$ and the data as $\boldsymbol{v}$. Row i of the design matrix A is $(r_i,\ \cos\alpha_i\cos\delta_i,\ \sin\alpha_i\cos\delta_i,\ \sin\delta_i)$, so the prediction is $A\boldsymbol{p}$. Setting the four derivatives of $\chi^2 = |\boldsymbol{v} - A\boldsymbol{p}|^2/\sigma_v^2$ to zero gives the normal equations $A^{\mathsf{T}}A\,\hat{\boldsymbol{p}} = A^{\mathsf{T}}\boldsymbol{v}$, and the covariance of the estimate is $\sigma_v^2(A^{\mathsf{T}}A)^{-1}$; both were derived in Section 3.10.2 ((3.87) and (3.88)). The first column is the one we already had; the other three are what a dipole on the sky looks like in Cartesian components. For the positions we use modern right ascensions and declinations. Precession has moved them by about a degree since 1929, far less than the fit can resolve.

11.14.3 Side by side with Hubble

Script `code/ch11/08_hubble1929.py` types in the table, builds A and solves the normal equations.

Listing 11.12: The two least-squares fits: the slope alone, and the slope with the solar-motion dipole (`code/ch11/08_hubble1929.py`)

```
def fit_origin(r, v):
    """(a) v = K r: K = sum(v r)/sum(r^2), error from the residual scatter."""
    K = np.sum(v * r) / np.sum(r ** 2)
    s2 = np.sum((v - K * r) ** 2) / (len(r) - 1)
    return K, np.sqrt(s2 / np.sum(r ** 2)), np.sqrt(s2)


def design_solar(r, ra, dec):
    return np.column_stack([r, np.cos(ra) * np.cos(dec), np.sin(ra) * np.cos(dec),
    np.sin(dec)])


def fit_solar(r, v, ra, dec):
    """(b) linear least squares for (K, X, Y, Z): normal equations A^T A p = A^T v."""
    A = design_solar(r, ra, dec)
    AtA = A.T @ A
    p = np.linalg.solve(AtA, A.T @ v)
    res = v - A @ p
    s2 = res @ res / (len(v) - A.shape[1])
    cov = s2 * np.linalg.inv(AtA)
    return p, np.sqrt(np.diag(cov)), np.sqrt(s2)
```

`fit_origin` is (11.91), with the residual scatter divided by $N-1$ because one parameter was fitted. `design_solar` writes the four columns, and `fit_solar` solves $A^{\mathsf{T}}A\,\boldsymbol{p} = A^{\mathsf{T}}\boldsymbol{v}$ and estimates σ_v^2 from the residuals with $N-4$ degrees of freedom.

quantity	Hubble (1929)	our fit
K [km/s/Mpc]	465 ± 50	465 ± 56
X [km/s]	-65 ± 50	-70 ± 79
Y [km/s]	$+226 \pm 95$	236 ± 152
Z [km/s]	-195 ± 40	-200 ± 82
apex of the solar motion	$\alpha \approx 286°$, $\delta \approx +40°$	$\alpha = 286°$, $\delta = +39°$
slope without solar motion	—	424 ± 42

The central values agree to a few km/s: we recover Hubble's slope of 465 from his own table. The apex is the direction in which the Sun moves; it is $-(X, Y, Z)$ converted to angles, and the speed is $\sqrt{X^2+Y^2+Z^2} = 317\,\mathrm{km/s}$. That is the Sun's motion relative to the galaxies of the Local Group and their neighbours, a quantity we meet again in Section 11.16. Our standard errors are of the same order as Hubble's uncertainties but larger, by factors between 1.1 and 2;

we do not try to match his error conventions, since the central values are what the reproduction is about. Grouping the nebulae instead of fitting them one by one, Hubble found 513 ± 60, and two years later, with distances reaching the Leo cluster, Hubble and Humason found about 560 [242]. Every version lies near 500.

Dropping the solar motion lowers the slope to 424. The reason is in the sky positions. Many of the galaxies at intermediate distances, 0.45 to 1.4 Mpc on Hubble's scale (NGC 5457, 5194, 4736, 4449, 5055, 7331, 4258), lie in the northern sky, roughly in the direction towards which the Sun is moving. The reflex lowers each of their measured velocities by 100 to 240 km/s, so a line through the uncorrected points comes out shallower. Once the dipole is fitted and added back, those galaxies move up and the slope rises.

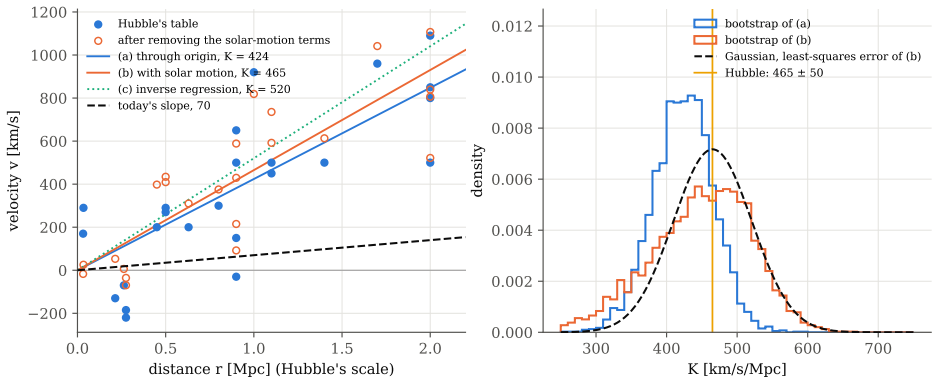


Figure 11.39: Left: Hubble's twenty-four galaxies (filled) and the same velocities with our fitted solar-motion dipole removed (open). The lines are the slope fitted through the origin (a), the slope fitted with the solar motion (b), the inverse regression of distance on velocity (c), and today's slope of 70 (dashed). Right: bootstrap distributions of the slopes (a) and (b), the Gaussian implied by the least-squares error of (b), and Hubble's value.

The error bar from resampling. The least-squares error assumes (H2): independent Gaussian peculiar velocities of a common spread. The bootstrap of Section 3.15 assumes only that the twenty-four galaxies are a fair sample of galaxies like them: draw twenty-four galaxies from the table with replacement, refit, repeat 4000 times. For the slope alone the bootstrap spread, 42, matches the formula. For the four-parameter fit it is 73, wider than 56 and skewed (Figure 11.39, right): a resample that happens to contain few galaxies in some part of the sky cannot pin down the dipole, and the slope wanders with it. With twenty-four points and four parameters the Gaussian error bar is optimistic.

11.14.4 When the distances are noisy too

Hubble's distances were not exact. A distance from the brightest stars of a galaxy is uncertain by a fixed *fraction*, because the luminosity of the brightest star varies from galaxy to galaxy and a magnitude error is a fractional distance error. So now replace the bookkeeping input by a physical one:

(H4) *Lognormal distance errors.* The measured distance is $r_i^{\mathrm{obs}} = r_i\, e^{\eta_i}$ with $\eta_i \sim \mathcal{N}(0, \Delta^2)$ independent of everything else; Δ is the fractional error.

Two simple limits bracket the answer. If all the noise is in the velocities, we regress v on r, which is (11.91). If all the noise is in the distances, the roles swap: the velocity is exact, the distance is the noisy variable, and we should regress r on v, $r = bv$, with $\hat{b} = \sum_i v_i r_i / \sum_i v_i^2$ by

the same chain with r and v exchanged, and $\hat{K} = 1/\hat{b} = \sum_i v_i^2 / \sum_i v_i r_i = 520$. The two lines differ by more than twice the error bar of either. The data have not changed; only the assumption about which number carries the noise has. That is the likelihood at work: it asks how plausible the noise needed to turn the prediction into the data is, and the answer depends on where that noise is allowed to sit.

Both noises at once. The honest model keeps both. To first order in the small η_i, $r_i = r_i^{\mathrm{obs}} e^{-\eta_i} \approx r_i^{\mathrm{obs}}(1-\eta_i)$, so

$$
\begin{aligned}
v_i &\overset{(1)}{=} K r_i + \epsilon_i \overset{(2)}{=} K r_i^{\mathrm{obs}} + \left(\epsilon_i - K r_i^{\mathrm{obs}} \eta_i\right), \\
\mathrm{Var}\left(v_i \mid r_i^{\mathrm{obs}}\right) &\overset{(3)}{=} \sigma_v^2 + K^2 \Delta^2 \left(r_i^{\mathrm{obs}}\right)^2 \equiv s_i^2(K).
\end{aligned}
\tag{11.92}
$$

(1) The model (H1)–(H2), without the solar motion for clarity.
(2) Substitute $r_i \approx r_i^{\mathrm{obs}}(1-\eta_i)$ and collect the noise.
(3) ϵ_i and η_i are independent, so their variances add; the second is multiplied by the square of its coefficient $K r_i^{\mathrm{obs}}$.

The noise of each point now grows with its distance, and its variance depends on the parameter K itself. The likelihood is no longer a plain χ^2:

$$
-2 \ln \mathcal{L}(K, \sigma_v) = \sum_i \left[\frac{\left(v_i - K r_i^{\mathrm{obs}}\right)^2}{s_i^2(K)} + \ln s_i^2(K) \right] + \text{const},
$$

and the $\ln s_i^2$ term, the normalisation of each Gaussian, must be kept: it depends on K, and dropping it would let the fit lower χ^2 simply by making K, and with it every s_i, large. We maximise it numerically over K and σ_v at a fixed Δ. At $\Delta = 0$ it returns 424, the slope through the origin; at $\Delta = 0.2$ (a 0.43 mag distance error) 420; at $\Delta = 0.4$, 397. As Δ grows the distant galaxies are down-weighted, since their variance $K^2\Delta^2 r^2$ grows fastest. The first-order expansion has also dropped something important: given a measured distance, the true distance is not equally likely to lie above and below it, because there are more galaxies at larger distances. That is Malmquist bias, and Section 11.18 treats it properly.

11.14.5 Why 465 and not 70

None of these fits moves the slope anywhere near 70. The reason is a property of the model, not of the estimator. Suppose every distance in the table is too small by the same factor c, so that the true distances are $c r_i$. Refit:

$$
\begin{aligned}
\hat{K}(c\,\boldsymbol{r}) &\overset{(1)}{=} \frac{\sum_i v_i\,(c r_i)}{\sum_i (c r_i)^2} \overset{(2)}{=} \frac{1}{c}\,\hat{K}(\boldsymbol{r}), \\
v_i - \hat{K}(c\,\boldsymbol{r})\, c r_i &\overset{(3)}{=} v_i - \hat{K}(\boldsymbol{r})\, r_i .
\end{aligned}
\tag{11.93}
$$

(1) (11.91) with r_i replaced by $c r_i$.
(2) One factor of c in the numerator, two in the denominator.
(3) The product $\hat{K}\, r_i$ is unchanged, so every residual is unchanged.

The residuals, and so $\chi^2_{\min}$, the estimate of σ_v and the fractional error $\sigma_{\hat{K}}/\hat{K}$, are identical for every c. The data cannot tell $c = 1$ from $c = 6.6$; they constrain only the product Kc. The error bar on Hubble's slope (±50 in his paper, ±56 in our refit) is an honest statement about the scatter of his points and says nothing about c. This is the H_0–M degeneracy of the supernova

Hubble diagram (Section 3.17.4) in its oldest form: a velocity–distance diagram measures H_0 only in units of its distance scale, and the scale has to come from somewhere else.

How large was the error of the scale? To turn 465 into 70, every distance must grow by $465/70 = 6.6$, which in the language of magnitudes is $5\log_{10} 6.6 = 4.1$ mag: Hubble's standard candles were assumed about four magnitudes fainter than they are. Hubble placed the Virgo cluster at 2 Mpc; today it is at about 16 Mpc. The error was found in pieces. Baade showed that the Cepheids used to calibrate the period–luminosity relation belonged to a different stellar population from those in the nebulae, which doubled the distances [243]. Sandage showed that many of Hubble's "brightest stars" in the more distant galaxies were clouds of ionised gas, intrinsically much brighter, and by 1958 the slope was near 75, still uncertain by a factor of two [244]. A further, quieter bias pushed the same way. The farther galaxies' distances came from the brightest objects that could be seen in them, and near the limit of a survey the objects that can be seen are the ones that happen to be unusually bright, so their distances come out too small; that is the classical Malmquist bias of Section T2.13. The modern version of the scale, Cepheids anchored by geometry and carried to supernovae, is the ladder of Section T2.12.

The lesson. Refitting Hubble's table reproduces his 465 to a few km/s, and shows that the error bar measures the scatter of the points, never the scale they are plotted on: a common multiplicative error is invisible to the fit and does not shrink with more galaxies.

11.15 Peculiar velocities: what moves the galaxies

The scatter in Hubble's diagram was 229 km/s (Section 11.14.2), and Andromeda, our nearest large neighbour, comes towards us at 220 km/s against the general recession. In the refit of Section 11.14 these peculiar velocities were noise: independent random numbers ϵ_i with a spread σ_v to be estimated from the residuals. Treating the velocities as independent neglects their common origin. Nearby galaxies accelerate in much the same gravitational field, so their peculiar velocities can be correlated. Once we know what produces the velocities, three things follow: how large they are in a universe with the measured power spectrum, how strongly the velocities of two galaxies a distance r apart are correlated, and, in the next two sections, what they do to every measurement of the local expansion.

The question.

1. How is a peculiar velocity read off from a redshift and a distance?
2. Given the density contrast everywhere, what is the velocity at one point?
3. How large is the typical velocity today, for the CAMB power spectrum, and which scales produce it?
4. How correlated are the velocities of two galaxies a distance r apart?

11.15.1 Reading a peculiar velocity from a redshift

A galaxy's redshift has two causes, composed one after the other (Section 11.5.1, (11.35)): the expansion stretches the light by $1 + z_c$, where z_c is the **cosmological redshift** the galaxy would have at rest, and its own line-of-sight velocity u_r Doppler-shifts it by $1 + u_r/c$. Solving for u_r,

$$1 + z_{\rm obs} = (1 + z_c)\Big(1 + \frac{u_r}{c}\Big) \quad\Longrightarrow\quad u_r \overset{(1)}{=} c\,\frac{(1 + z_{\rm obs}) - (1 + z_c)}{1 + z_c} \overset{(2)}{=} c\,\frac{z_{\rm obs} - z_c}{1 + z_c}, \tag{11.94}$$

(1) Divide by $1 + z_c$, subtract 1, multiply by c.
(2) The ones cancel in the numerator.

which is the definition Hogg uses [104, eq. (10)]. Near us $z_c \ll 1$ and $cz_c = H_0 r$, so $cz_{\text{obs}} \approx H_0 r + u_r$: the observed recession velocity is the Hubble flow plus the line-of-sight peculiar velocity. That is the model of Section 11.14.2, with ϵ_i renamed $u_{r,i}$.

To measure u_r for one galaxy we therefore need its distance from something other than its redshift: a Cepheid, a supernova, the Tully–Fisher relation. The error of $u_r = cz_{\text{obs}} - H_0 r$ is then $H_0 \sigma_r$, and it grows with distance.

Example: how far out can one velocity be measured? A Tully–Fisher distance is uncertain by about 20%. At $r = 10\,\text{Mpc}$, with $H_0 = 70\,\text{km/s/Mpc}$, the velocity error is $70 \times 0.2 \times 10 = 140\,\text{km/s}$, smaller than the 300 km/s we will find for a typical peculiar velocity. At 50 Mpc it is 700 km/s, larger. Beyond a few tens of megaparsecs individual peculiar velocities are lost in the distance errors, and only averages over many galaxies survive. That is why the local Universe is the place where velocities are data.

11.15.2 The velocity field of linear theory

Inputs. We take two results of cosmological perturbation theory as given (see the cosmology notes); they are the same inputs as in the redshift-space distortions of Section 11.5.2.

(V1) *Linear continuity.* With $\boldsymbol{u}$ the peculiar velocity, a the scale factor, H the expansion rate, δ the matter density contrast and $f = \mathrm{d}\ln D/\mathrm{d}\ln a$ the growth rate, $\nabla \cdot \boldsymbol{u} = -aHf\,\delta$, which is (11.38) multiplied by aH. Matter flows out of regions where the density falls.

(V2) *No curl.* Rotational velocity perturbations decay as the Universe expands, so $\nabla \times \boldsymbol{u} = 0$ on large scales.

The question: given δ everywhere, what is $\boldsymbol{u}$ at the point $\boldsymbol{x}$?

A field without curl is a gradient. Write $\boldsymbol{u} = \nabla\psi$. Then (V1) is an equation for ψ alone, $\nabla^2\psi = -aHf\,\delta$: Poisson's equation, with $-aHf\,\delta$ as the source. We solve it the way one solves for an electrostatic potential, with the potential of a point source.

The point source. Away from the origin, with $\rho = |\boldsymbol{x}|$, $\nabla^2(1/\rho) = \rho^{-2}\,\mathrm{d}/\mathrm{d}\rho\,\big(\rho^2\,\mathrm{d}(1/\rho)/\mathrm{d}\rho\big) = \rho^{-2}\,\mathrm{d}(-1)/\mathrm{d}\rho = 0$. At the origin it is singular, and its strength follows from the divergence theorem: the flux of $\nabla(1/\rho) = -\hat{\boldsymbol{\rho}}/\rho^2$ out of a sphere of any radius ε is $(-1/\varepsilon^2)(4\pi\varepsilon^2) = -4\pi$. So $\nabla^2(1/\rho)$ is zero everywhere except at the origin and integrates to -4π over any ball around it: $\nabla^2\big(1/|\boldsymbol{x} - \boldsymbol{x}'|\big) = -4\pi\,\delta_D(\boldsymbol{x} - \boldsymbol{x}')$, with δ_D the Dirac delta.

Superpose. A sum of point sources solves Poisson's equation for the summed source. Hence

$$\psi(\boldsymbol{x}) \overset{(1)}{=} \frac{aHf}{4\pi}\int \mathrm{d}^3x'\,\frac{\delta(\boldsymbol{x}')}{|\boldsymbol{x} - \boldsymbol{x}'|},$$
$$\boldsymbol{u}(\boldsymbol{x}) = \nabla\psi \overset{(2)}{=} \frac{aHf}{4\pi}\int \mathrm{d}^3x'\,\delta(\boldsymbol{x}')\,\frac{\boldsymbol{x}' - \boldsymbol{x}}{|\boldsymbol{x}' - \boldsymbol{x}|^3}. \tag{11.95}$$

(1) Check: $\nabla^2\psi = \frac{aHf}{4\pi}\int \mathrm{d}^3x'\,\delta(\boldsymbol{x}')\,\big(-4\pi\delta_D(\boldsymbol{x} - \boldsymbol{x}')\big) = -aHf\,\delta(\boldsymbol{x})$, which is (V1).
(2) $\nabla_{\boldsymbol{x}}|\boldsymbol{x} - \boldsymbol{x}'|^{-1} = -(\boldsymbol{x} - \boldsymbol{x}')/|\boldsymbol{x} - \boldsymbol{x}'|^3$; the gradient acts on $\boldsymbol{x}$, not on the integration variable.

Key idea: The linear velocity field

In linear theory the peculiar velocity at $\boldsymbol{x}$ is $\boldsymbol{u}(\boldsymbol{x}) = \frac{aHf}{4\pi}\int \mathrm{d}^3x'\,\delta(\boldsymbol{x}')\,(\boldsymbol{x}'-\boldsymbol{x})/|\boldsymbol{x}'-\boldsymbol{x}|^3$ [108, 245]: every overdense patch pulls with an inverse-square law, and the velocity points along the total pull. In Fourier space it is $\boldsymbol{u}_{\boldsymbol{k}} = i\,aHf\,\boldsymbol{k}\,\delta_{\boldsymbol{k}}/k^2$, the relation used for redshift-space distortions in (11.40).

The integral is Newton's gravitational field of the excess mass $\bar\rho\,\delta$, up to a constant. The velocity is parallel to the peculiar gravitational acceleration: in linear theory a galaxy moves in the direction it is pulled. The inverse-square fall-off is offset by the volume of a shell, which grows as the square of its radius, so distant matter is not negligible. A shell of uniform excess pulls equally in all directions and contributes nothing, but a shell whose density differs between one side and the other contributes as much as a nearby clump of the same total mass excess.

Example: one density wave. Take $\delta(\boldsymbol{x}) = A\cos(\boldsymbol{k}\cdot\boldsymbol{x})$ with a small amplitude A. The Fourier form gives $\boldsymbol{u} = -aHf\,A\,(\boldsymbol{k}/k^2)\sin(\boldsymbol{k}\cdot\boldsymbol{x})$: a velocity along $\boldsymbol{k}$, a quarter wavelength out of step with the density, with amplitude $aHfA/k = aHfA\lambda/2\pi$. Matter flows from the troughs towards the crests. For the same density amplitude, a longer wave gives a faster flow, in proportion to its wavelength. With $aH = 100$ km/s per h^{-1}Mpc today, $f = 0.527$, $A = 0.05$ and $\lambda = 100\,h^{-1}$Mpc, the flow amplitude is $100\times0.527\times0.05\times100/(2\pi) \approx 42$ km/s.

11.15.3 How fast, on average

The density contrast on large scales is a Gaussian random field with power spectrum $P(k)$ (Section 7.3.3): $\langle\delta_{\boldsymbol{k}}\rangle = 0$ and $\langle\delta_{\boldsymbol{k}}\delta^*_{\boldsymbol{k}'}\rangle = (2\pi)^3\delta_D(\boldsymbol{k}-\boldsymbol{k}')P(k)$, with $\delta(\boldsymbol{x}) = \int\frac{\mathrm{d}^3k}{(2\pi)^3}\delta_{\boldsymbol{k}}e^{i\boldsymbol{k}\cdot\boldsymbol{x}}$. The velocity is a linear function of δ, and a linear function of a Gaussian field is a Gaussian field (Section 7.2). So each velocity component is Gaussian with mean zero, and all we need is its covariance. For components i and j at two points separated by $\boldsymbol{r}$,

$$\begin{aligned}\big\langle u_i(\boldsymbol{x})\,u_j(\boldsymbol{x}+\boldsymbol{r})\big\rangle &\overset{(1)}{=} \int\frac{\mathrm{d}^3k}{(2\pi)^3}\int\frac{\mathrm{d}^3k'}{(2\pi)^3}\big\langle u_{i,\boldsymbol{k}}\,u^*_{j,\boldsymbol{k}'}\big\rangle\,e^{i\boldsymbol{k}\cdot\boldsymbol{x}-i\boldsymbol{k}'\cdot(\boldsymbol{x}+\boldsymbol{r})}\\ &\overset{(2)}{=} (aHf)^2\int\frac{\mathrm{d}^3k}{(2\pi)^3}\int\frac{\mathrm{d}^3k'}{(2\pi)^3}\,\frac{k_ik'_j}{k^2k'^2}\,(2\pi)^3\delta_D(\boldsymbol{k}-\boldsymbol{k}')P(k)\,e^{i\boldsymbol{k}\cdot\boldsymbol{x}-i\boldsymbol{k}'\cdot(\boldsymbol{x}+\boldsymbol{r})}\\ &\overset{(3)}{=} (aHf)^2\int\frac{\mathrm{d}^3k}{(2\pi)^3}\,P(k)\,\frac{k_ik_j}{k^4}\,e^{i\boldsymbol{k}\cdot\boldsymbol{r}}.\end{aligned}\tag{11.96}$$

(1) u_j is real, so we may write it as its own complex conjugate, $u_j(\boldsymbol{y}) = \int\frac{\mathrm{d}^3k'}{(2\pi)^3}u^*_{j,\boldsymbol{k}'}e^{-i\boldsymbol{k}'\cdot\boldsymbol{y}}$; the average passes inside the integrals.

(2) $u_{i,\boldsymbol{k}} = iaHf\,k_i\delta_{\boldsymbol{k}}/k^2$ and $u^*_{j,\boldsymbol{k}'} = -iaHf\,k'_j\delta^*_{\boldsymbol{k}'}/k'^2$; the factors i and $-i$ multiply to 1; then the spectrum of δ.

(3) The delta function sets $\boldsymbol{k}' = \boldsymbol{k}$ and cancels one $(2\pi)^3$; the $\boldsymbol{x}$ dependence cancels, as homogeneity requires. The sign of the exponent is immaterial because the rest of the integrand is even in $\boldsymbol{k}$.

One point. At $\boldsymbol{r} = 0$, sum over $i = j$: $k_ik_i = k^2$, and with $\mathrm{d}^3k = 4\pi k^2\mathrm{d}k$ for an isotropic integrand,

$$\sigma_v^2 \equiv \big\langle|\boldsymbol{u}|^2\big\rangle = (aHf)^2\int\frac{4\pi k^2\,\mathrm{d}k}{(2\pi)^3}\,\frac{P(k)}{k^2} = \frac{(aHf)^2}{2\pi^2}\int_0^\infty P(k)\,\mathrm{d}k, \qquad \sigma^2_{v,1\mathrm{D}} = \frac{\sigma_v^2}{3},\tag{11.97}$$

the second because isotropy shares the three components equally. Compare the density variance in a sphere of radius R (Section 7.6), $\sigma_R^2 = \frac{1}{2\pi^2}\int k^2P(k)\,W^2(kR)\,\mathrm{d}k$: the velocity carries two fewer

powers of k. That is the wave example again: velocity grows with wavelength, so velocities are made by larger scales than densities.

Example: the numbers today. Script `code/ch11/09_peculiar_velocities.py` integrates (11.97) with the CAMB linear spectrum at $z = 0$ (the one of Section T1.4, $\sigma_8 = 0.811$) and $f = 0.527$. Lengths are in h^{-1}Mpc, so $aH = 100\,\mathrm{km/s}$ per h^{-1}Mpc and h drops out. It finds $\sigma_v = 534\,\mathrm{km/s}$ in three dimensions and $\sigma_{v,1D} = 308\,\mathrm{km/s}$ along one line of sight: the "$\sim 300\,\mathrm{km/s}$" assumed for the peculiar velocities of supernova hosts in Section 3.17.1 and Section T2.11. Figure 11.40 (left) shows where it comes from. Half of σ_v^2 comes from $k < 0.060\,h/\mathrm{Mpc}$, wavelengths longer than $105\,h^{-1}$Mpc, while half of σ_8^2 comes from $k < 0.17$, wavelengths longer than $37\,h^{-1}$Mpc. A galaxy's velocity is set by the distribution of mass over a region several times larger than the region that sets its local density.

This number is the velocity of the smooth large-scale flow. Galaxies inside groups and clusters also orbit in the group's potential, at up to 1000 km/s in rich clusters; linear theory knows nothing of that, and it adds a non-Gaussian, small-scale part on top of (11.97).

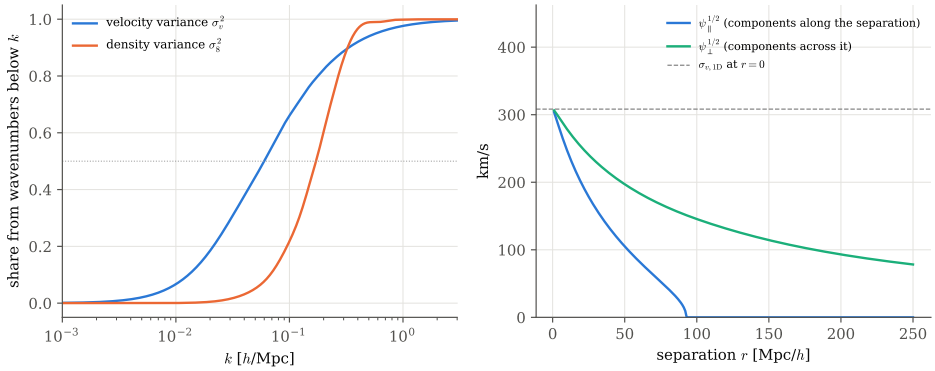


Figure 11.40: Left: the share of the velocity variance σ_v^2 and of the density variance σ_8^2 contributed by wavenumbers below k; velocities come from about three times larger scales. Right: the square roots of the velocity correlation functions for components along the separation ($\psi_\parallel$) and across it ($\psi_\perp$); both start at the one-dimensional rms velocity, and $\psi_\perp$ is still half of it at $100\,h^{-1}$Mpc.

11.15.4 How correlated: the velocity correlation functions

Now keep $\boldsymbol{r} \neq 0$ in (11.96). Isotropy says the answer can only be built from δ_{ij} and $\hat{r}_i\hat{r}_j$: the component of velocity along the separation and the components across it are the only distinguishable directions. So we need the angular average of $\hat{k}_i\hat{k}_j\, e^{i\boldsymbol{k}\cdot\boldsymbol{r}}$ over the directions of $\boldsymbol{k}$.

Start from the simplest average. Without the factors $\hat{k}_i\hat{k}_j$ the average over directions is a familiar integral: with μ the cosine of the angle between $\boldsymbol{k}$ and $\boldsymbol{r}$ and $x = kr$, $\frac{1}{4\pi}\int \mathrm{d}\Omega_k\, e^{i\boldsymbol{k}\cdot\boldsymbol{r}} = \frac{1}{2}\int_{-1}^{1} e^{ix\mu}\mathrm{d}\mu = \frac{\sin x}{x} = j_0(x)$, the spherical Bessel function of order zero.

Bring down the factors by differentiating. Each derivative $\partial/\partial r_i$ of $e^{i\boldsymbol{k}\cdot\boldsymbol{r}}$ brings down ik_i, so $k_ik_j\, e^{i\boldsymbol{k}\cdot\boldsymbol{r}} = -\partial_i\partial_j e^{i\boldsymbol{k}\cdot\boldsymbol{r}}$, and the angular average of $k_ik_j e^{i\boldsymbol{k}\cdot\boldsymbol{r}}$ is $-\partial_i\partial_j\, j_0(kr)$. For a function $g(r)$ of $r = |\boldsymbol{r}|$ alone, $\partial_i g = g'\hat{r}_i$ and $\partial_j\hat{r}_i = (\delta_{ij} - \hat{r}_i\hat{r}_j)/r$, so

$$-\partial_i\partial_j\, j_0(kr) \overset{(1)}{=} -\Big[g''\,\hat{r}_i\hat{r}_j + \frac{g'}{r}\big(\delta_{ij} - \hat{r}_i\hat{r}_j\big)\Big]_{g=j_0(kr)}$$
$$\overset{(2)}{=} k^2\Big[\Big(j_0(x) - \frac{2j_1(x)}{x}\Big)\hat{r}_i\hat{r}_j + \frac{j_1(x)}{x}\big(\delta_{ij} - \hat{r}_i\hat{r}_j\big)\Big],$$

$$\frac{1}{4\pi}\int \mathrm{d}\Omega_k\, \hat{k}_i\hat{k}_j\, e^{i\boldsymbol{k}\cdot\boldsymbol{r}} \overset{(3)}{=} \frac{j_1(x)}{x}\,\delta_{ij} + \Big(j_0(x) - \frac{3j_1(x)}{x}\Big)\hat{r}_i\hat{r}_j. \tag{11.98}$$

(1) The product rule on $\partial_j(g'\hat{r}_i)$.
(2) $j_0'(x) = -j_1(x)$ and $j_1'(x) = j_0(x) - 2j_1(x)/x$, with $j_1(x) = \sin x/x^2 - \cos x/x$; so $g' = -kj_1$, $g'' = -k^2(j_0 - 2j_1/x)$, and $g'/r = -k^2 j_1/x$.
(3) Divide by k^2 (since $k_ik_j = k^2\hat{k}_i\hat{k}_j$) and collect the $\hat{r}_i\hat{r}_j$ terms: $j_0 - 2j_1/x - j_1/x$.

A check: at $r \to 0$, $j_1(x)/x \to 1/3$ and $j_0 \to 1$, so the average becomes $\delta_{ij}/3$, the average of $\hat{k}_i\hat{k}_j$ over all directions. Putting (11.98) into (11.96) with $\mathrm{d}^3k = k^2\mathrm{d}k\,\mathrm{d}\Omega_k$, the k^2 cancels the k^2 left in $k_ik_j/k^4 = \hat{k}_i\hat{k}_j/k^2$, and for the components across and along the separation (take $i = j$ perpendicular to $\hat{\boldsymbol{r}}$, or $i = j$ along it):

$$\psi_\perp(r) = \frac{(aHf)^2}{2\pi^2}\int_0^\infty \mathrm{d}k\, P(k)\,\frac{j_1(kr)}{kr}, \qquad \psi_\parallel(r) = \frac{(aHf)^2}{2\pi^2}\int_0^\infty \mathrm{d}k\, P(k)\Big[j_0(kr) - \frac{2j_1(kr)}{kr}\Big], \tag{11.99}$$

the two velocity correlation functions of Górski [246]. Both equal $\sigma^2_{v,1\mathrm{D}}$ at $r = 0$.

Figure 11.40 (right) shows them. At $r = 100\,h^{-1}$Mpc, $\psi_\perp^{1/2} = 146\,\mathrm{km/s}$: two galaxies a hundred megaparsecs apart still share about half of their typical transverse velocity. Large regions move together, in **bulk flows**. $\psi_\parallel$ falls faster and turns slightly negative: the components along the separation of two distant galaxies tend to have opposite signs, as they do on the two sides of a region into which matter is falling, or out of which it is streaming.

Watch out: Peculiar velocities are not independent noise

Model (H2) of Section 11.14.2 treated the peculiar velocities of Hubble's galaxies as independent. Within a few tens of megaparsecs they are almost perfectly correlated. A sample of nearby galaxies shares a common flow, which does not average away as $1/\sqrt{N}$: it shifts every velocity the same way. The part of the common flow that grows linearly with distance looks exactly like a change of the Hubble constant (Section 11.17).

11.15.5 Velocities in a simulated box

The derivation assumed a Gaussian field with the CAMB spectrum. We now build such a field and check both results. Script `code/ch11/09_peculiar_velocities.py` draws 4 boxes of side $1000\,h^{-1}$Mpc with 256^3 cells, using the grid conventions of Section 7.4.1, and computes the velocity from the Fourier form of the key idea:

Listing 11.13: Velocity components from a Gaussian density field, and their one- and two-point averages (`code/ch11/09_peculiar_velocities.py`)

```
103         w = rng.standard_normal((N, N, N))
104         dk = np.fft.rfftn(w) * amp                  # delta_k / D^3 in the lib_lss convention
105         vx = np.fft.irfftn(1j * H100 * f * KX * inv_k2 * dk, s=(N, N, N), axes=(0, 1, 2))
106         vy = np.fft.irfftn(1j * H100 * f * KY * inv_k2 * dk, s=(N, N, N), axes=(0, 1, 2))
107         var_meas.append(0.5 * (vx.var() + vy.var()))
108         par_meas.append([np.mean(vx * np.roll(vx, -n, axis=0)) for n in nsep])
109         perp_meas.append([np.mean(vy * np.roll(vy, -n, axis=0)) for n in nsep])
110         hist_v.append(vx[::4, ::4, ::4].ravel().copy())
```

Line 104 colours white noise with $\sqrt{P(k)}$, which makes `dk` the Fourier transform of a Gaussian field with spectrum P on the grid. Lines 105–106 multiply by $iaHf\,k_x/k^2$ and $iaHf\,k_y/k^2$ (with $aH = 100$ and $1/k^2$ set to zero for the $\boldsymbol{k} = 0$ mode) and transform back: these are u_x and u_y in km/s. Line 107 records the one-point variance. Lines 108–109 shift the field along x by n cells

and average the products: u_x with u_x is $\psi_\parallel$ at separation $n \times 1000/256\, h^{-1}$Mpc, and u_y with u_y is $\psi_\perp$.

The theory to compare with is (11.96) restricted to the modes the box contains: the integral over $\boldsymbol{k}$ becomes a sum over the grid modes, divided by the volume (Section 7.4.1). It gives $\sigma_{v,1D} = 302\,\mathrm{km/s}$; the boxes give 298 ± 5, where the $\pm$ is the box-to-box scatter. Both are below the 308 of the full integral because the box has no waves longer than $1000\, h^{-1}$Mpc, and Figure 11.40 showed that such waves carry a few per cent of σ_v^2. The histogram of u_x is Gaussian, and the correlation functions follow the grid-mode sums (Figure 11.41). The points at large separations sit a little below the curves, by about one box-to-box scatter: with four boxes, the longest waves are sampled only a handful of times, which is the sample variance of Section 7.5.2 acting on a quantity dominated by large scales.

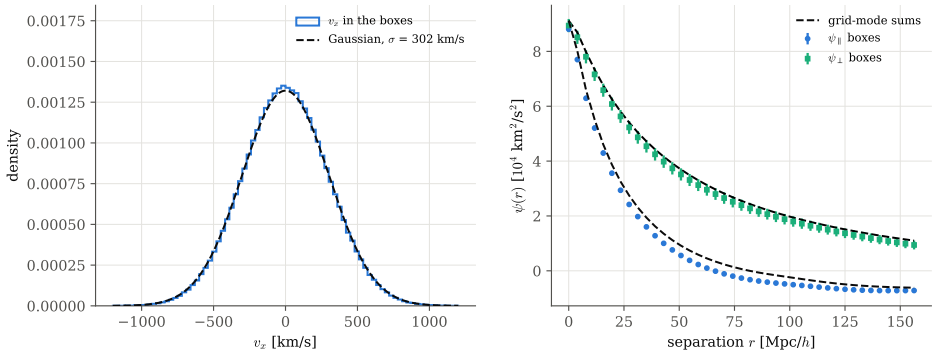


Figure 11.41: Velocities in Gaussian boxes. Left: the distribution of one velocity component against a Gaussian of the predicted width. Right: the two velocity correlation functions measured in the boxes (points, with the error of the mean of four boxes) and the corresponding sums over the box's Fourier modes (dashed).

The lesson. Peculiar velocities are the gradient of a potential sourced by the density contrast; that makes them Gaussian, about 300 km/s along a line of sight today, and coherent over a hundred megaparsecs, because they are made by longer waves than the densities are.

11.16 The local flow: falling towards Virgo

Four of Hubble's twenty-four galaxies were in the Virgo cluster, the nearest rich cluster, a few hundred galaxies at about 16 Mpc. A mass of that size pulls on everything around it, us included. Galaxies between us and Virgo are pulled away from us, towards the cluster; galaxies beyond it are pulled towards us. We ourselves fall towards Virgo, so every velocity we measure is measured from a moving platform. A measurement of H_0 from nearby galaxies that ignores all this fits a straight line to a pattern that is not straight. The linear velocity field of Section 11.15 turns the pattern into a model with one extra parameter, our infall speed, which can be fitted together with H_0.

Where we have met H_0 before. Every fit of H_0 so far treated peculiar velocities as noise. The mock supernovae of Section 3.17.1, made with $H_0 = 70$, carried independent Gaussian velocities of 300 km/s, and Hubble's twenty-four nebulae (Section 11.14) left residuals of 200 to 230 km/s that we fitted as an independent scatter, with only the Sun's motion modelled, as a dipole. The real ladder of Section T2.12 avoids the problem rather than modelling it: its Hubble-flow supernovae start at $z = 0.023$, about 7000 km/s, where a velocity of 250 km/s shifts a distance modulus by less than 0.1 mag (T2.99). Here the galaxies lie within 30 Mpc, where the whole

recession is at most a couple of thousand km/s and a coherent infall of 200 km/s is not small, so the velocities move from the noise into the model. The estimator is the same weighted least squares, now with two columns, H_0 and the infall speed, and the data are mock again, made with $H_0 = 70$ as in Section 3.17, so the question is once more whether a fit recovers a known truth. The low value it returns when the infall is ignored is therefore not a third measurement to set beside 73.04 and 67.36: it is the size of an omitted-variable bias, a measure of how far a local fit can miss without a model of the flow.

The question.

1. What velocity field does linear theory predict around a spherical overdensity?
2. What does an observer who is falling into it measure for the radial velocities of galaxies around them?
3. Can H_0 and the infall speed be fitted together from a local sample, and what goes wrong if the infall is ignored, or if the distances are poor?
4. How fast does the Local Group move with respect to the CMB, and how much of that is Virgo?

11.16.1 Infall around a spherical overdensity

Take a density contrast δ that is spherically symmetric about a centre, the cluster. Let $\boldsymbol{R}$ be the vector from the centre to a point, $R = |\boldsymbol{R}|$, and $\bar{\delta}(<R)$ the mean contrast inside the sphere of radius R. By symmetry the linear velocity $\boldsymbol{u}$ of Section 11.15.2 points along $\pm\hat{\boldsymbol{R}}$ and its size u_R depends only on R. Integrate the continuity equation (V1), $\nabla\cdot\boldsymbol{u} = -aHf\,\delta$, over the ball of radius R:

$$4\pi R^2\, u_R(R) \overset{(1)}{=} \oint_{|\boldsymbol{R}'|=R} \boldsymbol{u}\cdot \mathrm{d}\boldsymbol{S} \overset{(2)}{=} \int_{|\boldsymbol{R}'|<R} \nabla\cdot\boldsymbol{u}\;\mathrm{d}^3R' \overset{(3)}{=} -aHf\,\frac{4\pi}{3}R^3\,\bar{\delta}(<R),$$

$$u_R(R) = -\frac{aHf}{3}\,\bar{\delta}(<R)\,R. \tag{11.100}$$

(1) The velocity is radial and has the same size everywhere on the sphere, so the flux is u_R times the area.
(2) The divergence theorem.
(3) (V1), and the definition of the mean contrast: $\int_{<R}\delta\,\mathrm{d}^3R' = \frac{4\pi}{3}R^3\bar{\delta}(<R)$.

This is Newton's shell theorem in the language of velocities: only the excess inside R matters. An overdensity ($\bar{\delta}>0$) gives an inflow, a void an outflow. Inside a region of uniform contrast the flow is $-\frac{1}{3}aHf\bar{\delta}\,\boldsymbol{R}$, a change of the local expansion rate by $-\frac{1}{3}f\bar{\delta}$ in units of H: the local-void formula (T2.108), which we derived for the H_0 tension in Section T2.12 and will meet again in Section 11.17.

A power-law cluster. Around a cluster the mean contrast falls with radius. Model it as a power law, $\bar{\delta}(<R)\propto R^{-2}$, the shape used in classical models of the Virgocentric flow [247]. Then (11.100) gives a speed that falls as $1/R$. We name the constant by the infall speed of the Local Group, which sits at $R = d_V$, the distance of Virgo:

$$\boldsymbol{u}(\boldsymbol{x}) = -v_{\mathrm{inf}}\,\frac{d_V}{R}\,\hat{\boldsymbol{R}}, \qquad v_{\mathrm{inf}} = \frac{H_0 f}{3}\,\bar{\delta}(<d_V)\,d_V. \tag{11.101}$$

Example: how overdense is our neighbourhood? Literature values of our infall into Virgo lie around 200 km/s. With $H_0 = 70$, $f = 0.53$ and $d_V = 16.5$ Mpc, (11.101) gives $\bar{\delta}(< d_V) = 3 \times 200/(70 \times 0.53 \times 16.5) \approx 1.0$: the sphere centred on Virgo and reaching out to us contains about twice the mean density. A contrast of order one is the edge of linear theory. The model is good for the general shape of the flow and its dependence on H_0; it is not accurate near the cluster, where $\bar{\delta} \gg 1$, and we exclude galaxies closer than 5 Mpc to the centre.

11.16.2 What a falling observer sees

We sit at the origin and move with the Local Group. A galaxy at position $\boldsymbol{x}$, at distance $r = |\boldsymbol{x}|$ in the direction $\hat{\boldsymbol{r}}$, moves with the Hubble flow, $H_0\boldsymbol{x}$ relative to us, plus its own peculiar velocity $\boldsymbol{u}(\boldsymbol{x})$. We measure velocities relative to ourselves, and we have peculiar velocity $\boldsymbol{u}(0)$. The radial velocity we record is therefore

$$v_{\rm obs} \overset{(1)}{=} \left[H_0\boldsymbol{x} + \boldsymbol{u}(\boldsymbol{x}) - \boldsymbol{u}(0)\right] \cdot \hat{\boldsymbol{r}} \overset{(2)}{=} H_0 r + v_{\rm inf}\, g(\boldsymbol{x}),$$
$$g(\boldsymbol{x}) \equiv \left[-\frac{d_V}{R}\,\hat{\boldsymbol{R}} - \hat{\boldsymbol{s}}_V\right] \cdot \hat{\boldsymbol{r}}, \qquad \boldsymbol{R} = \boldsymbol{x} - d_V\hat{\boldsymbol{s}}_V. \tag{11.102}$$

(1) Velocities relative to a moving platform are differences; only the radial component is measured from the Doppler shift.

(2) $\boldsymbol{x} \cdot \hat{\boldsymbol{r}} = r$; insert (11.101), and note that at the origin $\boldsymbol{R} = -d_V\hat{\boldsymbol{s}}_V$, with $\hat{\boldsymbol{s}}_V$ the unit vector from us to Virgo, so $\boldsymbol{u}(0) = v_{\rm inf}\hat{\boldsymbol{s}}_V$: we fall towards the cluster.

The function g is fixed by geometry once the position of Virgo is known (Figure 11.42). It has two parts. The second, $-\hat{\boldsymbol{s}}_V \cdot \hat{\boldsymbol{r}}$, is the reflex of our own fall, a dipole on the sky exactly like the solar motion of Section 11.14.2. The first is the pattern of the other galaxies' fall towards the cluster: positive in front of Virgo, where galaxies are pulled away from us, and negative behind it, where they are pulled towards us. Along the line of sight through Virgo the two together give the S-shaped curve in Figure 11.43 (left, red).

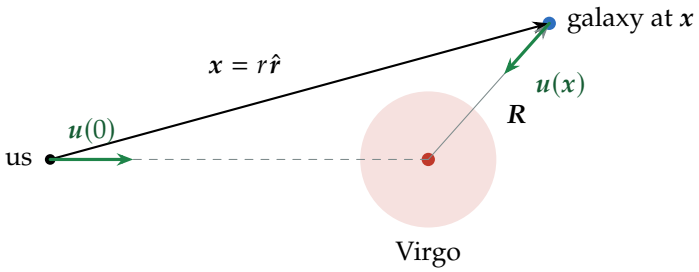


Figure 11.42: The geometry of the Virgocentric flow. In this model both the observed galaxy and our own fall towards the cluster, with infall speed proportional to $1/R$. The velocity we record for a galaxy is the radial part of its Hubble velocity plus the difference between its fall $\boldsymbol{u}(\boldsymbol{x})$ and ours $\boldsymbol{u}(0)$.

A model linear in its parameters. Equation (11.102) is linear in the two unknowns $(H_0, v_{\rm inf})$, with known regressors r_i and $g_i = g(\boldsymbol{x}_i)$. It is the solar-motion fit of Section 11.14.2 with a different second column. The noise has two parts. The galaxies' own motions about the smooth flow are small near us: within about ten megaparsecs, galaxies scatter around the local expansion by only some tens of km/s, a surprisingly "cold" flow, and we take $\sigma_{\rm th} = 60$ km/s. The distances carry fractional errors Δ, which by (11.92) add $(H_0\Delta r_i)^2$ to the variance. So the fit is weighted least squares with $w_i = 1/\left(\sigma_{\rm th}^2 + H_0^2\Delta^2 r_i^2\right)$: solve the 2×2 normal equations $A^{\mathsf{T}}WA\,\hat{\boldsymbol{p}} = A^{\mathsf{T}}W\boldsymbol{v}$, with rows (r_i, g_i) in A and $W = \mathrm{diag}(w_i)$, put the new $\hat{H}_0$ into the weights, and repeat until it stops changing; the covariance is $(A^{\mathsf{T}}WA)^{-1}$ (Section 3.10.5).

11.16.3 A simulated local sample

Script `code/ch11/10_virgo_infall.py` places 400 galaxies at random in the volume between 2 and 30 Mpc, outside 5 Mpc of Virgo, which sits at $d_V = 16.5$ Mpc in its true direction on the sky. It gives them velocities from (11.102) with $H_0 = 70$ and $v_{\rm inf} = 200$ km/s, adds the thermal scatter, scatters the distances lognormally, and fits. The observer computes g at the position implied by the *measured* distance, since the true one is unknown.

Listing 11.14: The weighted fit of $(H_0, v_{\rm inf})$, and one mock local sample (`code/ch11/10_virgo_infall.py`)

```
74 def fit(r_obs, g_obs, v, Delta, with_infall=True):
75     """Iterated weighted least squares for (H0, v_inf); returns estimates and covariance."""
76     A = np.column_stack([r_obs, g_obs]) if with_infall else r_obs[:, None]
77     H = 70.0
78     for _ in range(5):
79         w = 1.0 / (SIGTH ** 2 + (H * Delta * r_obs) ** 2)
80         F = A.T @ (w[:, None] * A)
81         p = np.linalg.solve(F, A.T @ (w * v))
82         H = p[0]
83     return p, np.linalg.inv(F)
84
85
86 def mock(rng, Delta, cone=None):
87     x = draw_positions(rng, NGAL, cone)
88     r = np.linalg.norm(x, axis=1)
89     v = H0 * r + VINF * infall_g(x) + rng.normal(0, SIGTH, len(r))
90     r_obs = r * np.exp(Delta * rng.standard_normal(len(r)))
91     # the observer evaluates g at the position implied by the measured distance
92     x_obs = x * (r_obs / r)[:, None]
93     return x, r, v, r_obs, infall_g(x_obs)
```

`fit` builds the two columns, iterates the weights five times and returns the estimate and the inverse of the weighted normal matrix. In `mock`, line 89 is (11.102) plus thermal noise, line 90 scatters the distances by $e^{\Delta\eta}$, and lines 92–93 evaluate g at the measured position.

Good distances. With $\Delta = 0.05$, the precision of distances from the tip of the red giant branch, one mock sample gives $\hat{H}_0 = 69.8 \pm 0.3$ and $\hat{v}_{\rm inf} = 187 \pm 7$ km/s, with a correlation coefficient of 0.61 between the two estimates. Over 2000 mock samples the estimates average 69.84 and 195, and scatter by 0.28 and 7; the error bars from $(A^{\mathsf{T}}WA)^{-1}$ are 0.27 and 7. The fit works, and its errors are right.

Ignoring the infall. Fit the same samples with H_0 alone, as if the flow were pure expansion. The estimates average 65.45 ± 0.38: biased low by more than four km/s/Mpc, sixteen times the statistical error. Where does the bias come from? The omitted term $v_{\rm inf} g_i$ is not zero on average over the sample. Our own fall is a dipole, which averages away over a full sky. The fall of the other galaxies does not: behind Virgo they move towards us, and there is much more volume behind the cluster, out to 30 Mpc, than in front of it. Least squares weights each galaxy by r_i^2, as in (11.91), so these distant infalling galaxies dominate, and the slope comes out low. A sample confined to within 40° of Virgo, as many early studies of the region were, does far worse: the one-parameter fit gives 54.94, while the joint fit on the same samples gives 68.14 ± 0.76.

Poor distances. With $\Delta = 0.2$, the precision of Tully–Fisher distances, the joint fit returns $\hat{H}_0 = 69.73 \pm 0.93$, still close to the truth, but $\hat{v}_{\rm inf} = 145 \pm 25$ km/s, two standard deviations below the true 200. The reason is the regressor: g was computed at the measured position,

which is the true position scattered by 20%. A noisy regressor pulls its coefficient towards zero, the regression dilution of (T2.101). Near Virgo, where g changes fastest with position, the effect is strongest. The same mechanism, galaxies assigned to the wrong place along a line of sight where the density changes rapidly, produces the spurious flows of Section 11.18.

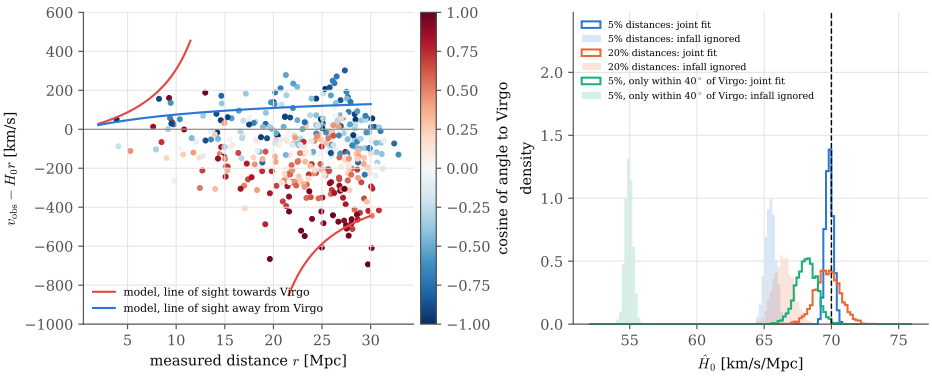


Figure 11.43: Left: one mock local sample with 5% distances. Each point is a galaxy's velocity minus the pure Hubble flow, coloured by the cosine of its angle from Virgo; the curves are the model along the line of sight towards Virgo (red, the S-shape through the cluster) and away from it (blue, mostly the reflex of our own infall). Right: $\hat{H}_0$ over 2000 mock samples, from the joint fit (outlines) and with the infall ignored (filled), for 5% and 20% distances over the whole sky and for 5% distances within 40° of Virgo. The dashed line is the truth.

11.16.4 Our motion through the CMB

There is one frame in which "at rest" has a meaning on cosmological scales: the frame in which the cosmic microwave background looks the same in every direction. An observer moving through it with speed $v = \beta c$ sees the radiation ahead Doppler-shifted to a higher temperature and the radiation behind to a lower one. To first order in β the temperature in a direction at angle θ to the motion is $T(\theta) = T_0(1 + \beta\cos\theta)$ (see the cosmology notes): a pure dipole of amplitude βT_0.

Example: the Sun's speed from the dipole. Planck measures a dipole amplitude of 3362.08 μK and a monopole $T_0 = 2.7255$ K [186, Table 3]. Then $\beta = 3362.08 \times 10^{-6}/2.7255 = 1.2336 \times 10^{-3}$, and $v = \beta c = 369.8$ km/s, towards Galactic longitude and latitude $(l, b) = (264.0°, 48.3°)$: Planck's 369.82 ± 0.11 km/s.

The Sun's velocity is the sum of its motion relative to the Local Group and the Local Group's motion relative to the CMB. The first is measured from the velocities of the Local Group members, as Hubble's solar motion was from his nebulae; Planck adopts 299 ± 15 km/s towards $(98.4°, -5.9°)$ [186, Table 3]. Converting each speed and direction to a vector, with the unit vector $(\cos b\cos l, \cos b\sin l, \sin b)$, and subtracting,

$$\boldsymbol{v}_{\mathrm{LG-CMB}} = \boldsymbol{v}_{\mathrm{Sun-CMB}} - \boldsymbol{v}_{\mathrm{Sun-LG}},$$

gives 620 km/s towards $(272°, 30°)$, which is Planck's 620 ± 15 km/s towards $(271.9°, 29.6°)$. The Local Group moves through the CMB at more than six hundred kilometres per second.

How much of it is Virgo? The direction of Virgo is $(283.8°, 74.5°)$, 45° from the Local Group's motion, so the component of the motion towards Virgo is 437 km/s. An infall of 200 km/s accounts for less than half of that, and subtracting a 200 km/s vector pointing at Virgo leaves 500 km/s pointing elsewhere. That remainder is produced by mass beyond the Local

Supercluster: in the 1980s, the peculiar velocities of elliptical galaxies revealed a large-scale streaming towards a concentration of mass named the Great Attractor [248], and the slow fall-off of the velocity correlation $\psi_\perp$ in Figure 11.40 says that contributions from a hundred megaparsecs and more are expected. Our own velocity is a sum over the whole volume, as (11.95) says it must be.

The lesson. The local flow is a field, not a scatter: modelling the infall into Virgo removes a bias many times larger than the statistical error on H_0, and poor distances bias the infall itself, because the regressor that describes the flow is evaluated at the wrong place.

11.17 How well can a local measurement know H_0?

Every measurement of H_0 from distances and redshifts uses tracers inside some region around us: Hubble's out to a couple of megaparsecs on his scale, the supernovae of the modern distance ladder out to $z \approx 0.15$, some 450 h^{-1}Mpc. Inside that region the matter density is not exactly the cosmic mean, and Section 11.16 showed that a mean contrast inside a sphere changes the expansion rate seen from its centre. So the "local H_0" we measure differs from the global one by an amount that depends on where we happen to live. That amount is random: had we lived elsewhere, it would be different. Its spread over all possible observers is the **cosmic variance** of a local H_0. It sets a floor under every local measurement, no matter how many tracers it uses, and it decides whether the gap between the ladder and the CMB could be the accident of living in an underdense region.

Where we have met H_0 before. The question was first asked in Section T2.12.3. There the ladder's 73.04 ± 1.04 lay 8.4% above Planck's 67.36 ± 0.54, and linear continuity, $\delta H_0/H_0 = -f\delta/3$, said that a void explaining it would have to be almost half empty, $|\delta| \approx 0.45$, over the whole volume of the Hubble-flow supernovae, against an expected scatter of 0.5 to 1% quoted from the literature. The physics here is the same, and so is that relation. What is new is the statistics: there the void was one sphere of chosen depth, here the density is a random field drawn from the CAMB spectrum, the same field that moved the galaxies of Section 11.15 and pulled us towards Virgo in Section 11.16, and the local H_0 is an estimator whose spread over observers we compute rather than quote. The mock samples of Section 3.17 and Section 11.16 asked whether a fit recovers a known $H_0 = 70$ from noisy tracers; here even perfect tracers give a local value that differs from the global one, and the question is by how much. The required depth comes out as 0.48 in Section 11.17.6, not 0.45, only because the gap is kept at 8.4% there instead of being rounded to 8%.

The sentence "the cosmic variance of H_0 inside a sphere of radius R" sounds complete, and is not. It hides a choice: the estimator that turns the tracers' velocities into one number. Two reasonable estimators, applied to the same tracers in the same sphere, have cosmic variances that differ by 12 to 17% at every radius, and two well-known papers quote, in effect, different estimators.

The question.

1. An observer fits H to the radial velocities of tracers inside a sphere of radius R. By how much does the answer differ from the global H_0, as a function of the density field?

2. What is the spread of that difference over observers, for the CAMB spectrum, and how does it depend on R and on the estimator?

3. Do we reproduce the numbers of Wojtak et al. (2014) and Marra et al. (2013)?

4. Could it explain the Hubble tension?

11.17.1 The estimator decides the window

The setup. We sit at the origin. Tracers fill the sphere $|\boldsymbol{x}| < R$ uniformly and densely. Velocities are measured in the frame of the CMB, so our own motion is removed (Section 11.16.4), and tracer i at distance r_i has radial velocity $v_i = H_0 r_i + \boldsymbol{u}(\boldsymbol{x}_i)\cdot\hat{\boldsymbol{r}}_i$, with $\boldsymbol{u}$ the linear velocity field of Section 11.15.2. We assume the distances are exact, so that the only departure from the Hubble law is the flow.

The least-squares slope. The estimator of (11.91) is $\hat{H} = \sum_i v_i r_i / \sum_i r_i^2$. Since $\boldsymbol{u}\cdot\hat{\boldsymbol{r}}\, r = \boldsymbol{u}\cdot\boldsymbol{x}$,

$$\hat{H} - H_0 = \frac{\sum_i \boldsymbol{u}(\boldsymbol{x}_i)\cdot\boldsymbol{x}_i}{\sum_i r_i^2} \;\longrightarrow\; \frac{\int_V \boldsymbol{u}\cdot\boldsymbol{x}\,\mathrm{d}^3x}{\int_V r^2\,\mathrm{d}^3x},$$

where the arrow is the limit of many tracers with uniform number density: each sum becomes the density times an integral over the sphere V, and the density cancels.

What could make the numerator nonzero? Only the divergence of the flow, as we now show. The product rule gives $\nabla\cdot(r^2\boldsymbol{u}) = r^2\,\nabla\cdot\boldsymbol{u} + 2\,\boldsymbol{u}\cdot\boldsymbol{x}$, since $\nabla r^2 = 2\boldsymbol{x}$. So

$$\begin{aligned}
\int_V \boldsymbol{u}\cdot\boldsymbol{x}\,\mathrm{d}^3x &\overset{(1)}{=} \frac{1}{2}\Big[\oint_{r=R} r^2\,\boldsymbol{u}\cdot\mathrm{d}\boldsymbol{S} - \int_V r^2\,\nabla\cdot\boldsymbol{u}\,\mathrm{d}^3x\Big] \overset{(2)}{=} \frac{1}{2}\int_V \big(R^2 - r^2\big)\,\nabla\cdot\boldsymbol{u}\,\mathrm{d}^3x \\
&\overset{(3)}{=} -\frac{H_0 f}{2}\int_V \big(R^2 - r^2\big)\,\delta\,\mathrm{d}^3x, \\
\frac{\hat{H} - H_0}{H_0} &\overset{(4)}{=} -\frac{f}{2}\,\frac{5}{4\pi R^5}\int_V \big(R^2 - r^2\big)\,\delta\,\mathrm{d}^3x \overset{(5)}{=} -\frac{f}{3}\,\delta_W, \qquad \delta_W \equiv \int \mathrm{d}^3x\; W_{\rm LS}(\boldsymbol{x})\,\delta(\boldsymbol{x}).
\end{aligned} \tag{11.103}$$

(1) Integrate the product rule over V and use the divergence theorem on $\nabla\cdot(r^2\boldsymbol{u})$.
(2) On the surface $r = R$, so $r^2 = R^2$ comes out, and $\oint \boldsymbol{u}\cdot\mathrm{d}\boldsymbol{S} = \int_V \nabla\cdot\boldsymbol{u}\,\mathrm{d}^3x$ by the divergence theorem again.
(3) The continuity equation (V1) of Section 11.15.2 today, $a = 1$, $H = H_0$.
(4) Divide by $\int_V r^2\mathrm{d}^3x = 4\pi\int_0^R r^4\mathrm{d}r = 4\pi R^5/5$ and by H_0.
(5) Define the normalised **window** $W_{\rm LS}(\boldsymbol{x}) = \frac{15}{8\pi R^5}(R^2 - r^2)$ for $r < R$ and zero outside. Its integral is $\frac{15}{8\pi R^5}\,4\pi\big(\frac{R^5}{3} - \frac{R^5}{5}\big) = 1$, so $\int_V (R^2 - r^2)\delta = \frac{8\pi R^5}{15}\delta_W$, and $\frac{f}{2}\cdot\frac{5}{4\pi R^5}\cdot\frac{8\pi R^5}{15} = \frac{f}{3}$.

Key idea: The local Hubble rate measures a weighted density

To linear order, a local estimate of H_0 is off by $\delta H/H_0 = -\frac{1}{3}f\,\delta_W$, where δ_W is the density contrast averaged with a window set by the estimator. For the least-squares slope through the origin the window is $\propto R^2 - r^2$ inside the sphere: the sphere's centre counts most.

The surprise is the shape. One might have guessed that the least-squares slope, which weights each tracer by r_i^2 and so favours the distant ones, measures the density near the edge. It measures the opposite. A tracer at the edge feels the mass inside it; a tracer near the centre feels the mass inside its own smaller sphere; the mass near the centre is felt by every tracer, so it accumulates the largest weight.

Other estimators, other windows. Averaging the individual ratios v_i/r_i, the **mean of ratios** $\hat{H} = N^{-1}\sum_i v_i/r_i$, goes the same way with $\boldsymbol{u}\cdot\boldsymbol{x}/r^2 = \boldsymbol{u}\cdot\nabla\ln r$ in place of $\boldsymbol{u}\cdot\boldsymbol{x}$, and the product rule for $\nabla\cdot(\boldsymbol{u}\ln r)$ gives the window $W_{\rm MR} = \frac{9}{4\pi R^3}\ln(R/r)$, even more concentrated at the centre (it diverges there, slowly). And if all tracers sat on the surface $r = R$, the slope would be $u_R(R)/R$, which by (11.100) is $-\frac{1}{3}H_0 f\,\bar{\delta}(<R)$: the mean contrast inside the sphere, a uniform

top-hat window $W_{\rm TH} = 3/(4\pi R^3)$. This is the "Hubble bubble" of Marra et al. and the local-void example of Section T2.12. All three share the same $-\frac{1}{3}f$; they differ only in the window (Figure 11.44, left).

11.17.2 From the window to the variance

The contrast δ_W is a linear functional of a Gaussian field, so it is Gaussian with mean zero and variance $\langle \delta_W^2 \rangle = \int \frac{\mathrm{d}^3k}{(2\pi)^3} P(k)\,|\tilde W(\boldsymbol{k})|^2$, where $\tilde W$ is the Fourier transform of the window (Section 7.6, where the same step gave σ_8). For a spherical window $\tilde W(k) = 4\pi \int_0^R W(r)\, r^2 j_0(kr)\,\mathrm{d}r$, the angular integral being the one that gave j_0 in Section 11.15.4. So

$$\sigma^2\!\left(\frac{\delta H}{H_0}\right) = \left(\frac{f}{3}\right)^2 \frac{1}{2\pi^2} \int_0^\infty \mathrm{d}k\; k^2 P(k)\, \tilde W^2(kR). \tag{11.104}$$

The transform of the parabolic window. We need one identity for spherical Bessel functions, $\frac{\mathrm{d}}{\mathrm{d}x}\big[x^{n+1} j_{n+1}(x)\big] = x^{n+1} j_n(x)$. For $n=0$: $x^2 j_1 = \sin x - x\cos x$, whose derivative is $x\sin x = x^2 j_0$. For $n=1$: $x^3 j_2 = (3-x^2)\sin x - 3x\cos x$, whose derivative is $-2x\sin x + (3-x^2)\cos x - 3\cos x + 3x\sin x = x\sin x - x^2\cos x = x^3 j_1$. With $x = kr$, $X = kR$,

$$\begin{aligned}
\int_0^R (R^2 - r^2) r^2 j_0(kr)\,\mathrm{d}r &\overset{(1)}{=} \frac{1}{k^5}\int_0^X (X^2 - x^2) x^2 j_0(x)\,\mathrm{d}x \\
&\overset{(2)}{=} \frac{1}{k^5}\Big[X^2 \cdot X^2 j_1(X) - \Big(X^4 j_1(X) - 2\int_0^X x^3 j_1(x)\,\mathrm{d}x\Big)\Big] \\
&\overset{(3)}{=} \frac{2X^3 j_2(X)}{k^5}, \\
\tilde W_{\rm LS}(kR) &\overset{(4)}{=} 4\pi\,\frac{15}{8\pi R^5}\,\frac{2X^3 j_2(X)}{k^5} = \frac{15\, j_2(kR)}{(kR)^2}.
\end{aligned} \tag{11.105}$$

(1) Substitute $r = x/k$: $R^2 - r^2 = (X^2 - x^2)/k^2$ and $r^2\mathrm{d}r = x^2\mathrm{d}x/k^3$.
(2) The first term is the $n=0$ identity integrated. For the second, integrate $x^2 \cdot x^2 j_0$ by parts, with $x^2 j_1$ as the antiderivative of $x^2 j_0$.
(3) The $X^4 j_1$ terms cancel, and the $n=1$ identity integrates $x^3 j_1$ to $X^3 j_2$.
(4) Multiply by 4π and the normalisation of $W_{\rm LS}$; $X^3/(R^5 k^5) = 1/X^2$.

A check: for small X, $j_2(X) \approx X^2/15$, so $\tilde W_{\rm LS} \to 1$, as a normalised window must. The same steps give $\tilde W_{\rm TH} = 3 j_1(kR)/(kR)$ (only the $n=0$ identity) and, integrating $\ln(R/r)$ by parts, $\tilde W_{\rm MR} = 9\big[\mathrm{Si}(kR) - \sin kR\big]/(kR)^3$, with Si the sine integral. Figure 11.44 (right) shows that the more a window concentrates at the centre, the further out in k its transform reaches, and the more small-scale power it lets into the variance.

Example: the numbers. Script `code/ch11/11_local_h0_variance.py` evaluates (11.104) with the CAMB spectrum and $f = 0.527$. At $R = 50\,h^{-1}$Mpc the rms of $\delta H/H_0$ is 2.73% for the top hat, 3.31% for the least-squares slope and 3.72% for the mean of ratios; at $100\,h^{-1}$Mpc, 1.17, 1.45 and 1.67%. The least-squares value falls to 0.83% at 150 and 0.28% at $300\,h^{-1}$Mpc, roughly as $R^{-1.5}$ (Figure 11.45, left). A percent-level local H_0 needs tracers out to at least $100\,h^{-1}$Mpc.

11.17.3 Checking the window on Gaussian boxes

The derivation used two idealisations: a continuum of tracers and the linear velocity field. The second we cannot test with a Gaussian simulation, but the first, and all the algebra, we can. The

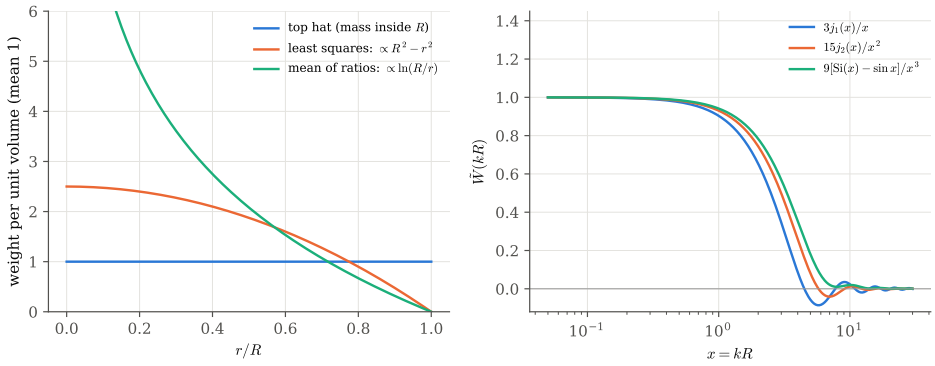


Figure 11.44: The three windows. Left: the weight each estimator gives to the density at radius r inside the sphere, normalised to unit mean over the volume. Right: their Fourier transforms. The parabolic window of the least-squares slope reaches higher k than the top hat, and the logarithmic window of the mean of ratios higher still.

script draws four Gaussian boxes of side 1500 h^{-1}Mpc with 256^3 cells, computes the velocity field as in Section 11.15.5, places 1000 observers at random, and lets each fit the least-squares slope to 3000 tracers inside spheres of five radii.

Listing 11.15: Observers at random positions fit H to tracers inside a sphere (`code/ch11/11_local_h0_variance.py`)

```
        for _ in range(nobs):
            x0 = rng.uniform(0, L, 3)
            for R in Rs:
                rr = R * rng.random(npts) ** (1 / 3)
                mu_ = rng.uniform(-1, 1, npts)
                ph = rng.uniform(0, 2 * np.pi, npts)
                sn = np.sqrt(1 - mu_ ** 2)
                nhat = np.column_stack([sn * np.cos(ph), sn * np.sin(ph), mu_])
                pos = (x0 + rr[:, None] * nhat) % L
                idx = np.floor(pos / D).astype(int) % N
                ur = sum(u[c][idx[:, 0], idx[:, 1], idx[:, 2]] * nhat[:, c] for c in
    range(3))
                meas[R].append(np.sum(ur * rr) / np.sum(rr ** 2) / 100.0)
```

Lines 136–140 draw tracers uniformly in the volume of the sphere: the radius as $R\,U^{1/3}$, so that the number within r grows as r^3, and an isotropic direction. Line 142 finds the cell of each tracer, line 143 takes the radial velocity there, and line 144 is the least-squares slope, divided by 100 km/s per h^{-1}Mpc to give $\delta H/H_0$.

The rms over observers is 3.31, 1.48, 0.86, 0.55 and 0.28% at $R = 50, 100, 150, 200, 300\,h^{-1}$Mpc, against 3.31, 1.45, 0.83, 0.54 and 0.27% from (11.104) summed over the box's modes. The simulation agrees with the prediction to within a few per cent, about what four boxes allow: the observers in one box share its long waves and are far from independent, so the effective number of independent samples is much smaller than 1000. The mean of $\delta H/H_0$ is zero within that scatter: its largest value, 0.16% at $R = 50\,h^{-1}$Mpc, is a small fraction of the 3.31% rms there.

11.17.4 Reproduction: Wojtak et al. (2014)

Their question. By 2013 the supernova measurement of H_0 had a 3% error and differed from Planck's by 2.4σ. Wojtak and collaborators asked how large the scatter of a local H_0 is in a realistic simulated universe, and how much it depends on where the observers sit: at random points, in dark-matter haloes like the one we live in, or in the centres of voids [249].

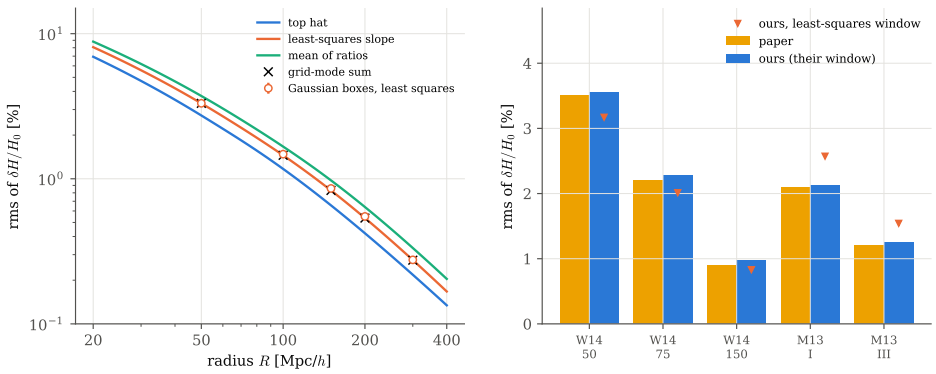


Figure 11.45: Left: the rms deviation of a local Hubble constant from the global one, for the three windows, with the least-squares estimator measured on Gaussian boxes (circles) and the grid-mode sum (crosses). Right: our reproduction of Wojtak et al. (2014) at three radii and of Marra et al. (2013), cases I and III. Orange bars: the paper; blue bars: our calculation with the paper's window; triangles: the same calculation with the least-squares window.

Their setup. An N-body simulation of a $(6\,h^{-1}\mathrm{Gpc})^3$ box in a WMAP-like cosmology ($\Omega_m = 0.27$, $\Omega_b = 0.044$, $h = 0.70$, $\sigma_8 = 0.80$, $n_s = 0.96$), with haloes identified in it. For each of 10^5 observers they fit their eq. (1), $\boldsymbol{u}\cdot\hat{\boldsymbol{r}} + H_0 r = H_{\mathrm{loc}} r$, to all haloes within $r_{\max}$, with equal weights. Their Table 1 gives the scatter of $H_{\mathrm{loc}}/H_0 - 1$ at $r_{\max} = 50$, 75 and 150 h^{-1}Mpc, from the simulation and from a linear-theory formula, their eq. (3), with $f = \Omega_m^{0.55}$.

Their equation, derived. Their eq. (3) reads, in our notation,

$$\sigma^2\!\left(\frac{\delta H}{H_0}\right) = \frac{f^2}{2\pi^2 R^2}\int_0^\infty \mathrm{d}k\, P(k)\left[\frac{3(\sin x - \mathrm{Si}\, x)}{x^2}\right]^2, \qquad x = kR.$$

This is (11.104) with the mean-of-ratios window:

$$\left(\frac{f}{3}\right)^2 \frac{1}{2\pi^2}\int \mathrm{d}k\, k^2 P\, \tilde{W}_{\mathrm{MR}}^2 \overset{(1)}{=} \frac{f^2}{9}\,\frac{1}{2\pi^2}\int \mathrm{d}k\, k^2 P\, \frac{81(\mathrm{Si}\, x - \sin x)^2}{x^6} \overset{(2)}{=} \frac{f^2}{2\pi^2 R^2}\int \mathrm{d}k\, P\, \frac{9(\sin x - \mathrm{Si}\, x)^2}{x^4}.$$

(1) Insert $\tilde{W}_{\mathrm{MR}} = 9(\mathrm{Si}\, x - \sin x)/x^3$.
(2) $k^2/x^6 = 1/(R^2 x^4)$ since $x = kR$; $81/9 = 9$; the square removes the sign.

So the formula they quote describes averaging the ratios u_r/r; the fit they ran on the haloes was the least-squares slope, whose window is the parabola.

What we compute. The CAMB linear spectrum for their cosmology, normalised to $\sigma_8 = 0.80$, and (11.104) with both windows and $f = 0.27^{0.55} = 0.487$.

Side by side.

$r_{\max}$ [h^{-1}Mpc]	50	75	150
Wojtak et al., linear theory (their eq. 3)	3.5%	2.2%	0.9%
ours, same formula (mean-of-ratios window)	3.55%	2.28%	0.97%
ours, least-squares window	3.17%	2.01%	0.83%
Wojtak et al., simulation, observers at random points	3.3%	2.1%	0.9%

Why they differ. Our evaluation of their formula agrees with their linear row to within the last digit they quote; the remaining few per cent are what a different transfer function or a rounding of their table would produce. The least-squares window, which is the estimator actually applied to their haloes, gives values 11–14% lower. Their simulation lies between

the two: at 150 h^{-1}Mpc, where linear theory should be good, it is 0.9% against our 0.83%; at 50 h^{-1}Mpc non-linear growth adds variance, and the authors note that voids, which fill most of the volume, push the mean of $H_{\rm loc}$ above H_0 there.

The lesson. "The cosmic variance of H_0 within R" is a property of an estimator, not of the sphere: the same tracers give a 12–17% larger scatter when their ratios are averaged than when a slope is fitted, because the two weight the density near us differently.

11.17.5 Reproduction: Marra et al. (2013)

Their question. In 2013 the local value was 73.8 ± 2.4 km/s/Mpc and the first Planck release gave 67.8 ± 0.8, a gap of $(73.8 - 67.8)/\sqrt{2.4^2 + 0.8^2} = 6.0/2.53 = 2.4$ standard deviations (the variance of a difference of independent measurements is the sum of the variances). Today's values, 73.04 ± 1.04 and Planck's final 67.36 ± 0.54, leave a gap of about the same size in km/s/Mpc but a much smaller error, so the same gap is 4.8 standard deviations (Section 11.20). How much of it can the local gravitational potential absorb, and what systematic error should every local measurement carry for it [250]?

Their setup. A Hubble bubble: a top hat, so that their eq. (1) is $\delta H/H = -\frac{1}{3} f\,\delta$ (in the linear regime, where their non-linear correction Θ equals one), which is (11.100). The contrast inside a sphere of radius R is Gaussian with the top-hat variance σ_R^2 (their eqs. 2–3). The supernovae lie at redshifts between $z_{\rm min}$ and 0.1, and each redshift z is assigned the sphere of comoving radius $R(z)$. The resulting variances are averaged over the redshift distribution $W_{\rm SN}(z)$ of the supernovae:

$$\sigma_{H_0}^2 = \int_{z_{\rm min}}^{z_{\rm max}} {\rm d}z\; W_{\rm SN}(z) \left(\frac{f}{3}\right)^2 \sigma_{R(z)}^2 \qquad \text{(their eq. 5)}.$$

Their Table I gives 2.1% for $z_{\rm min} = 0.010$ (case I) and 1.2% for $z_{\rm min} = 0.023$ (case III).

Their equation, derived. σ_R^2 is (11.104) without the $(f/3)^2$ and with $\tilde{W}_{\rm TH}$; eq. (5) is the variance of $\delta H/H$ averaged over the redshifts of the supernovae, each probing the sphere out to its own distance.

What we compute. The one input not given in the paper is $W_{\rm SN}(z)$. They state its medians: 0.025 for $z_{\rm min} = 0.010$ and 0.033 for $z_{\rm min} = 0.023$. A lognormal distribution in z, truncated to $[z_{\rm min}, 0.1]$, has two parameters, and the two medians fix them: median 0.0245 and logarithmic width 0.50 before truncation. We use the Planck 2018 spectrum rather than their Planck 2013 one; the matter densities differ by 2%.

Side by side. Case I: 2.13% against their 2.1%. Case III: 1.25% against 1.2%. With the least-squares window instead of the top hat the same averages would be 2.57% and 1.54%.

Why they differ. At the quoted precision they do not; our $W_{\rm SN}$ is a reconstruction, and a different shape with the same medians would move the results by a few hundredths of a per cent.

The lesson. A two-line linear-theory calculation, with the supernova distribution rebuilt from two quoted medians, reproduces the published systematic floor; the floor is about 1% once the nearest supernovae are excluded.

How this is used in real research

The Gaussian boxes *verify* (11.104): they check an analytic result in the regime where it should hold exactly. Wojtak et al.'s N-body simulation is the *research instrument*. It answers what no formula can: the scatter when observers live in haloes rather than at random points (lower $H_{\rm loc}$, because haloes sit in infall regions), when velocities are measured relative to the halo rather than the CMB, and when non-linear voids dominate the small spheres. The

same division of labour runs through the book: closed forms where they exist, checked by simulation; simulation alone where they do not (Chapter 6).

11.17.6 Could we live in a hole?

The ladder value 73.04 exceeds Planck's 67.36 by 8.4%. To produce that by (11.103), the density averaged over the ladder's window would need $\delta_W = -3 \times 0.084/f = -0.48$: a region half empty out to hundreds of megaparsecs, as Section T2.12 estimated. How unlikely is that? The ladder uses supernovae between $z = 0.023$ and 0.15. With Marra's distribution extended to $z_{\max} = 0.15$, the averaged scatter is 1.25% for the top hat and 1.54% for the least-squares window, so the gap is 5.5 standard deviations of the larger one. For a Gaussian, a fluctuation of more than five standard deviations in one direction has probability below 3×10^{-7}. A lognormal density field makes deep voids somewhat more probable than a Gaussian, as Marra et al. show, and observers at the centres of the ten largest voids in Wojtak et al.'s simulation see $H_{\rm loc}$ about 5% high out to $100\,h^{-1}$Mpc, falling to 1% at $200\,h^{-1}$Mpc; but at the ladder's distances both effects fall far short. Cosmic variance is a real systematic error of order 1% and should be in every local error budget [105, Sec. IV.F]; it is not the explanation of the Hubble tension.

11.18 Where is a galaxy whose distance we measured? Malmquist bias

A distance indicator such as the Tully–Fisher relation is calibrated so that it is right on average: take many galaxies at the same true distance r, measure them, and the logarithms of their inferred distances average to $\ln r$. The plain claim that follows seems harmless: *an unbiased indicator gives unbiased distances.* The everyday question it answers is: *if I take all the galaxies that my indicator places at distance d, are they on average at distance d?*

The question contains a hidden assumption. "Unbiased" was established by averaging over galaxies at a fixed *true* distance. The question averages over galaxies at a fixed *inferred* distance. Those are two different sampling rules, and they select two different sets of galaxies. The galaxies that end up labelled d came from a range of true distances, and how many came from each depends on how many galaxies there were at each true distance to begin with. In a uniform universe there are more galaxies at larger distances, simply because a shell of larger radius has more volume, so more of them scatter *in* to d from beyond than from in front. Near a cluster, where the number of galaxies changes rapidly along the line of sight, the effect is larger and changes sign from one side of the cluster to the other. Read as peculiar velocities, the mislabelled distances produce flows that do not exist. We met the first sign of this in Section 11.16.3, where poor distances weakened the fitted infall towards Virgo.

This is the **inhomogeneous Malmquist bias**. It is a cousin of the classical Malmquist bias of Section T2.13, where a flux limit selected the unusually bright objects; here no selection is needed at all.

The question.

1. Given an inferred distance d, what is the probability distribution of the true distance r?
2. What is the mean true distance $E(r \mid d)$ in a uniform universe, and near a density gradient?
3. Does a simulated line of sight through a cluster reproduce the bias found by Strauss and Willick?

4. What does the bias do to inferred peculiar velocities, and how should a peculiar velocity be estimated from a lognormal distance?

11.18.1 The distribution of the true distance

Inputs.

(M1) *Physical: lognormal errors.* For a galaxy at true distance r, the inferred distance is $d = re^{\Delta\eta}$ with $\eta \sim \mathcal{N}(0,1)$. A magnitude error σ in a distance modulus is a fractional distance error $\Delta = (\ln 10/5)\sigma \approx 0.46\sigma$, so 0.35 mag gives $\Delta = 0.161$.

(M2) *Physical: galaxies along the line of sight.* In a narrow cone around the line of sight, the number of galaxies between r and $r + \mathrm{d}r$ is proportional to $r^2 n(r)\,\mathrm{d}r$, with $n(r)$ their number density.

(M3) *Selection does not depend on the true distance.* Galaxies enter the sample through observed quantities (an apparent size, a line width) whose selection is the same at every true distance. Strauss and Willick show that the selection then cancels from what follows [251, Sec. 6.5.2]; when it does not, the classical Malmquist bias of Section T2.13 appears on top.

The event. Our indicator says "d". *What could have produced it?* Any true distance r: a galaxy at r produces the label d if its error happened to be $\eta = \ln(d/r)/\Delta$. *How probable was the label under each candidate?* By (M1), the density of d given r is lognormal, $p(d \mid r) = \exp\left[-\ln^2(d/r)/2\Delta^2\right]/(\sqrt{2\pi}\,\Delta\, d)$. *Only then the prior*: how many galaxies were at r to begin with, $r^2 n(r)$ by (M2). Bayes' theorem (5.5) puts them together:

$$p(r \mid d) = \frac{r^2 n(r)\, \exp\left[-\ln^2(r/d)/2\Delta^2\right]}{\displaystyle\int_0^\infty \mathrm{d}r'\, r'^2 n(r')\, \exp\left[-\ln^2(r'/d)/2\Delta^2\right]}, \qquad E(r \mid d) = \int_0^\infty r\, p(r \mid d)\, \mathrm{d}r. \tag{11.106}$$

The factor $1/d$ and the constants of the lognormal cancel between numerator and denominator, and the square of $\ln(d/r)$ equals that of $\ln(r/d)$. This is eq. (182) of Strauss and Willick, and the mean is their eq. (183). It depends on the density all along the line of sight, wherever the lognormal reaches: the bias at d is not a local property of the point d.

11.18.2 A uniform universe, and a slowly varying one

Uniform density. Let n be constant and change variable to $y = \ln r$, so $\mathrm{d}r = r\,\mathrm{d}y$ and the prior $r^2\mathrm{d}r$ becomes $r^3\mathrm{d}y = e^{3y}\mathrm{d}y$. Then

$$\begin{aligned} p(y \mid d) &\overset{(1)}{=\!\propto} e^{3y} \exp\left[-\frac{(y - \ln d)^2}{2\Delta^2}\right] \overset{(2)}{=\!\propto} \exp\left[-\frac{\left(y - \ln d - 3\Delta^2\right)^2}{2\Delta^2}\right], \\ E(r \mid d) &\overset{(3)}{=} E\left(e^y \mid d\right) \overset{(4)}{=} \exp\left(\ln d + 3\Delta^2 + \frac{\Delta^2}{2}\right) = d\, e^{7\Delta^2/2}. \end{aligned} \tag{11.107}$$

(1) (11.106) with n constant, written in y; $\ln(r/d) = y - \ln d$.

(2) Complete the square: $3y - \frac{(y-\ln d)^2}{2\Delta^2} = -\frac{(y-\ln d-3\Delta^2)^2}{2\Delta^2} + \text{const}$. So y is Gaussian with mean $\ln d + 3\Delta^2$ and variance Δ^2.

(3) $r = e^y$.

(4) For a Gaussian y with mean m and variance s^2, $E(e^y) = e^{m+s^2/2}$, the lognormal mean of (3.124).

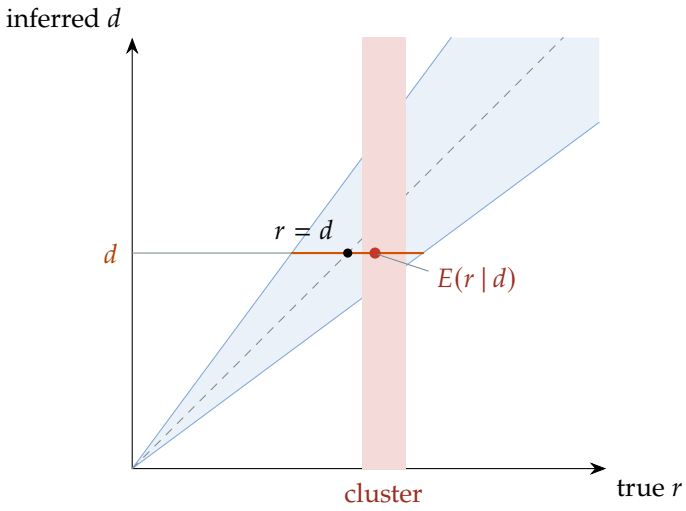


Figure 11.46: Conditioning on the inferred distance. Galaxies at true distance r are scattered vertically into the blue band of inferred distances. Fixing the label d keeps only the horizontal orange segment: the galaxies that could have produced that label. Where the segment crosses the cluster (red strip), there are many more galaxies than elsewhere, so the mean true distance of the galaxies labelled d is pulled towards the cluster.

This is eq. (184) of Strauss and Willick, first derived for galaxy distances by Lynden-Bell et al. [248]. For $\Delta = 0.161$ the factor is 1.095: galaxies labelled d are on average 9.5% farther away. Of the 7/2, three halves come from the volume factor and one half from the asymmetry of the lognormal.

A slowly varying density. If n changes little over the width of the lognormal, approximate it near $r = d$ by a power law, $n(r) \approx n(d)(r/d)^{\gamma}$, with the local logarithmic slope $\gamma(d) = \mathrm{d}\ln n/\mathrm{d}\ln r$ at $r = d$. The prior becomes $e^{(3+\gamma)y}$, the same chain runs with 3 replaced by $3 + \gamma$, and

$$E(r \mid d) \approx d\, e^{(7/2+\gamma)\Delta^2} \approx d\left[1 + \left(\frac{7}{2} + \gamma(d)\right)\Delta^2\right], \tag{11.108}$$

the second form for small Δ^2: Strauss and Willick's eq. (185). In front of a cluster the density rises with distance, $\gamma > 0$, and the bias is larger; behind it, $\gamma < 0$, and the bias is reduced or reversed.

When the approximation fails. A power law must hold across the whole lognormal, from $de^{-\Delta}$ to de^{Δ}, about ±16% of d. A cluster of width 500 km/s at 5000 km/s varies over 10% of its distance, and its $|\gamma|$ reaches 18: (11.108) then predicts corrections of order $\gamma\Delta^2 \approx 0.5$, absurdly large, and the full integral is needed.

11.18.3 Reproduction: Strauss and Willick's line of sight through a cluster

Their question. How large is the Malmquist bias near a realistic density peak, how good is the slowly varying approximation, and does the exact formula (11.106) predict it? [251, Sec. 6.5.2, Fig. 15]

Their setup. One line of sight. A uniform background density plus a Gaussian peak centred at 5000 km/s (distances in velocity units, $H_0 r$), of full width about 1200 km/s and maximum contrast 25. Galaxies follow pure Hubble flow: no true peculiar velocities at all. Distances come from a Tully–Fisher relation with 0.35 mag of scatter. They bin the galaxies in 150 km/s bins of inferred distance and plot the mean true distance in each bin.

Their equations. (11.106), (11.107) and (11.108), derived above, are their eqs. (183)–(185).

What we compute. We read "full width" as the full width at half maximum, which gives a Gaussian of standard deviation 510 km/s, and draw true distances with density $\propto r^2 n(r)$ so that the peak holds about 800 galaxies per 200 km/s bin, as in their upper panel; that makes 13446 galaxies in all. Each is given $d = re^{\Delta\eta}$. The exact curve is (11.106) integrated numerically:

Listing 11.16: The exact mean true distance at given inferred distance (code/ch11/12_malmquist.py)

```
45 def exact_mean(d, rr=np.linspace(RLO, RHI, 20000)):
46     """E(r|d) from the posterior p(r|d) ~ r^2 n(r) p(d|r), p(d|r) lognormal around r."""
47     out = []
48     for dd in np.atleast_1d(d):
49         lik = np.exp(-0.5 * (np.log(dd / rr) / DELTA) ** 2) / dd
50         w = rr ** 2 * n_of_r(rr) * lik
51         out.append(np.trapezoid(rr * w, rr) / np.trapezoid(w, rr))
52     return np.array(out)
```

Line 49 is the lognormal $p(d \mid r)$ on a grid of true distances, line 50 multiplies by the number of galaxies at each r, and line 51 takes the ratio of the first moment to the normalisation.

Side by side. Figure 11.47 has the two panels of their figure, with a third added. The simulated bins follow the exact curve: over the 44 bins with at least five galaxies, $\chi^2 = 66$, where a correct curve gives 44 ± 9. Some realisations give a value on the high side, and that is expected: each error bar is itself estimated from the few galaxies in its bin, and an error bar that comes out too small adds a lot to χ^2. Far from the peak they follow the uniform-density line $de^{7\Delta^2/2}$. In front of the peak, galaxies scattered out of the cluster pile up at smaller inferred distances, so the mean true distance at given d is too large; the excess over the uniform prediction reaches 655 km/s at $d \approx 3256$. Behind the peak the sign reverses, -1185 km/s at $d \approx 6583$. The slowly varying approximation oscillates wildly in between, exactly as theirs does. In their figure the amplitude of the effect is "500–1000 km/s"; ours is in that range in front and slightly larger behind.

Why they differ. Three things are not specified exactly: the width (we assumed FWHM), the number of background galaxies (their sample had a diameter limit, which thins the far background and so their high-d bins), and the scatter of a single realisation. None changes the shape: an S-shaped departure from the uniform-density line, centred on the peak.

The spurious flow. The galaxies have no peculiar velocities. An observer who takes the inferred distance at face value computes $v_{\rm pec} = cz - H_0 d = H_0(r - d)$, and the bottom panel shows what they find: galaxies in front of the peak appear to recede faster than the Hubble flow, galaxies behind it appear to approach. That is exactly the signature of infall onto a mass concentration, the S-shape of Figure 11.43, and its amplitude, hundreds of km/s, is that of a real infall. Uncorrected, the bias makes a peak look more massive than it is, and inflates the infall it is meant to measure. Every velocity analysis must either correct the distances with (11.106), which needs the density field along each line of sight, or work in redshift space, where the flow is modelled and the distances are the noisy data.

The lesson. An indicator that is unbiased at fixed true distance is biased at fixed inferred distance, because conditioning on what we measured reweights the possible true distances by how many galaxies were there.

11.18.4 A peculiar velocity from a lognormal distance

There is a second, smaller trap in $v_{\rm pec} = cz - H_0 d$, even for one galaxy at a known true distance. Take a galaxy in pure Hubble flow at r, so $cz = H_0 r$. Then $v_{\rm pec} = H_0 r\left(1 - e^{\Delta\eta}\right)$, a nonlinear

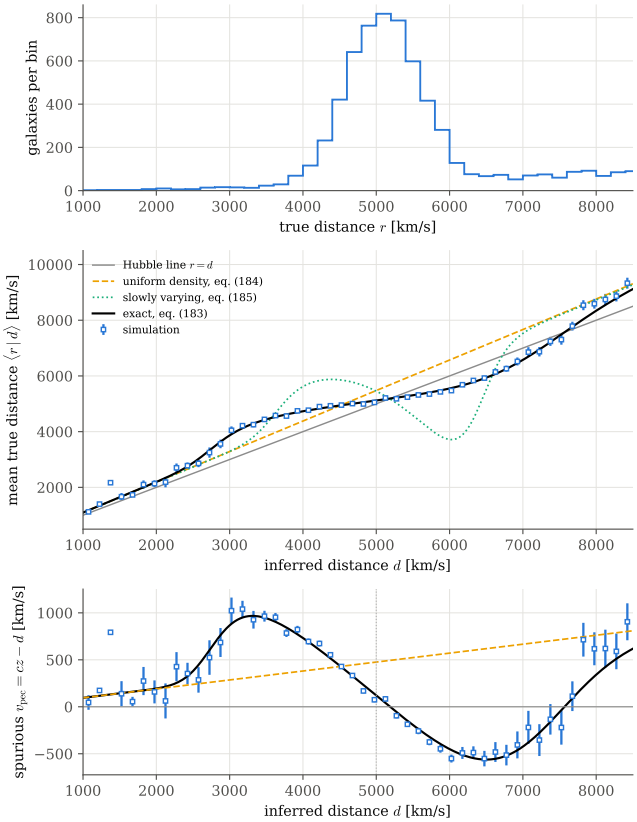


Figure 11.47: A line of sight through a density peak, after Strauss and Willick's Fig. 15. Top: the number of galaxies per 200 km/s bin of true distance. Middle: the mean true distance in 150 km/s bins of inferred distance (squares), with the Hubble line $r = d$, the uniform-density bias, the slowly varying approximation and the exact integral. Bottom: the same as an apparent peculiar velocity $cz - d$ for galaxies that in truth have none; the dotted line marks the peak.

function of the Gaussian η. Its mean is $H_0 r\left(1 - e^{\Delta^2/2}\right) < 0$ and its distribution has a long tail to negative values: at $r = 5000\,\mathrm{km/s}$ the mean is $-64\,\mathrm{km/s}$ and the skewness -0.49. The fix is to estimate the velocity from the logarithm,

$$\hat{v}_{\rm pec} = cz\,\ln\frac{cz}{H_0 d} \;=\; H_0 r\,\ln\frac{r}{d} = -H_0 r\,\Delta\eta,$$

which is linear in η: Gaussian, with mean zero (our simulation gives 2 km/s and skewness 0.00). For small velocities it agrees with $cz - H_0 d$ to first order, since $\ln(1+\epsilon) \approx \epsilon$. This is the estimator of Watkins and Feldman [252], used for the Cosmicflows-4 peculiar velocities [253]. It keeps the noise Gaussian, which is what a likelihood with Gaussian errors assumes; it does nothing about the Malmquist bias, which is a property of the population, not of the estimator.

11.19 Cosmic chronometers: the expansion rate from the ages of galaxies

Every method so far measured a distance. Hubble's slope was wrong because his distance scale was, and (11.93) showed that no fit can see a common error of scale. The modern ladder fixes the scale with geometry, rung by rung (Section T2.12), and the local flows and Malmquist bias of the last three sections are corrections to distances. A completely different route needs no distance at all, only a clock. The Universe expands as it ages, so if we could read the cosmic

time at two nearby redshifts, the rate at which redshift changes with time would give the expansion rate directly. Galaxies that stopped forming stars long ago are such clocks: their stars age passively, and their spectra change in a predictable way as they do.

Where we have met H_0 before. The chronometers first appeared in Section 3.17.6, as 30 mock points spread uniformly over $0.07 < z < 2$, made with $H_0 = 70$ and $\Omega_{m0} = 0.3$, each with an independent 8% error. The weighted ratio (3.142) found H_0 to 1.0 km/s/Mpc at fixed Ω_{m0}, and to about three times worse with Ω_{m0} profiled. The estimator and the profile are the same here; the data are not. The 32 real measurements of Moresco et al. lie mostly below $z = 1$, their errors run from about 5% to over 60% instead of a uniform 8%, and the stellar-population models they share add an error common to all of them, so the covariance is no longer diagonal and the weighted ratio becomes generalised least squares. No truth is supplied: the question is no longer whether the estimator recovers 70 but where the real H_0 lies relative to the ladder's 73.04 ± 1.04 and the CMB's 67.36 ± 0.54, 4.8σ apart (Section T2.12.3). Unlike the ladder, which borrows M (Section 3.17.4), or the CMB (Section 10.6.3) and the inverse ladder (Section T2.15), which borrow a sound horizon, a chronometer borrows only the rate at which a spectrum ages. The error at fixed Ω_{m0} comes out 1.6 rather than 1.0 because, weighted together, the 32 real points carry the information of about a dozen mock points with 8% errors, and the shared model error then adds a floor that no number of points removes.

The question.

1. Why does an age difference measure $H(z)$, and what sets the error of one measurement?
2. With a covariance between the points, what are the estimator of H_0 and its error, and what does a systematic error shared by all the points cost?
3. What do the 32 published measurements give for H_0 in flat ΛCDM, and how does it compare with the CMB and with the ladder?

11.19.1 A clock instead of a ruler

Why ages measure the expansion rate. The redshift of light emitted at cosmic time t and received today is $1 + z = 1/a(t)$, with a the scale factor normalised to one today, and the expansion rate is $H = \dot{a}/a$ (see the cosmology notes). Then

$$\frac{\mathrm{d}z}{\mathrm{d}t} \overset{(1)}{=} \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{a}\right) \overset{(2)}{=} -\frac{\dot{a}}{a^2} \overset{(3)}{=} -\frac{1}{a}\,\frac{\dot{a}}{a} = -(1+z)\,H(z), \qquad \text{so} \qquad H(z) = -\frac{1}{1+z}\,\frac{\mathrm{d}z}{\mathrm{d}t}. \tag{11.109}$$

(1) $z = 1/a - 1$, and the constant drops out of the derivative.
(2) The chain rule.
(3) Split a^2 and use $1/a = 1 + z$.

This is the starting point of Jimenez and Loeb [254, eq. (2)], and (3.141). Redshift is measured to a part in a thousand or better; the whole difficulty is $\mathrm{d}t$.

What is measured. Take two samples of the most massive galaxies that stopped forming stars early, one at redshift z_1 and one at a slightly larger z_2. If both formed their stars at the same early time, the difference of their stellar ages, Δt, is the cosmic time that elapsed between z_2 and z_1. With $\Delta z = z_2 - z_1$, (11.109) becomes $H(\bar{z}) \approx \Delta z / \left[(1+\bar{z})\,\Delta t\right]$ at the mean redshift $\bar{z}$. Two inputs are hidden in "the difference of their ages is the elapsed time": that the galaxies' clocks started together, and that no new stars were added since. A selection that admits younger galaxies at low redshift breaks the first (the progenitor bias), and a small young component breaks the second; both are checked against the data rather than assumed [255, Sec. 3.1.4].

Why a difference. The absolute age of a stellar population depends on the models used to interpret its spectrum, and different models disagree by a large fraction of a gigayear. An error that shifts both ages by the same amount cancels in Δt. That is the whole point of the differential method: Jimenez and Loeb proposed relative ages precisely because they are far more robust than absolute ones.

Example: the size of the clock signal. At $\bar{z} = 0.4$ in a flat universe with $H_0 = 68$ and $\Omega_{m0} = 0.315$, $H = 68\,E(0.4) \approx 85\,\mathrm{km/s/Mpc}$. The Hubble time $1/H$ is $977.8/85 \approx 11.6\,\mathrm{Gyr}$ ($977.8\,\mathrm{Gyr}$ is $1/(1\,\mathrm{km\,s^{-1}Mpc^{-1}})$). For $\Delta z = 0.1$, $\Delta t = \Delta z/[(1+\bar{z})H] = 0.1/1.4 \times 11.6 \approx 0.83\,\mathrm{Gyr}$. An error of 0.1 Gyr on the age difference is a 12% error on H. Two populations several gigayears old must be dated relative to each other to about a hundred million years.

The ratio, again. H is inversely proportional to Δt, and Δt is the noisy quantity. The inverse of a noisy number is biased: for a fractional error ε, $1/(1+\varepsilon) \approx 1 - \varepsilon + \varepsilon^2$, so the mean of $1/\Delta t$ is high by the fractional variance, 1.5% for a 12% error (Section 3.6.4). The published analyses work with the ages directly; the $H(z_i)$ and their errors that come out are treated as Gaussian, which is what we do too.

A spectral clock. In practice the age is not measured directly. Moresco and collaborators use the strength of the break in the spectrum at 4000 Å, which grows linearly with age for these galaxies, $D_n4000 = A \cdot \mathrm{age} + B$ [256, eq. (2.1)]. Then $\mathrm{d}t = \mathrm{d}D_n4000/A$ and $H = -A\,(1+z)^{-1}\,\mathrm{d}z/\mathrm{d}D_n4000$. The slope A comes from stellar population synthesis models. An error in A multiplies every $H(z)$ measured with the same models by the same factor. Unlike a common shift of the ages, a common error of the slope does not cancel in a difference. It is a scale error, exactly like Hubble's.

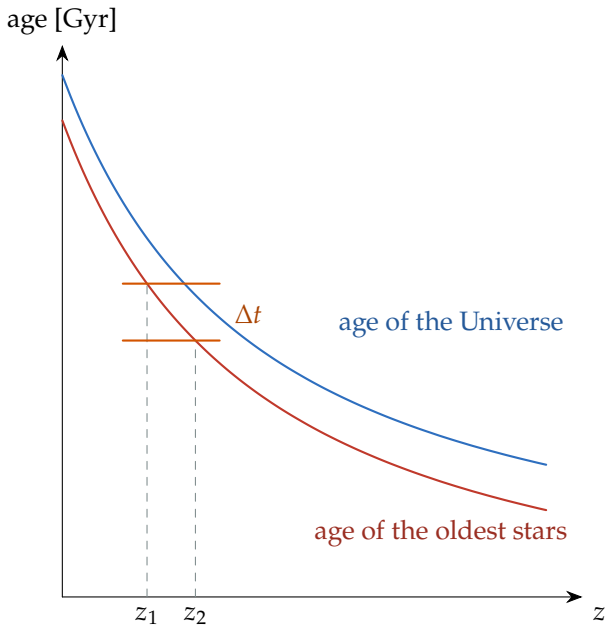


Figure 11.48: The differential-age idea, schematically. Passive galaxies formed their stars soon after the Big Bang, so the age of their oldest stars (red) runs parallel to the age of the Universe (blue). The difference Δt between the ages at two nearby redshifts z_1, z_2 is the cosmic time between them, and $\Delta z/\Delta t$ gives the expansion rate. A model error that shifts both ages by the same amount cancels; an error in the rate at which the spectra age does not.

11.19.2 The estimator with a covariance

A compilation delivers N redshifts z_i, measured rates H_i and a covariance matrix C. In flat ΛCDM the prediction is $H_0 E_i$, with $E_i = E(z_i;\Omega_{m0}) = \sqrt{\Omega_{m0}(1+z_i)^3 + 1 - \Omega_{m0}}$. For Gaussian errors the likelihood is $-2\ln\mathcal{L} = \chi^2 + \text{const}$, with $\chi^2 = (\boldsymbol{H} - H_0\boldsymbol{E})^{\mathsf{T}} C^{-1} (\boldsymbol{H} - H_0\boldsymbol{E})$ (Section 3.10.5). At fixed Ω_{m0}, H_0 enters linearly, so

$$\chi^2(H_0) \overset{(1)}{=} \boldsymbol{H}^{\mathsf{T}} C^{-1} \boldsymbol{H} - 2H_0\, \boldsymbol{E}^{\mathsf{T}} C^{-1} \boldsymbol{H} + H_0^2\, \boldsymbol{E}^{\mathsf{T}} C^{-1} \boldsymbol{E},$$
$$\hat{H}_0 \overset{(2)}{=} \frac{\boldsymbol{E}^{\mathsf{T}} C^{-1} \boldsymbol{H}}{\boldsymbol{E}^{\mathsf{T}} C^{-1} \boldsymbol{E}}, \qquad \operatorname{Var} \hat{H}_0 \overset{(3)}{=} \frac{1}{\boldsymbol{E}^{\mathsf{T}} C^{-1} \boldsymbol{E}}. \tag{11.110}$$

(1) Expand the quadratic form. The two cross terms are equal because C^{-1} is symmetric: $\boldsymbol{H}^{\mathsf{T}} C^{-1} \boldsymbol{E} = \boldsymbol{E}^{\mathsf{T}} C^{-1} \boldsymbol{H}$.
(2) Set $\mathrm{d}\chi^2/\mathrm{d}H_0 = 0$.
(3) $\hat{H}_0$ is a fixed linear combination $\boldsymbol{c}^{\mathsf{T}}\boldsymbol{H}$ with $\boldsymbol{c} = C^{-1}\boldsymbol{E}/(\boldsymbol{E}^{\mathsf{T}} C^{-1}\boldsymbol{E})$, so its variance is $\boldsymbol{c}^{\mathsf{T}} C \boldsymbol{c} = \boldsymbol{E}^{\mathsf{T}} C^{-1} C C^{-1} \boldsymbol{E} / (\boldsymbol{E}^{\mathsf{T}} C^{-1} \boldsymbol{E})^2$.

For a diagonal C this is the weighted ratio (3.142) and its variance (3.143). It is the same generalised least squares as the distance ladder (T2.106), with one parameter.

A systematic error that every point shares. Suppose the stellar-population slope A is uncertain by a fraction η. Then every rate is off by the same fraction: on top of independent errors with covariance $D = \operatorname{diag}(\sigma_i^2)$ there is a fully correlated part,

$$C = D + \eta^2\, \boldsymbol{m}\boldsymbol{m}^{\mathsf{T}}, \qquad m_i = H_0 E_i,$$

because a common fractional shift $\eta\xi$ (with ξ a standard Gaussian) moves point i by $\eta\xi\, m_i$, and the covariance of two such shifts is $\eta^2 m_i m_j$. What does it cost?

Inverting a matrix plus a rank-one term. For an invertible D and a vector $\boldsymbol{m}$,

$$\left(D + \eta^2 \boldsymbol{m}\boldsymbol{m}^{\mathsf{T}}\right)^{-1} = D^{-1} - \frac{\eta^2\, D^{-1}\boldsymbol{m}\boldsymbol{m}^{\mathsf{T}} D^{-1}}{1 + \eta^2\, \boldsymbol{m}^{\mathsf{T}} D^{-1} \boldsymbol{m}},$$

the Sherman–Morrison formula. To check it, multiply by $D + \eta^2\boldsymbol{m}\boldsymbol{m}^{\mathsf{T}}$ and write $s = \boldsymbol{m}^{\mathsf{T}} D^{-1}\boldsymbol{m}$: the product is $\mathbb{1} + \eta^2\boldsymbol{m}\boldsymbol{m}^{\mathsf{T}} D^{-1} - \eta^2 \boldsymbol{m}\boldsymbol{m}^{\mathsf{T}} D^{-1}(1+\eta^2 s)/(1+\eta^2 s) = \mathbb{1}$. Now let $S = \boldsymbol{E}^{\mathsf{T}} D^{-1} \boldsymbol{E}$, the Fisher information without the systematic. Since $\boldsymbol{m} = H_0 \boldsymbol{E}$,

$$\boldsymbol{E}^{\mathsf{T}} C^{-1} \boldsymbol{E} \overset{(1)}{=} S - \frac{\eta^2\, (\boldsymbol{E}^{\mathsf{T}} D^{-1} \boldsymbol{m})^2}{1 + \eta^2\, \boldsymbol{m}^{\mathsf{T}} D^{-1} \boldsymbol{m}} \overset{(2)}{=} S - \frac{\eta^2 H_0^2 S^2}{1 + \eta^2 H_0^2 S} \overset{(3)}{=} \frac{S}{1 + \eta^2 H_0^2 S},$$
$$\operatorname{Var} \hat{H}_0 \overset{(4)}{=} \frac{1}{S} + \eta^2 H_0^2. \tag{11.111}$$

(1) Sandwich the Sherman–Morrison formula between $\boldsymbol{E}^{\mathsf{T}}$ and $\boldsymbol{E}$.
(2) $\boldsymbol{E}^{\mathsf{T}} D^{-1} \boldsymbol{m} = H_0 S$ and $\boldsymbol{m}^{\mathsf{T}} D^{-1} \boldsymbol{m} = H_0^2 S$.
(3) Common denominator: $S(1 + \eta^2 H_0^2 S) - \eta^2 H_0^2 S^2 = S$.
(4) Invert and use (11.110).

The same steps applied to $\boldsymbol{E}^{\mathsf{T}} C^{-1} \boldsymbol{H}$ divide it by the same factor $1 + \eta^2 H_0^2 S$, so the estimate $\hat{H}_0$ itself does not change.

Key idea: A shared fractional error does not average down

If every measurement shares a common fractional error η, the variance of $\hat{H}_0$ is the statistical variance plus $(\eta H_0)^2$. Adding points shrinks the first term as $1/N$ and leaves the second untouched. It is the error of Hubble's distance scale, written as a covariance.

11.19.3 The data, and H_0 from them

Moresco et al. compile 32 measurements of $H(z)$ from cosmic chronometers between $z = 0.07$ and 1.97, obtained with three ways of reading the ages: fitting the full spectrum, Lick absorption indices, and the calibrated 4000 Å break [255, Table 1]. Their errors are the diagonal part of the covariance; the correlated part from the stellar-population models must be added. For that part we use one fully correlated fraction, $\eta = 4.5\%$, the average size of the stellar-population-model term found by Moresco et al. [257]; their recipe gives a separate, redshift-dependent value for each component, which we do not attempt.

z	H	σ	z	H	σ	z	H	σ	z	H	σ
0.07	69.0	19.6	0.27	77.0	14.0	0.4783	80.9	9.0	0.88	90.0	40.0
0.09	69.0	12.0	0.28	88.8	36.6	0.48	97.0	62.0	0.9	117.0	23.0
0.12	68.6	26.2	0.352	83.0	14.0	0.593	104.0	13.0	1.037	154.0	20.0
0.17	83.0	8.0	0.38	83.0	13.5	0.68	92.0	8.0	1.3	168.0	17.0
0.179	75.0	4.0	0.4	95.0	17.0	0.75	98.8	33.6	1.363	160.0	33.6
0.199	75.0	5.0	0.4004	77.0	10.2	0.781	105.0	12.0	1.43	177.0	18.0
0.20	72.9	29.6	0.425	87.1	11.2	0.875	125.0	17.0	1.53	140.0	14.0
			0.445	92.8	12.9				1.75	202.0	40.0
			0.47	89.0	49.6				1.965	186.5	50.4

Table 11.5: The cosmic-chronometer compilation of Moresco et al. (2022), Table 1: redshift, $H(z)$ and its error in km/s/Mpc.

Script `code/ch11/13_chronometers.py` types in the table and evaluates (11.110) directly:

Listing 11.17: The generalised least-squares estimate of H_0 at fixed Ω_{m0} (`code/ch11/13_chronometers.py`)

```
62 def gls(Hd, om, eta=0.0, H0ref=70.0):
63     """Closed-form GLS estimate of H0 at fixed Om; returns (H0_hat, sigma, chi2_min)."""
64     e = E(z, om)
65     for _ in range(3 if eta > 0 else 1):              # m depends weakly on H0: iterate
66         Ci = np.linalg.inv(cov(om, H0ref, eta))
67         S = e @ Ci @ e
68         h = (e @ Ci @ Hd) / S
69         H0ref = h
70     r = Hd - h * e
71     return h, 1 / np.sqrt(S), r @ Ci @ r
```

Lines 67–68 are (11.110). With the systematic switched on, the vector $\boldsymbol{m}$ in the covariance needs H_0, so lines 65–69 iterate: start from 70, compute $\hat{H}_0$, put it back into $\boldsymbol{m}$. Line 71 returns the estimate, its error $1/\sqrt{\boldsymbol{E}^{\mathsf{T}} C^{-1} \boldsymbol{E}}$ and $\chi^2_{\min}$.

At fixed matter density. With $\Omega_{m0} = 0.315$, $\hat{H}_0 = 68.3 \pm 1.6$ km/s/Mpc from the diagonal errors. With the correlated 4.5%, the estimate stays at 68.3 and the error grows to 3.5, exactly $\sqrt{1.6^2 + (0.045 \times 68.3)^2} = 3.5$ as (11.111) says. The systematic floor, 3.1 km/s/Mpc, is twice the statistical error: more chronometers of the same kind would not help.

Profiling the matter density. Ω_{m0} is not known in advance. At each value on a grid the closed form gives $\hat{H}_0(\Omega_{m0})$ and its $\chi^2_{\min}$, and the profile over Ω_{m0} (Section 3.14) gives $H_0 = 68.1$ with the $\Delta\chi^2 = 1$ interval $[65.1, 71.2]$, and $\Omega_{m0} = 0.32$ in $[0.26, 0.38]$. The interval on H_0 is twice as wide as at fixed Ω_{m0}: the data reach $z \approx 2$, where a higher matter density raises $E(z)$, and a lower H_0 compensates. With the correlated systematic the interval becomes $[63.9, 72.6]$ (Figure 11.49).

The χ^2 is too small. At the best fit, $\chi^2 = 14.6$ for 30 degrees of freedom. If the quoted errors were right and independent, would a χ^2 this small, or even smaller, be something we would reasonably expect? The χ^2_{30} distribution puts only 0.008 of its probability below 14.6 (a lower-tail test of the kind in Chapter 4). The points scatter about the curve by only 0.70 of their error bars. The likely reason is in the errors themselves: the published diagonal errors already include systematic contributions (metallicity, star-formation history), and those are not random from point to point. An honest reading is that the diagonal errors overstate the statistical scatter, and that the H_0 intervals above are, if anything, conservative.

Checking the error bars on mock compilations. The script draws 4000 compilations from $H_0 = 70$, $\Omega_{m0} = 0.315$ with the same redshifts and errors. Without the systematic, the estimates average 69.94 and scatter by 1.58, against 1.57 from (11.110). With a common fractional shift drawn once per compilation, they scatter by 3.48, against 3.52 from (11.111).

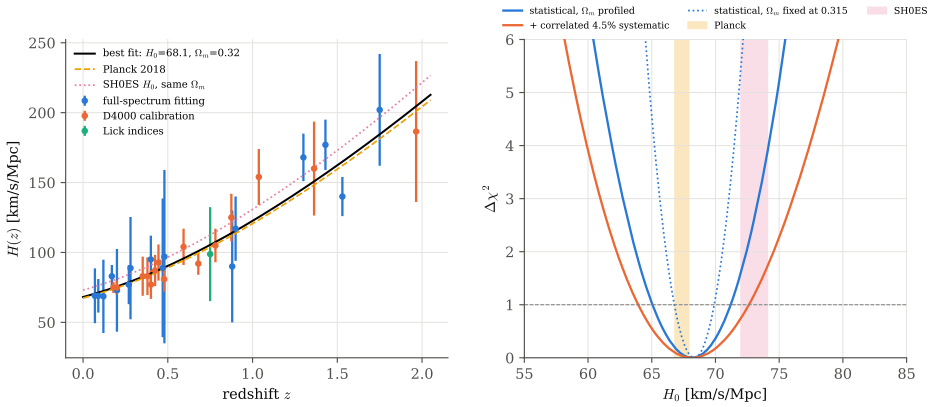


Figure 11.49: Left: the 32 chronometer measurements, coloured by how the ages were read, with the best fit of flat ΛCDM, the Planck 2018 expansion history, and the same history with the SH0ES value of H_0. Right: $\Delta\chi^2$ as a function of H_0 with Ω_{m0} fixed (dotted), profiled (blue), and profiled with the correlated systematic (orange); the bands are Planck and SH0ES at 68%.

Against the CMB and the ladder. Treat the chronometer interval as Gaussian with half-width 3.1 (statistical) or 4.3 (with the systematic). It differs from Planck's 67.36 ± 0.54 by 0.2 standard deviations of the difference and from SH0ES's 73.04 ± 1.04 by 1.5 (1.1 with the systematic), the variance of a difference of independent estimates being the sum of the variances. The chronometers side with the CMB, but at less than two standard deviations from the ladder: on their own they cannot decide the tension, as their compilers note [255, Sec. 3.1.5.1]. What they offer is independence: no distance ladder, and no early-Universe physics.

Three estimators, one structure. The supernova estimator of Section 3.17, Hubble's slope and the chronometer estimator are the same calculation. In each, the parameter enters the model linearly (as a shift of magnitudes, a multiplier of distances, a multiplier of $E(z)$), the

maximum-likelihood estimate is a weighted mean or ratio in closed form, and it sits on the Cramér–Rao bound. And in each, something else enters as a common factor that the fit cannot see: the absolute magnitude of the supernovae (Section 3.17.4), Hubble's distance scale, the stellar-population slope A. The statistics delivers H_0 only in units of a calibration that must come from outside the data.

The lesson. Chronometers measure $H(z)$ with a clock instead of a ruler and give $H_0 = 68.1 \pm 3.1$ km/s/Mpc in flat ΛCDM (±4.3 with the shared model error); their precision is limited not by the number of galaxies but by a shared model error, which a covariance carries and a weighted mean alone would hide.

11.20 One constant, many routes

The Hubble constant has come up in this book as 465, as 70, as 73.04, as 67.36, as 68.1 and as 69.9. Hubble's own slope, Section 11.14, was off from today's by a factor of 6.6, and nobody speaks of a tension about it. The ladder and the CMB differ by 8%, and the disagreement is called a crisis. What separates the two cases? Not the size of the gap. In Hubble's case the error bar was about 10% and the distance scale was unknown; in the modern case the error bars are 1.4% and 0.8%.

So a list of numbers cannot settle anything. For each number we need two answers. First, *what was actually observed, and what had to be brought in from outside to turn it into a value of H_0?* Second, *given the errors, would two such numbers be expected to differ by this much if they measured the same quantity?* The first is physics, and decides which routes can disagree for a reason. The second is statistics, and decides how surprised to be. We take them in that order.

11.20.1 What each route observes, and what it must borrow

H_0 is a velocity per length, km/s per Mpc. A redshift gives the velocity (at low redshift, $v = cz$, up to the peculiar velocities of Section 11.15). The length is never read off the sky. Every route observes something that is not H_0 and combines it with one quantity borrowed from outside the data. Four examples show the pattern, one for each way a length or a time enters.

A brightness, and the true brightness of the candle. The supernovae of Section 3.17.4 observe the apparent magnitude m at redshift z. At low redshift the luminosity distance is $D_L = cz/H_0$, and the distance modulus relates it to the absolute magnitude M of the candle, with D_L in Mpc. Call $\alpha \equiv m - 5\log_{10}(cz)$ the quantity the Hubble-flow supernovae determine:

$$\alpha \overset{(1)}{=} M + 5\log_{10} D_L + 25 - 5\log_{10}(cz) \overset{(2)}{=} M + 25 - 5\log_{10} H_0,$$
$$H_0 \overset{(3)}{=} 10^{(M+25-\alpha)/5}, \qquad \frac{\delta H_0}{H_0} \overset{(4)}{=} \frac{\ln 10}{5}\,\delta M. \tag{11.112}$$

(1) The distance modulus $m = M + 5\log_{10} D_L + 25$, subtracting $5\log_{10}(cz)$ from both sides.
(2) Put $D_L = cz/H_0$: the $5\log_{10}(cz)$ cancels, leaving $-5\log_{10} H_0$.
(3) Solve for H_0, with the observed α fixed.
(4) With α fixed, $\delta \ln H_0 = \ln 10\ \delta \log_{10} H_0 = \ln 10\, \delta M/5$, since $\log_{10} H_0 = (M + 25 - \alpha)/5$.

The data fix the combination $M - 5\log_{10} H_0$ and nothing else, the degeneracy of Section 3.17.4. The absolute magnitude M is the borrowed quantity. It comes from Cepheids anchored by geometry (Section T2.12). A shift of $\delta M = 0.18$ mag moves H_0 by about 8%, the size of the

Hubble tension. Hubble's 465 is the same equation with a wrong M: the distances he assumed, from the brightest stars of each nebula, were short by a factor 6.6 (Section 11.14.5), and $K = sH_0$ for a distance scale off by a factor s.

An angle, and the length of a ruler. A feature of known comoving length r (the sound horizon r_d for baryon acoustic oscillations, r_s at recombination for the CMB peaks) at comoving distance D_M subtends the angle $\theta = r/D_M$. In a flat universe $D_M = c\int_0^z \mathrm{d}z'/H(z') = (c/H_0)\,I(z)$, where $I(z) = \int_0^z \mathrm{d}z'/E(z')$ is a number once the shape of $E(z) = H(z)/H_0$ is fixed. Then

$$\theta \overset{(1)}{=} \frac{r}{D_M} \overset{(2)}{=} \frac{rH_0}{c\,I(z)}, \qquad H_0 \overset{(3)}{=} \frac{c\,I(z)\,\theta}{r}, \qquad \frac{\delta H_0}{H_0} \overset{(4)}{=} -\frac{\delta r}{r}. \tag{11.113}$$

(1) The definition of the subtended angle of a ruler of length r at distance D_M.
(2) Substitute $D_M = cI/H_0$.
(3) Solve for H_0.
(4) Hold the observed θ and the integral I fixed: $\ln H_0 = \ln(cI\theta) - \ln r$.

The borrowed quantity is the ruler r, a length computed from the physics before recombination (Section T2.15). The inverse ladder of Section T2.15 is this equation with $r_d = 147.09$ Mpc. For the CMB the integral I also depends on H_0 through $\Omega_{m0} = \omega_m/h^2$ (Section 10.6.3), so H_0 is found by solving $\theta_{\rm MC} = r_s/D_M$ rather than by rearranging it; the lever on the ruler remains. The CMB borrows a second thing: to carry the result from $z \approx 1100$ to today it assumes the flat ΛCDM expansion history.

A redshift step, and a time step. Chronometers observe the difference Δz between two passively ageing galaxy populations and the difference Δt of their ages, and give $H(z) = -\Delta z/[(1+z)\Delta t]$ (Section 3.17.6); $H_0 = H(z)/E(z)$ then needs the shape of E. The borrowed quantity is the age scale: stellar-population models turn a spectrum into an age, and their error is a common fractional factor $\eta \approx 4.5\%$ (Section 11.19.2). There is no distance in it at all.

A velocity profile, and the flow we are caught in. A fit of velocity against distance in our own neighbourhood needs, besides H_0, the peculiar flow. The joint fit of Section 11.16 recovers H_0 together with the infall speed, which is the borrowed quantity here; the same sample fitted without the infall comes out low. The finite volume adds a random error on top, the cosmic variance of Section 11.17.

The shape they share. Write any route as $H_0 = O/X$, the observable O divided by a borrowed quantity X. Take logarithms: $\ln H_0 = \ln O - \ln X$. If the errors of O and X are independent and small, their variances add:

$$\left(\frac{\sigma_{H_0}}{H_0}\right)^2 = \left(\frac{\sigma_O}{O}\right)^2 + \left(\frac{\sigma_X}{X}\right)^2. \tag{11.114}$$

Observing more supernovae, more galaxies or more acoustic peaks shrinks the first term as $1/N$. The second term is the error on one number that every measurement in the route shares, and it does not shrink at all, as in the shared factor η of Section 11.19.2. The routes differ in which number they borrow, and each route's borrowed number is a different physical quantity with its own independent error. Table 11.6 lists them; Table 11.7 gives the values obtained in the book and in the literature.

Route	Observed	Borrowed from outside	Derived in
Toy Hubble diagram, Hubble 1929	velocities and brightnesses of galaxies	the absolute brightness of the candle (the distance scale)	Sections 3.17 and 11.14
Distance ladder	magnitudes of Cepheids and of Hubble-flow supernovae	M from geometric anchors, one common value across rungs	Sections T2.12 and 3.17.4
CMB	angles and heights of the acoustic peaks	r_s from early-Universe physics, and flat ΛCDM from $z \approx 1100$ to now	Section 10.6.3
BAO, inverse ladder	angle and redshift extent of the baryon feature in galaxy catalogues	r_d, from the CMB or from nucleosynthesis, and the shape of $E(z)$	Sections T2.15 and 11.4.6
Cosmic chronometers	redshift steps and age steps of old galaxies	stellar-population models for the ages ($\eta \approx 4.5\%$)	Sections 3.17.6 and 11.19
Local infall	velocities and distances in our neighbourhood	the infall pattern $v_{\rm inf}\, g(\boldsymbol{x})$	Section 11.16
Local cosmic variance	the density contrast in the volume used	an error floor, not a route	Section 11.17
Combining a chain with a prior on H_0	a posterior sample and a new Gaussian measurement	reweighting, valid only if the two overlap	Section 9.23.1

Table 11.6: The routes to H_0 in the book: what each observes and what it must bring in from outside to become a Hubble constant.

Route	The book's own value	Published value
Toy Hubble diagram, supernovae mock	mock truth 70, recovered at the Cramér–Rao bound	–
Hubble 1929, 24 nebulae	465 ± 56	465 ± 50 (his table, solar-motion solution)
Local infall, mock sample	69.9 ± 0.3 with the infall, 65.5 without	–
Distance ladder	mock truth 73, fit 73.90 ± 1.10	SH0ES: 73.04 ± 1.04
CCHP TRGB ladder	–	69.8 ± 1.7
CMB	Fisher forecast $\sigma = 0.53$ (TT,TE,EE+lowE)	Planck 2018: 67.36 ± 0.54 (0.60 without lensing)
BAO, inverse ladder	69.2 ± 0.9 with the CMB r_d	DESI: 69.29 ± 0.87 (CMB r_d), 68.53 ± 0.80 (BBN r_d)
Cosmic chronometers	68.1 ± 3.1 (±4.3 with η)	–
Chain of Section 9.23.1	67.06 ± 0.93	–

Table 11.7: The values of H_0 in km/s/Mpc. A mock value is a number the script was told to find and then found, not a measurement. The BAO and CMB values share the sound horizon when the CMB r_d is used, and are not independent.

11.20.2 Is the gap too big? The same gap in three sets of error bars

Two routes give $H_A \pm \sigma_A$ and $H_B \pm \sigma_B$, and the errors are independent. Suppose both measure the same constant. Then $D = H_A - H_B$ has mean zero. The variance of a difference of independent variables is the sum of their variances, so D has standard deviation $\sigma_D = \sqrt{\sigma_A^2 + \sigma_B^2}$, and the surprise is $|D|/\sigma_D$ standard deviations. For the ladder and the CMB:

$$\frac{|D|}{\sigma_D} \overset{(1)}{=} \frac{73.04 - 67.36}{\sqrt{1.04^2 + 0.54^2}} \overset{(2)}{=} \frac{5.68}{\sqrt{1.082 + 0.292}} \overset{(3)}{=} \frac{5.68}{1.172} = 4.8. \tag{11.115}$$

(1) SH0ES against Planck, with their quoted errors.
(2) The squares of the errors.
(3) Their sum has square root 1.172.

For a Gaussian, a difference this many standard deviations from zero in either direction has probability of order 10^{-6}. This is the value we use for the Hubble tension throughout. Other places in the book get different numbers, with the same gap and different errors:

Pair	Gap [km/s/Mpc]	σ_D	$\lvert D\rvert/\sigma_D$	Where
SH0ES 73.04 ± 1.04 against Planck 67.36 ± 0.54	5.68	1.17	4.8	Section T2.12.3
SH0ES against the book's chain 67.06 ± 0.93	5.98	1.40	4.3	Section 9.23.1
Riess 2011 73.8 ± 2.4 against Planck 2013 67.8 ± 0.8	6.0	2.53	2.4	Section 11.17.5

Table 11.8: One gap of about 6 km/s/Mpc, three significances. The gap has barely moved in a decade. The error has shrunk from 2.5 to 1.2.

The 4.3 differs from 4.8 only because our chain, fitted to one simulated temperature spectrum, is wider than Planck's final result (0.93 against 0.54): a larger σ_D under the same gap. The 2.4 is the same arithmetic on the numbers of 2013: the ladder then had an error of 2.4 and Planck 0.8. A gap of 6 which was 2.4 standard deviations in 2013 has become 4.8 as the error bars shrank, while the gap stayed. That is the sense in which the tension is a statement about precision.

The other routes, against both camps. The same arithmetic gives where each independent route sits relative to the two:

Route	H_0	against SH0ES	against Planck
CCHP TRGB	69.8 ± 1.7	1.6σ	1.4σ
BAO with BBN r_d (DESI)	68.53 ± 0.80	3.4σ	1.2σ
Chronometers, ours (statistical)	68.1 ± 3.1	1.5σ	0.2σ
Chronometers, ours (with η)	68.1 ± 4.3	1.1σ	0.2σ

Table 11.9: Significance of each route's difference from the ladder and from the CMB, each with the rule $|D|/\sqrt{\sigma_A^2 + \sigma_B^2}$.

For instance, BAO with a big-bang-nucleosynthesis ruler against the ladder is $(73.04 - 68.53)/\sqrt{1.04^2 + 0.80^2} = 4.51/1.31 = 3.4$. The chronometers are consistent with everyone, and so decide nothing. The independent BAO value, which uses neither the CMB peaks nor the ladder, sits with the CMB.

11.20.3 Two calibrations, not two expansion rates

The expansion rate today is one number. What the tension compares is two chains of inference, each ending in that number, each with its own borrowed quantity. The late route, the ladder, borrows the absolute magnitude M and a flow model, and assumes that supernovae are the same kind of candle in Cepheid hosts and in the Hubble flow. The early route borrows the sound horizon r and the history of expansion from $z \approx 1100$. If the two disagree, by (11.112) and (11.113) some borrowed quantity is off:

$$\delta M \overset{(1)}{=} \frac{5}{\ln 10} \ln \frac{73.04}{67.36} = 2.171 \times 0.0810 = 0.18 \text{ mag},$$
$$r_d \overset{(2)}{=} 147.09 \times \frac{69.29}{73.04} = 139.5 \text{ Mpc} \quad \text{(a 5.1\% shorter ruler)}. \tag{11.116}$$

(1) Invert (11.112): $\delta H_0/H_0 = \ln(73.04/67.36) = 0.0810$ must come from a shift of M; this is $5 \log_{10}(73.04/67.36) = 0.18$ in magnitudes.

(2) By (11.113), $H_0 \propto 1/r_d$ with θ fixed, so a ruler of $147.09 \times 69.29/73.04$ would turn the inverse ladder's 69.29 into 73.04.

Either a candle 0.18 mag brighter than assumed, or a ruler 5% shorter. These are the two sizes of the problem. Neither can be tested on the ladder or on the CMB alone, because each calibration is a single number that all its data share (Eq. (11.114)). That is why the way forward is a third route with a different borrowed quantity: chronometers borrow neither, BAO borrows r_d but from nucleosynthesis, and the tip of the red-giant branch borrows a different candle. They sit between, closer to the CMB, with errors still too large to say that the ladder is wrong.

What local cosmic variance can explain. The ladder's volume is finite, and a density contrast averaged over it changes the expansion seen from inside (Eq. (11.100)). For the ladder's supernovae out to $z = 0.15$ the spread of that effect is 1.25 to 1.5% (Section 11.17.6), about 1.1 km/s/Mpc, the size of the ladder's own error. Suppose none of it were in the ladder's quoted 1.04. Then the variances add: $\sqrt{1.04^2 + 1.12^2} = 1.53$, and

$$\frac{5.68}{\sqrt{1.53^2 + 0.54^2}} = \frac{5.68}{1.62} = 3.5.$$

The tension would fall from 4.8 to 3.5 standard deviations. What cosmic variance cannot do is produce the gap itself: 8.4% is 5.5 times the spread, which would need the ladder's volume to be half empty, a fluctuation whose Gaussian probability is below 10^{-7}. Nor does it touch the CMB, the BAO ruler or the chronometers at z up to 2, which use volumes far larger. Local effects bite hardest on the nearest samples: in the mock sample of Section 11.16, within 30 Mpc, ignoring a 200 km/s infall moved H_0 from 69.9 to 65.5, 6%, as large as the tension, but the ladder's supernovae lie beyond $z = 0.023$, where the same 200 km/s is a few per cent of each redshift, and are corrected with a flow model (Section 11.21.3).

The one-line lesson. Every route to H_0 is an observable divided by a borrowed calibration, the calibration does not average down, and the Hubble tension is the statement that two calibrations, a luminosity and a ruler, disagree by 0.18 mag, or 5%, at 4.8 standard deviations.

11.21 How the local Universe is measured in research

The Cosmicflows-4 catalogue lists 55,877 galaxies gathered into 38,065 groups [253]. Each line holds a position on the sky, a redshift in the frame of the cosmic microwave background, a distance with its error, and a peculiar velocity. Every idea we have met since Hubble's diagram is somewhere behind those columns: Hubble's fit of velocity against distance (Section 11.14), the velocity field that structure induces (Section 11.15), our own fall towards Virgo and our motion through the microwave background (Section 11.16), the scatter of a locally measured H_0 (Section 11.17) and the bias of a distance read off a noisy indicator (Section 11.18). The SH0ES distance ladder and the chronometer compilation use the same ideas. Our scripts ran each of them once, on Hubble's 24 nebulae, on a few hundred simulated galaxies, or on a table of 32 measurements. The question, in words:

> *At each step from a catalogue of distances, redshifts or ages to a published H_0 or velocity field, what does a research analysis do that our scripts left out, and which part of its error can more data shrink?*

Figure 11.50 shows the three routes. The top row is the distance ladder: geometry calibrates nearby stars, nearby stars calibrate brighter indicators, and those reach the Hubble flow. The middle row is the map of galaxies that predicts velocities, which the ladder uses to clean its redshifts. The bottom row is the chronometer route, which uses no distances at all.

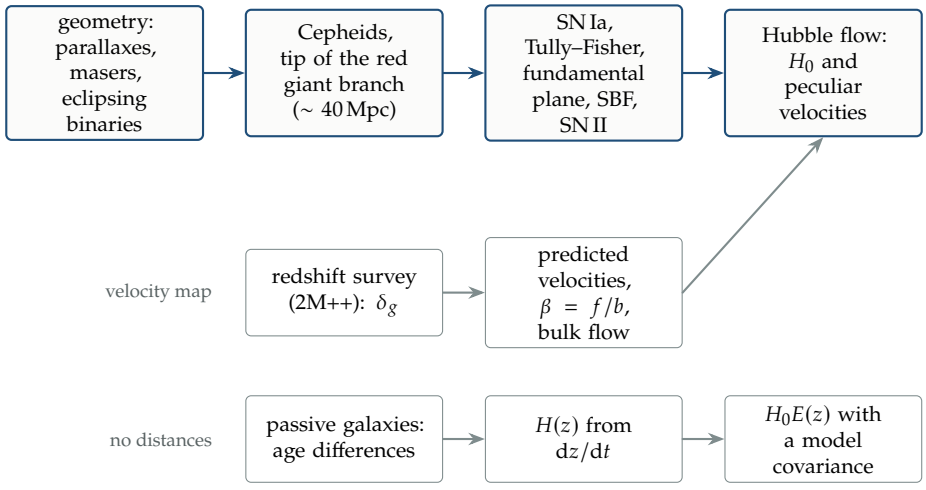


Figure 11.50: Three routes to the local expansion. Top: the distance ladder, each rung calibrating the next. Middle: a map of galaxy positions predicts peculiar velocities, which are subtracted from the redshifts of the Hubble-flow objects. Bottom: cosmic chronometers turn age differences of passive galaxies into $H(z)$ without any distance.

11.21.1 Tens of thousands of distances: Cosmicflows-4

Cosmicflows-4 combines eight ways of measuring a distance [253, Sec. 1]. Most galaxies have a distance from the Tully–Fisher relation [the luminosity of a spiral galaxy predicted from its rotation speed] or the fundamental plane [the same idea for an elliptical galaxy, using its velocity dispersion and surface brightness]; each such distance is uncertain by 20–25%. Type Ia supernovae give about 7% out to $z = 0.1$, type II supernovae about 15%, surface-brightness fluctuations about 5% out to $z \approx 0.03$. Cepheids and the tip of the red giant branch give about 5% but only within about 20 Mpc; masers give six geometric distances.

Putting eight methods on one scale. Every subsample s (one Tully–Fisher survey, say) measures distance moduli with a zero point of its own. Call $\mathrm{DM}_n^{(s)}$ the modulus of object n (a group or a single galaxy) in subsample s, $\Delta\mu_s$ an offset added to every modulus of that subsample, and $\langle \mathrm{DM} \rangle_n$ the weighted mean of the adjusted moduli $\mathrm{DM}_n^{(s)} + \Delta\mu_s$ over all subsamples that contain object n. With independent Gaussian errors $\sigma_{n,s}$ on each difference, the log-likelihood of the offsets is [253, eqs. (1)–(2)]

$$-2 \ln \mathcal{L}(\Delta\boldsymbol{\mu}) = \sum_n \sum_{s \ni n} \frac{\left(\mathrm{DM}_n^{(s)} + \Delta\mu_s - \langle \mathrm{DM} \rangle_n\right)^2}{\sigma_{n,s}^2} + \text{const.} \tag{11.117}$$

For given offsets it asks how plausible the fluctuations are that must be added to the adjusted moduli to make every object have one distance, the likelihood of Chapter 3. With a flat prior the posterior is this likelihood, and the Cosmicflows team samples it with `emcee`, 128 walkers of 10^4 steps.

Two things follow from what we already know. First, each offset enters (11.117) linearly, and so does $\langle \mathrm{DM} \rangle_n$, which is a weighted mean of terms linear in the offsets. The exponent is therefore quadratic in $\Delta\boldsymbol{\mu}$, the posterior is exactly Gaussian, and its mean and covariance follow from one linear solve, as for every linear model since Section 3.17; the chain is a check of that solve. Second, the overlaps cannot fix the overall zero point. Add the same constant c to every offset. With weights $w_{n,s}$,

$$\langle \mathrm{DM} \rangle_n(\Delta\boldsymbol{\mu} + c) \overset{(1)}{=} \frac{\sum_s w_{n,s}\left(\mathrm{DM}_n^{(s)} + \Delta\mu_s + c\right)}{\sum_s w_{n,s}} \overset{(2)}{=} \langle \mathrm{DM} \rangle_n(\Delta\boldsymbol{\mu}) + c,$$

$$\mathrm{DM}_n^{(s)} + \Delta\mu_s + c - \langle \mathrm{DM} \rangle_n(\Delta\boldsymbol{\mu} + c) \overset{(3)}{=} \mathrm{DM}_n^{(s)} + \Delta\mu_s - \langle \mathrm{DM} \rangle_n(\Delta\boldsymbol{\mu}). \tag{11.118}$$

(1) The definition of the weighted mean, with every offset shifted by c.
(2) Split the sum: the c terms give $c \sum_s w_{n,s} / \sum_s w_{n,s} = c$.
(3) Subtract the first line: the two c's cancel.

Every residual, and so the likelihood, is unchanged: one direction in the space of offsets is flat. Cosmicflows fixes it by setting the offset of its own Tully–Fisher sample to zero. This is the third time the same thing has happened. Hubble's slope could not see a common factor in his distances, (11.93), and a supernova Hubble diagram cannot see the absolute magnitude (Section 3.17.4). The absolute scale has to come from outside: 542 calibrator galaxies with distances from the tip of the red giant branch, Cepheids or masers [253, Sec. 10].

Example: how small is the statistical error on H_0? Each group i gives a Hubble parameter $H_i = f_i V_i / d_i$, its recession velocity V_i in the microwave background frame over its distance d_i, with f_i a small correction to the low-redshift Hubble law. Beyond 4000 km/s, where a peculiar velocity of 300 km/s is less than 8% of V_i, the mean of $\log_{10} H_i$ over about 35,000 groups is 1.8725, that is $H_0 = 74.6$ km/s/Mpc, and the scatter is 0.091 in $\log_{10} H$ [253, Sec. 13]. How well does the mean pin down H_0 if the groups are independent?

$$\sigma_{\log H_0} \overset{(1)}{=} \frac{0.091}{\sqrt{35\,000}} = 4.9 \times 10^{-4},$$

$$\frac{\sigma_{H_0}}{H_0} \overset{(2)}{=} \ln 10 \; \sigma_{\log H_0}, \qquad \sigma_{H_0} = 74.6 \times 2.303 \times 4.9 \times 10^{-4} = 0.08 \text{ km/s/Mpc}. \tag{11.119}$$

(1) The error of a mean of N independent values is the scatter over $\sqrt{N}$.
(2) $\log_{10} H = \ln H / \ln 10$, so $\mathrm{d} \log_{10} H = \mathrm{d}H / (H \ln 10)$.

A scatter of 0.091 in $\log_{10} H$ is $\ln 10 \times 0.091 = 21\%$ per group, yet the mean is fixed to about a thousandth. Cosmicflows quotes ±0.2, more than twice the naive figure; groups that fall together in one flow are not independent (Section 11.15.4), so the naive figure is a floor. Either way the number is small beside the ±0.8 of the zero-point calibration and the possible systematic of about 3 km/s/Mpc that the authors themselves state.

The calibration route decides the answer. The value 74.6 comes from merging all eight methods onto one scale and tying that scale to 121 calibrator galaxies. Tying only the supernovae, through the 44 supernova hosts with Cepheid, red-giant or maser distances, gives 72.1, close to the SH0ES value [253, Sec. 14]. The two routes use largely the same data, and they differ by 2.5 km/s/Mpc, three times the statistical error. Hubble's lesson of Section 11.14.5 survives ninety years and a factor of two thousand in the number of galaxies: the fit delivers the expansion rate in units of a calibration, and the error of the calibration does not average down.

Peculiar velocities from the catalogue. The velocities in the catalogue use the difference $cz - H_0 d$ nearby and change over, by 3000 km/s, to the logarithmic estimator of Watkins and Feldman [252], the one we derived in Section 11.18.4 to keep the noise Gaussian [253, eqs. (9)–(11)]. For groups whose modulus is known to 0.10 mag or better, the velocities have a roughly Gaussian distribution with standard deviation 300 km/s. Linear theory with the Planck spectrum gave $\sigma_{v,1\mathrm{D}} = 308$ km/s in Section 11.15.3. The two numbers are not a like-for-like comparison: a modulus error of 0.10 mag is a distance error of $(\ln 10/5) \times 0.10 = 4.6\%$, which at 4000 km/s is 184 km/s of velocity noise, so the measured width contains both the motions and the errors. What the agreement in size does say is that the galaxies move at the speeds linear theory predicts. Cosmicflows also finds larger Hubble parameters in one part of the sky and smaller ones in the opposite part, the dipole pattern of a bulk motion of the sample that we met in Section 11.16.4.

11.21.2 Velocities predicted from a map of galaxies

A distance catalogue measures velocities and asks what mass moves them. The reverse route needs no distances: take a redshift survey, which maps where the galaxies are, and predict the velocities they must have. Carrick et al. did this with the 2M++ compilation, a redshift survey that maps the galaxy density out to 200 h^{-1}Mpc [258].

The prediction. The linear velocity field of Section 11.15.2 is $\boldsymbol{u}(\boldsymbol{x}) = \frac{aHf}{4\pi} \int \mathrm{d}^3x'\, \delta(\boldsymbol{x}')\, (\boldsymbol{x}' - \boldsymbol{x})/|\boldsymbol{x}' - \boldsymbol{x}|^3$, equation (11.95), where δ is the matter density contrast and f the growth rate. A survey sees galaxies, not matter. With the linear bias of Section 11.5.2, $\delta_g = b\,\delta$, the galaxy contrast is b times the matter contrast. Today $a = 1$ and $H = H_0$, so

$$\boldsymbol{u}(\boldsymbol{x}) \overset{(1)}{=} \frac{H_0 f}{4\pi} \int \mathrm{d}^3x'\, \frac{\delta_g(\boldsymbol{x}')}{b}\, \frac{\boldsymbol{x}' - \boldsymbol{x}}{|\boldsymbol{x}' - \boldsymbol{x}|^3} \overset{(2)}{=} \beta\, \frac{H_0}{4\pi} \int \mathrm{d}^3x'\, \delta_g(\boldsymbol{x}')\, \frac{\boldsymbol{x}' - \boldsymbol{x}}{|\boldsymbol{x}' - \boldsymbol{x}|^3}, \qquad \beta \equiv \frac{f}{b}. \tag{11.120}$$

(1) (11.95) at $a = 1$, $H = H_0$, with $\delta = \delta_g/b$.
(2) A constant comes out of the integral; name the ratio f/b.

The map of galaxies predicts the whole velocity field up to one number, β. Call p_i the line-of-sight component of the integral at galaxy i, the velocity it would have if β were one.

The fit is Hubble's fit again. A measured peculiar velocity u_i with error σ_i (the distance error turned into velocity, plus a small-scale dispersion that linear theory does not describe) is

modelled as $u_i = \beta p_i$+noise. Which β makes the required noise most plausible?

$$\chi^2(\beta) \overset{(1)}{=} \sum_i \frac{(u_i - \beta p_i)^2}{\sigma_i^2}, \qquad \frac{\mathrm{d}\chi^2}{\mathrm{d}\beta} \overset{(2)}{=} -2\sum_i \frac{p_i\,(u_i - \beta p_i)}{\sigma_i^2} = 0,$$
$$\hat\beta \overset{(3)}{=} \frac{\sum_i p_i u_i/\sigma_i^2}{\sum_i p_i^2/\sigma_i^2}, \qquad \operatorname{Var}\hat\beta \overset{(4)}{=} \Big(\sum_i \frac{p_i^2}{\sigma_i^2}\Big)^{-1}. \tag{11.121}$$

(1) Gaussian noise: $-2\ln\mathcal{L}$ is the χ^2 up to a constant.
(2) Differentiate each square; the chain rule brings down $-p_i$.
(3) Solve the linear equation for β.
(4) The second derivative is $2\sum p_i^2/\sigma_i^2$; the variance of a maximum-likelihood estimate in a linear-Gaussian model is the inverse of half of it (Chapter 3).

This is Hubble's slope (11.91) with $r_i \to p_i$, and the chronometer estimator (11.110) with $E_i \to p_i$: a parameter that multiplies a known template is estimated by a weighted ratio. Carrick et al. add a uniform velocity $\boldsymbol{V}_{\mathrm{ext}}$, a **bulk flow** [a velocity shared by the whole mapped volume, caused by mass outside it], so that $u_i = \beta p_i + \boldsymbol{V}_{\mathrm{ext}}\cdot\hat{\boldsymbol{r}}_i$. The model is still linear, now in four parameters, and the estimate is still one matrix solve. Comparing with Tully–Fisher and supernova velocities they find $\beta^* = 0.431 \pm 0.021$ for their reference galaxies, which translates into $f\sigma_8 = 0.401 \pm 0.024$, the same combination that redshift-space distortions measure (Section 11.5.4), and a bulk flow of 159 ± 23 km/s, preferred at 5.1 standard deviations [258].

Example: does the map explain our motion? The 2M++ volume alone predicts a Local Group velocity of 540 km/s towards galactic $(l, b) = (268^\circ, 38^\circ)$; the external bulk flow is 159 km/s towards $(304^\circ, 6^\circ)$. The motion through the microwave background found in Section 11.16.4 is 620 km/s towards $(272^\circ, 30^\circ)$. Velocities add as vectors, so we need the angle between the two directions. A direction (l, b) is the unit vector $(\cos b\cos l, \cos b\sin l, \sin b)$, and

$$\cos\theta \overset{(1)}{=} \sin b_1\sin b_2 + \cos b_1\cos b_2\cos(l_1 - l_2) = 0.698, \qquad \theta = 46^\circ,$$
$$|\boldsymbol{v}_1 + \boldsymbol{v}_2|^2 \overset{(2)}{=} v_1^2 + v_2^2 + 2v_1v_2\cos\theta = 540^2 + 159^2 + 2\times 540\times 159\times 0.698,$$
$$|\boldsymbol{v}_1 + \boldsymbol{v}_2| = 661 \text{ km/s}. \tag{11.122}$$

(1) The dot product of the two unit vectors; the $\cos b_1\cos b_2$ terms combine through $\cos l_1\cos l_2 + \sin l_1\sin l_2 = \cos(l_1 - l_2)$.
(2) $|\boldsymbol{v}_1 + \boldsymbol{v}_2|^2 = (\boldsymbol{v}_1 + \boldsymbol{v}_2)\cdot(\boldsymbol{v}_1 + \boldsymbol{v}_2)$, expanded.

The sum points to $(278^\circ, 32^\circ)$, 5° from the dipole direction, and its size is within 7% of the measured 620 ± 15 km/s. Most of our motion is pulled by structure within $200\,h^{-1}$Mpc; the rest, a bulk flow of 159 km/s, comes from farther away.

11.21.3 Flow corrections in the supernova ladder

The third rung of the SH0ES ladder is 277 supernovae with $0.0233 < z < 0.15$ [24]. Each one's redshift contains a peculiar velocity, which the Hubble law reads as a wrong distance.

How far does an unknown velocity move one supernova? The distance modulus of a source at luminosity distance d is $\mu = 5\log_{10}(d/10\,\mathrm{pc})$; at low redshift the predicted distance is $d = cz/H_0$. By (11.94), the observed cz is $H_0 r + u_r$, the Hubble flow plus the line-of-sight

peculiar velocity u_r. Treating u_r as a small change of cz,

$$\mu \overset{(1)}{=} 5\log_{10}\frac{cz}{H_0 \cdot 1\,\mathrm{Mpc}} + 25, \qquad \delta\mu \overset{(2)}{=} \frac{5}{\ln 10}\frac{\delta(cz)}{cz} \overset{(3)}{=} \frac{5}{\ln 10}\frac{u_r}{cz},$$
$$\sigma_\mu \overset{(4)}{=} \frac{5}{\ln 10}\frac{\sigma_v}{cz} \approx 2.17\,\frac{\sigma_v}{cz}. \tag{11.123}$$

(1) d in Mpc: $5\log_{10}(d \times 10^6\,\mathrm{pc}/10\,\mathrm{pc}) = 5\log_{10} d + 25$.
(2) $\mathrm{d}\log_{10} x = \mathrm{d}x/(x \ln 10)$, the same identity as in (11.119).
(3) The change of cz is the peculiar velocity.
(4) A velocity scatter σ_v gives a scatter in μ larger by the same factor.

For $\sigma_v = 250\,\mathrm{km/s}$ this is 0.181 mag at $z = 0.01$, 0.078 mag at $z = 0.0233$ and 0.012 mag at $z = 0.15$ (Figure 11.51, left). After standardisation, nearby supernovae scatter about the Hubble line by 0.16–0.17 mag all told, velocities included [259]. At $z = 0.01$ a velocity of the typical size is as large as that whole scatter; at $z = 0.0233$ it is half of it in magnitudes, a quarter in variance. The lower cut keeps the supernovae whose distances the velocities barely disturb; the upper cut, $z = 0.15$, keeps the Hubble law nearly independent of the matter density. Pantheon+ puts the remaining velocity scatter into each supernova's error as a term $\sigma_{v_{\mathrm{pec}}}$, computed from 240 km/s and converted to magnitudes as in (11.123) [26, eq. (3), Sec. 3.1.3].

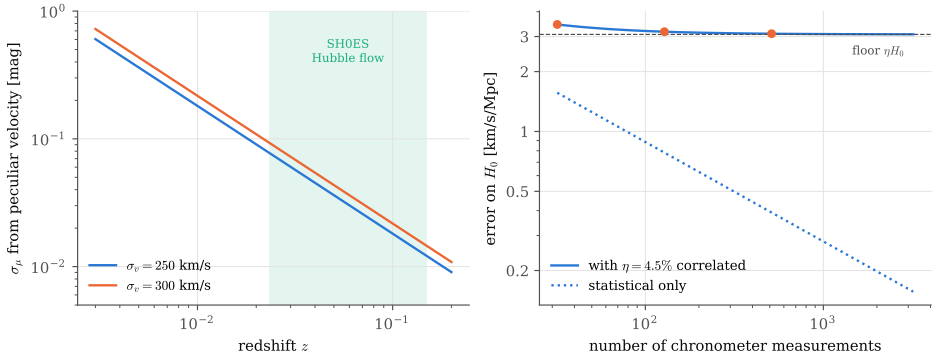


Figure 11.51: Left: the scatter in a supernova's distance modulus caused by an unmodelled peculiar velocity, for two velocity dispersions; the shaded band is the redshift range of the SH0ES Hubble-flow sample. Right: the error on H_0 from cosmic chronometers as the compilation grows, with only statistical errors (dotted) and with a 4.5% error shared by all points (solid), which levels off at ηH_0; the dots are the full generalised least-squares fit to 1, 4 and 16 stacked copies of the data.

Correcting the redshifts. Pantheon+ corrects the redshifts in two steps [259]. A supernova in a group gets the group's mean redshift, which removes the orbital motion inside the group. Then the velocity predicted from the 2M++ map, (11.120), is subtracted. Together the corrections reduce the scatter of 584 nearby supernovae about the Hubble line from 0.167 to 0.157 mag and raise H_0 by about 0.4 km/s/Mpc; the choice among velocity maps leaves a systematic uncertainty of only 0.06–0.11 km/s/Mpc in H_0. In the SH0ES fit, dropping the corrections lowers H_0 by 0.5, and dropping every supernova below $z = 0.06$ raises it by 0.3 [24, Secs. 6.8, 6.12]. As the SH0ES team notes, both changes go the wrong way for the idea that we sit in an underdense region that raises the local H_0 (Section 11.17.6): in such a hole, removing the nearest supernovae would lower H_0, and so would correcting for the outflow. What the corrections cannot remove is the part of the scatter in H_0 that comes from structure beyond the map, the cosmic variance of Section 11.17; for this redshift range it is estimated at 0.5–1% [105, Sec. IV.F].

Example: why not use the Cepheid hosts' own redshifts? The 37 galaxies with Cepheids have a mean cz of about 2000 km/s [24, Sec. 6.12]. If their redshifts were used directly, a velocity of 250 km/s would be 250/2000 = 12.5% of each one's Hubble velocity. For 37 independent velocities the mean would be uncertain by $12.5\%/\sqrt{37} = 2.1\%$. The velocities are not independent: galaxies within a few tens of megaparsecs share most of their motion (Section 11.15.4). The SH0ES team, using the spatial covariance of the flows, finds that this two-rung route is limited to 3–4%. If the velocities alone set that number, the 37 hosts are worth about $(12.5/3.5)^2 \approx 13$ independent ones. A third rung reaching $z > 0.023$ is what makes a 1% measurement possible.

11.21.4 Chronometers: what the error budget is made of

A cosmic chronometer (Section 11.19) reads $H(z) = -(1+z)^{-1}\mathrm{d}z/\mathrm{d}t$ from the age difference $\mathrm{d}t$ of passively evolving galaxies at nearby redshifts. The ages come from spectral features, such as the 4000 Å break, through slopes computed with **stellar population synthesis** models [models that predict the spectrum of a population of stars from its age, metallicity and history]. Moresco et al. split the covariance of the $H(z)$ points into a statistical part and four systematic ones [255, Sec. 3.1.4, eqs. (19)–(21)]:

$$C = C^{\mathrm{stat}} + C^{\mathrm{met}} + C^{\mathrm{young}} + C^{\mathrm{model}}, \qquad C^{\mathrm{model}} = C^{\mathrm{SFH}} + C^{\mathrm{IMF}} + C^{\mathrm{lib}} + C^{\mathrm{SPS}}. \tag{11.124}$$

Each term answers a question about the galaxies. *Is the metallicity right?* A 10% error in the stellar metallicity gives about a 10% error in H; it belongs to each spectrum separately, so it is diagonal. *Did the stars form in one burst?* Assuming so when the star formation lasted up to 0.3 Gyr costs 2–3%, again diagonal. *Is there a young component hiding in an old galaxy?* If 10% (1%) of the light came from young stars, H would be off by 5% (0.5%); in the samples used so far the contamination is below detection and the effect below 0.5%. *Is the population model right?* Changing the synthesis model shifts H by about 4.5%, the initial mass function by less than 0.4%, the stellar library by somewhat more [257]. These shifts move every point the same way, so they are treated as fully correlated across redshift. The three diagonal terms are already inside the published errors of Table 11.5. Errors that include a conservative systematic allowance do not show up as scatter, which is the reading we gave of the small χ^2 in Section 11.19.3. The correlated model term must be added by the user; we added it as a single $\eta = 4.5\%$.

More galaxies do not help. Section 11.19.2 showed that a fully correlated fractional error η adds in quadrature, $\sigma^2_{H_0} = \sigma^2_{\mathrm{stat}} + (\eta H_0)^2$, equation (11.111). Suppose we had k compilations of the same quality instead of one. The statistical variance falls as $1/k$; the shared model error is the same in every copy:

$$\sigma^2_{H_0}(k) \overset{(1)}{=} \frac{\sigma^2_{\mathrm{stat}}}{k} + (\eta H_0)^2 \xrightarrow{k\to\infty} (\eta H_0)^2. \tag{11.125}$$

(1) (11.111) with $\boldsymbol{E}^{\mathsf{T}} D^{-1} \boldsymbol{E}$ multiplied by k, since k independent copies add their Fisher information.

With $\sigma_{\mathrm{stat}} = 1.57$ and $\eta H_0 = 3.08$ km/s/Mpc, the error is 3.45 for the present data, 3.17 with four times as many and 3.10 with sixteen times as many; the generalised least-squares fit on stacked copies gives the same numbers (Figure 11.51, right). The floor is 4.5% of H_0. Moresco et al. reach the same conclusion by simulation: with the current data they find $H_0 = 66.5 \pm 5.4$ in flat ΛCDM, and their optimistic forecast reaches 69.0 ± 2.1 only if the synthesis-model term is resolved, leaving the initial-mass-function term [255, Sec. 3.1.6, Table 2]. In (11.125) a 0.4% floor is 0.27 km/s/Mpc, below the statistical error, so the data would again be limited

by their number. Their current interval is wider than ours (±4.3): their covariance contains the stellar-library term and a redshift-dependent synthesis-model term on top of our single constant. The centres, 66.5 and 68.1, differ by a third of either error.

A bias the differential method avoids. At high redshift, a selection for old passive galaxies may pick objects that are the ancestors of a different population from the one selected nearby, and so flatten the age–redshift relation. Comparing only galaxies at nearly the same redshift limits this **progenitor bias**; Moresco et al. estimate it at about 1% in H, small beside the model term.

11.21.5 What our shortcuts cost

Table 11.10 lists, for each calculation, what a research analysis does instead and what our reduction costs. The reductions are of two kinds. Some only cost precision and would shrink with computing time or data. Others leave out a physical ingredient, and no amount of computing inside our model brings it back.

section	research	ours	what it costs
Hubble 1929	5.6×10^4 galaxies, eight methods, zero points by MCMC, 542 calibrators	Hubble's 24 nebulae, one velocity scatter, solar motion fitted	nothing for the lesson; H_0 itself was off by a factor 6.6 in 1929
velocities	N-body and constrained simulations, nonlinear motions, galaxy bias	linear theory from CAMB; 4 Gaussian boxes of 1000 h^{-1}Mpc	no motions inside groups and clusters; matter only
local flow	velocity field from 2M++ with β and a bulk flow; Cosmicflows velocities	one spherical attractor, power-law profile, 400 simulated galaxies	no external bulk flow (the 159 km/s of [258]); one attractor only
local H_0	N-body box of 6 h^{-1}Gpc, observers in haloes and voids, the real supernova sky	linear $P(k)$, three windows, Gaussian boxes of 1500 h^{-1}Mpc	no observer selection, no nonlinear density; linear rows agree within 0.1 percentage point
Malmquist	joint likelihood of distances and a 3-D density map over the sky	one line of sight through one cluster, lognormal errors	no selection function, no sky coverage; the mechanism only
chronometers	full covariance (11.124), redshift-dependent terms, sampled posteriors	diagonal errors plus one constant 4.5%, grid profile over Ω_{m0}	interval ±4.3 instead of ±5.4 km/s/Mpc

Table 11.10: Each section's analysis at research scale and in the scripts, with the cost of the reduction.

The row that matters most is the local flow. Our Virgo model has one attractor; the real local Universe has the Great Attractor, Perseus–Pisces, the Shapley concentration and a bulk flow from beyond them all. Everything statistical we derived (the linear velocity field, the joint linear fit, the dilution by distance errors, the Malmquist correction) carries over unchanged to a three-dimensional map; what does not carry over is the claim that one parameter, $v_{\rm inf}$, describes the flow. Research replaces it with (11.120) evaluated on a real galaxy map, plus a bulk flow.

How this is used in real research: Checks, and instruments

Four calculations on the local Universe checked an analytic result by simulation: the velocity dispersion from $P(k)$ against Gaussian boxes (298±5 against 298 km/s), the window variance of a local H_0 against observers in boxes (1.48% against 1.45% at 100 h^{-1}Mpc), the Malmquist

factor against a simulated catalogue (1.096 against 1.095), and the chronometer errors against 4000 mock compilations (3.48 against 3.52). Each time the simulation verified a formula we already had. Research uses the same tools where no formula exists: Wojtak et al. needed an N-body simulation because observers sitting in haloes see a nonlinear density; Moresco et al. sized the synthesis-model error by simulating spectra with one model and analysing them with the others; Pantheon+ measures the effect of its velocity model by refitting with an alternative map, because no formula relates two maps.

The lesson. Every local measurement of the expansion is a weighted ratio of a measured quantity to a template, whose statistical error falls as $1/\sqrt{N}$, multiplied by a calibration that does not fall at all; the work of research is in the calibration and in the correlated errors, which is where Hubble's 500 km/s/Mpc went wrong and where the present tension will be decided.

Recap

- Hubble's 24 nebulae, refitted with a solar motion, give his $K = 465 \pm 56$. The slope of a fit through the origin is $\sum v_i r_i / \sum r_i^2$; distance errors add $K^2\Delta^2 r_i^2$ to the variance; a common factor in all distances is invisible to the fit, which is why 465 became 70 (Section 11.14).
- A redshift gives $u_r = c(z_{\rm obs} - z_c)/(1+z_c)$. Linear theory gives $\boldsymbol{u} = \frac{aHf}{4\pi} \int \delta\,(\boldsymbol{x}' - \boldsymbol{x})/|\boldsymbol{x}' - \boldsymbol{x}|^3$, a velocity dispersion $\sigma_{v,1\mathrm{D}} = 308\,\mathrm{km/s}$ from $P(k)/k^2$, and velocities correlated over $\sim 100\,h^{-1}$Mpc (Section 11.15).
- Infall towards a spherical overdensity is $-\frac{1}{3} aHf\,\bar{\delta}(<R)R$; a joint linear fit recovers H_0 and the infall speed, and ignoring the infall biases H_0. The Local Group moves at 620 km/s through the microwave background (Section 11.16).
- A local least-squares H_0 sees the density through the window $15 j_2(kR)/(kR)^2$, with $\delta H/H = -f\,\delta_W/3$; its scatter is about 1% at $150\,h^{-1}$Mpc, reproducing Wojtak et al. and Marra et al., and far too small to explain the 8.4% tension (Section 11.17).
- A distance measured with lognormal error Δ in a density $n(r)$ has $E(r \mid d) \approx d[1 + (7/2 + \gamma)\Delta^2]$: inhomogeneous Malmquist bias, which makes spurious outflows around clusters; the logarithmic velocity estimator keeps the noise Gaussian (Section 11.18).
- Chronometers give $H(z) = -(1+z)^{-1}\mathrm{d}z/\mathrm{d}t$; the generalised least-squares estimate is $\boldsymbol{E}^{\mathsf{T}} C^{-1} \boldsymbol{H} / \boldsymbol{E}^{\mathsf{T}} C^{-1} \boldsymbol{E}$, and a shared 4.5% model error sets a floor $\eta H_0 = 3.08\,\mathrm{km/s/Mpc}$; flat ΛCDM gives $H_0 = 68.1$ with $[63.9, 72.6]$ (Section 11.19).
- Research adds tens of thousands of distances on one zero point, velocity maps with $\beta = f/b$, flow-corrected redshifts beyond $z = 0.023$ where $\sigma_\mu = 2.17\,\sigma_v/cz$ is small, and full systematic covariances (Section 11.21).

Reading the literature

- **Statistics of the fits.** Cowan, *Statistical Data Analysis*, Chapters 6 and 7 on maximum likelihood and least squares [Cowan Chs. 6–7]; Hobson et al., *Bayesian Methods in Cosmology*, on survey design and forecasts [HobsonBMC §5.3].
- **Galaxies as points, and two-point estimators.** Peebles on the large-scale structure of the Universe [217]; Peebles and Hauser, Davis and Peebles, Hamilton, Landy and Szalay and Kerscher et al. on correlation-function estimators [218–222]; Feldman, Kaiser and Peacock on the power spectrum [223]; Xavier et al. on lognormal fields [55].

- **The acoustic peak and redshift-space distortions.** Eisenstein and Hu on the transfer function [109]; Seo and Eisenstein on forecasts [260]; Eisenstein et al. on the first detection [42]; Percival and White, Anderson et al. and Alam et al. on BOSS [224, 261, 262]; DESI 2024 [106]; Kaiser on redshift-space clustering [108].

- **Beyond two points.** Bernardeau et al. on perturbation theory [263]; Cooray and Sheth on the halo model [107]; Navarro, Frenk and White on halo profiles [94]; Takahashi et al. on simulated skies [51].

- **The local Universe.** Hubble, and Hubble and Humason, on the expansion [241, 242]; Baade and Sandage on the recalibration [243, 244]; Hogg on distance measures [104]; Peebles on the linear velocity field [245]; Górski on velocity correlations [246]; Schechter on Virgocentric infall [247]; Lynden-Bell et al. on the Great Attractor [248]; Strauss and Willick's review of peculiar velocities and Malmquist bias [251]; Watkins and Feldman on the logarithmic estimator [252]; Carrick et al. on velocities from 2M++ [258]; Tully et al., Cosmicflows-4 [253]; Planck 2018 for the dipole [186].

- H_0 **and the tension.** Riess et al. (SH0ES) and Brout et al. (Pantheon+) [24, 26]; Peterson et al. on flow corrections [259]; Marra et al. and Wojtak et al. on the cosmic variance of a local H_0 [249, 250]; Abdalla et al. for a review [105]; Planck 2018 for the early-Universe value [25].

- **Cosmic chronometers.** Jimenez and Loeb on the idea [254]; Moresco et al. on the 4000 Å break, on systematics, and the review [255–257].

Chapter summary

A galaxy survey is a list of points, and no theory predicts any one of them. What a model predicts is how the points cluster, so every number a survey yields comes from asking one question about a list: if the points are a random sample of a smooth field, what is the sampling distribution of a summary of their positions, and how does the survey distort it? Poisson sampling adds the shot noise $1/\bar{n}$; a random catalogue cancels the survey's edges in the Landy–Szalay estimator; a weight $1/(1+\bar{n}P)$ minimises the variance of the power spectrum; a dilation of a template measures the acoustic ruler; the motions of galaxies squash the pattern by $(b+f\mu^2)^2$ and so measure the growth rate. The two-point function does not see everything: fields with the same $P(k)$ can look entirely different, and haloes, their abundance, their profiles, and the voids between them carry the rest. In the local Universe the velocities are measured directly, and the same estimators return in their simplest form: Hubble's slope, the infall speed, β and the chronometer H_0 are all weighted ratios, each multiplied by a calibration the data cannot see. In the pipeline

$$\text{physical model} \rightarrow \boxed{\text{field or data} \rightarrow \text{summary } S \rightarrow P(S \mid \theta)} \rightarrow \text{infer physics},$$

galaxy clustering works the middle arrows for points rather than maps: how a sample of a field becomes a summary, and what sampling, masks and motions do to its distribution. Chapter 12 takes the structures the two-point function misses into topological summaries, Chapter 13 turns to the Planck data that set the early-Universe side of the H_0 comparison, and Chapter 14 collects the ways in which an estimator and its calibration part company.

Clustering of galaxies

concept	operational meaning	used later in
Cox process, shot noise	Poisson given $\bar{n}(1+\delta_g)$; counts-in-cells variance $\langle N\rangle + \langle N\rangle^2\sigma_R^2$; power $P(k) + 1/\bar{n}$	Chapters 12 and 14
lognormal field	$1+\delta = e^{G-\sigma_G^2/2}$: a positive density with a chosen spectrum, for mocks	Chapter 12
Landy–Szalay	$(DD - 2DR + RR)/RR$ against a random catalogue; edge terms cancel to first order	Chapter 14
integral constraint	$\hat{\bar{\xi}}$ is low by the mean of ξ over the survey	Chapter 14
FKP estimator	weighted data-minus-randoms field, $w = 1/(1+\bar{n}P)$; information counted by V_{eff}	Chapter 14
BAO dilation α	template fitted at αr; α measures D_V/r_d; significance $\sqrt{\Delta\chi^2}$ peak against no peak	Chapters 13 and 14
Kaiser formula	$P_s(k,\mu) = (b + f\mu^2)^2 P(k)$; multipoles P_0, P_2, P_4	Chapter 14
$f\sigma_8$	the growth rate times the amplitude, measured by the quadrupole-to-monopole ratio	Chapters 13 and 14
fingers of God	small-scale velocities stretch clusters along the line of sight; a damping of P_s	Chapter 14
mocks	a thousand simulated surveys for the covariance; error bars only as good as the mocks	Chapter 14

Beyond two points: haloes and voids

concept	operational meaning	used later in
same $P(k)$, different field	two fields with identical spectra and different phases; the difference is in higher-order statistics	Chapter 12
halo mass function	number of haloes per unit mass and volume, from the fraction of the smoothed field that has collapsed	Chapter 14
halo bias	how much more strongly haloes of a given mass cluster than matter	Chapter 14
halo profile	the density around a halo centre and its Fourier transform $u(k\,\vert\,m)$	Chapter 14
halo model	$P(k)$ as a one-halo (pairs in one halo) plus a two-halo (pairs in two haloes) term	Chapter 14
halo occupation	the mean number of galaxies in a halo of mass m, central and satellites	Chapter 14
voids	large underdense regions; their abundance, profiles and shapes as summaries	Chapters 12 and 14

The local Universe

concept	operational meaning	used later in
fit through the origin	$\hat{K} = \sum v_i r_i / \sum r_i^2$; distance errors add $K^2\Delta^2 r_i^2$ to the variance	Chapter 14
invisible calibration	a common factor in all distances, all magnitudes or all zero points leaves χ^2 unchanged	Chapters 13 and 14
peculiar velocity	$u_r = c(z_{\rm obs} - z_c)/(1+z_c)$; linear theory $\boldsymbol{u} \propto f \int \delta\,(\boldsymbol{x}' - \boldsymbol{x})/\lvert\boldsymbol{x}' - \boldsymbol{x}\rvert^3$	Chapter 14
velocity statistics	$\sigma_v^2 = (aHf)^2 \int P\,\mathrm{d}k/2\pi^2$; correlations $\psi_\parallel, \psi_\perp$	Chapter 14
infall and dipole	joint linear fit of H_0 and $v_{\rm inf}$; the Local Group moves at 620 km/s	Chapter 13
local H_0 variance	$\delta H/H = -f\delta_W/3$, window $15 j_2(kR)/(kR)^2$; ~1% at 150 h^{-1}Mpc	Chapters 13 and 14
Malmquist bias	$E(r \mid d) \approx d[1 + (7/2 + \gamma)\Delta^2]$; use logarithmic velocities	Chapter 14
$\beta = f/b$	velocities predicted from a galaxy map up to one factor, fitted as a weighted ratio	Chapter 14
flow corrections	$\sigma_\mu = 2.17\,\sigma_v/cz$; group redshifts and a velocity map; $z > 0.023$	Chapter 13
cosmic chronometers	$H = -(1+z)^{-1}\mathrm{d}z/\mathrm{d}t$; $\hat{H}_0 = \boldsymbol{E}^{\mathsf{T}}C^{-1}\boldsymbol{H}/\boldsymbol{E}^{\mathsf{T}}C^{-1}\boldsymbol{E}$; floor ηH_0	Chapters 13 and 14
borrowed calibration	every route is $H_0 = O/X$, an observable over a quantity from outside (M, a sound horizon r_d, an age scale, a distance scale); $\delta H_0/H_0 = (\ln 10/5)\,\delta M$ or $-\delta r_d/r_d$; it does not average down	Section 11.20 and Chapters 13 and 14
significance of a gap	$\lvert H_A - H_B\rvert/\sqrt{\sigma_A^2 + \sigma_B^2}$ for independent routes: ladder against Planck 4.8, the same gap in 2013 errors 2.4	Section 11.20 and Chapter 13

Part IV

Beyond Two-Point Statistics, and Real Data

Chapter 12

Topological Data Analysis: The Shape of a Field as Data

Where this chapter is going

A theory of the early universe does not predict the temperature map we observe. It predicts the statistics of maps like it, and the sky gives us one of them, seen through a beam, with noise, behind a mask, after the light has been bent by every galaxy cluster it passed. Until now we compressed such a map into its power spectrum (Chapters 7 and 8) and fed that summary to likelihoods, Markov chains and Fisher matrices (Chapters 9 and 10). The power spectrum keeps the amount of fluctuation on each scale and throws away everything else: in Section 7.7 two fields with identical spectra looked nothing alike, one made of blobs and the other of rings. The difference between blobs and rings suggests a measurement of shape. To use such a measurement on an observed field, we need to know which aspects of its geometry survive the effects of propagation and of the instrument. *Which features survive the changes we regard as nuisances, and how can we quantify them?*

The answer starts from a simple construction. Pick a threshold and colour every point where the field is above it. The coloured region, the excursion set, is a genuine geometric object: it has separate pieces, and the pieces can have holes. Counting pieces and holes gives the Betti numbers, and pieces minus holes gives the Euler characteristic. These counts do not change when the region is stretched or bent without tearing, which is exactly the kind of change a smooth remapping of the sky produces. On a pixel grid the counting needs a convention for which pixels touch, and the convention for the region and for its complement have to fit together. For a Gaussian field the average Euler characteristic at each threshold can be computed exactly by counting maxima, saddles and minima, and the answer depends on the power spectrum through a single number, the variance of the gradient. That has two consequences: for a Gaussian sky the Euler characteristic adds nothing to the power spectrum, and a measured curve that departs from the formula is a direct sign that the sky is not Gaussian. We check the formula by Monte Carlo on the plane, in three dimensions and on the sphere, and reproduce the classic curves of Hamilton, Gott and Weinberg, of Schmalzing and Górski, of Matsubara and of the *Planck* team (Figure 12.1).

A single threshold is an arbitrary choice. Part 12b follows every piece and every hole as the threshold sweeps through all values, which gives persistent homology: birth and death of each feature, persistence diagrams, and the distances between them. Part 12w runs the whole pipeline on weak-lensing mass maps, from shear to topology. Part 12c closes the loop: a topological summary is a random observable like any other, with a sampling distribution obtained by simulation, a covariance, a likelihood, a Fisher forecast and a posterior, and it can tell apart fields that the power spectrum cannot.

In the pipeline of the book (physical model → field or data set → choose what information matters → construct a summary → derive its sampling distribution → infer physics), this chapter lives on the third and fourth arrows. The earlier chapters answered *how* to do inference once a summary is chosen; this one asks *which* summary is worth choosing, and builds it.

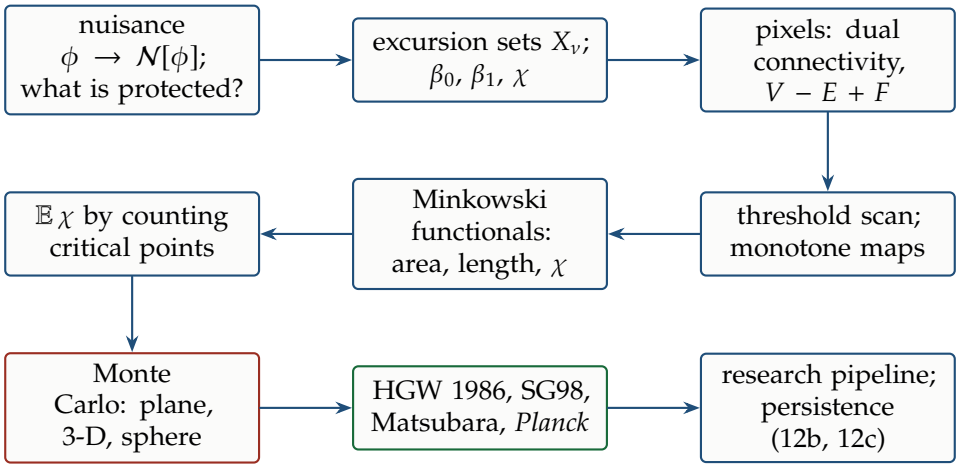


Figure 12.1: The route through the ideas. A nuisance transformation motivates looking for protected quantities; the excursion set turns a field into a shape whose pieces and holes can be counted; pixels force a counting convention; scanning the threshold and the three Minkowski functionals organise the counts; the expected Euler characteristic of a Gaussian field follows from counting critical points. The red box is the numerical test of that formula, the green box the reproduction of the published results.

12.1 What survives a nuisance transformation?

Take one temperature map of the microwave sky. Before it reached us, the light was deflected by the matter it passed, so every hot spot is displaced by a few arcminutes and slightly stretched. The telescope then blurred it with its beam and the detectors added noise. Each of these steps changes the numbers in the map. If we summarise the map by its power spectrum, each of them also changes the summary: lensing moves power between scales and turns some E-mode polarization into B modes (Chapter T1), the beam suppresses the high multipoles (Section 7.6.4), and the noise adds a bias (Section 8.3). In Chapter 8 we handled this the honest way: model each step and simulate it. There is a second strategy, and this chapter is about it. Instead of modelling a nuisance, look for a quantity it cannot change.

The question. Suppose a physical theory produces a field $\phi(\boldsymbol{x})$ and an experiment gives us one imperfect realisation of it. The values of the field are altered by processes we regard as nuisance: perturbations along the way, noise, smoothing, propagation. Is there information about the *shape and structure* of the data that these alterations leave intact?

12.1.1 A nuisance is an operator; a protected statistic ignores it

Write the effect of the nuisance as an operator acting on the whole field,

$$\phi \longrightarrow \phi' = \mathcal{N}[\phi]. \tag{12.1}$$

Here ϕ is the field as the physics made it, ϕ' the field we receive, and $\mathcal{N}$ the **nuisance operator**: lensing, a beam, a monotone change of units, added noise, or a combination. A summary

statistic T is a rule that turns a whole field into a number (or a short list of numbers), like the variance, the power at one wavenumber, or the count of hot spots. The ideal situation is a statistic that does not notice the nuisance,

$$T\big[\mathcal{N}[\phi]\big] = T[\phi] \qquad \text{for every field } \phi \text{ the theory can produce.} \tag{12.2}$$

We call such a T **protected** against $\mathcal{N}$. This protection means that we may measure T on the field we received and interpret it as if we had measured it on the field the physics made, without ever modelling the nuisance. That is a physical reason to care about T, not just one more way of compressing data.

The power spectrum is not protected against any of the nuisances above, as we just saw. What could be? Think about what lensing actually does. It does not change the temperature of a photon; it changes the direction from which the photon seems to arrive. To first order the observed map is the true map evaluated at a displaced point,

$$\phi'(\boldsymbol{x}) = \phi\big(\boldsymbol{x} + \boldsymbol{s}(\boldsymbol{x})\big), \tag{12.3}$$

with $\boldsymbol{s}$ a small, smooth displacement field (the lensing deflection; Chapter T1 and Section T3.4). Every value present in ϕ is still present in ϕ'; only its location has moved. A hot region is carried along, stretched or squeezed, but if the displacement is smooth and never folds the sky onto itself, the region is not torn into two and no two regions are glued together. So any statistic that counts *how regions are connected*, rather than how large or how long they are, has a chance to be protected.

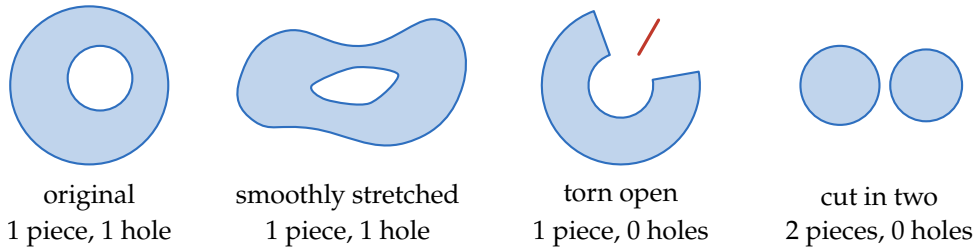


Figure 12.2: What a smooth one-to-one deformation can and cannot do. Stretching the ring (second panel) changes its area and the length of its boundary, but it is still one piece with one hole. Only an operation that tears (third panel, the red cut) or cuts the region (fourth panel) changes these counts.

Figure 12.2 makes the distinction concrete. The number of pieces and the number of holes of a region survive any deformation that neither tears nor glues. Mathematicians call such counts **topological invariants**; the precise statement, and why the counts are well defined, is in the topology notes. For us the working content is short: *a count of pieces and holes can change only at an event where something is torn or glued*.

12.1.2 The protection is conditional

Equation (12.2) is an ideal, and the honest version of it always comes with a condition. A topological count is protected against a stated *class* of transformations, the smooth one-to-one ones, and against nothing else. Noise is not in that class: it adds new small hot and cold spots everywhere, which creates pieces and holes out of nothing. A beam is not in it either: it averages neighbouring values, so two nearby hot spots can merge into one. A remapping that is too strong can fold, sending two different points to the same place, and then regions can be glued. Even a perfect remapping of the full sky becomes imperfect on a masked sky, because a region can slide across the edge of the mask. Finding out exactly what is protected, and how well, is part of the physics problem, not a detail to be waved away.

Example: four nuisances on one map. The script `code/ch12/01_nuisance.py` makes one Gaussian random field on a periodic 256×256 grid with the recipe of Section 7.4 (white noise smoothed on 6 pixels, then scaled to zero mean and unit variance) and colours the pixels where the field exceeds one standard deviation. It then applies four nuisance operators and recounts. (i) A smooth remapping (12.3) whose displacement has an r.m.s. of 14 pixels and varies on a scale of about 45 pixels; the smallest value of the Jacobian $\det(\mathbb{1} + \nabla \boldsymbol{s})$ over the map is 0.33, positive, so the map never folds. (ii) A monotone change of the values, $\phi \to e^{\phi} + \phi^3$, with the threshold moved along with it (Section 12.3.2). (iii) White noise of r.m.s. 0.3 added to every pixel. (iv) Extra smoothing by a Gaussian of 6 pixels. The numbers are:

	area [%]	boundary length [pix]	pieces β_0	holes β_1	$\beta_0 - \beta_1$
original	15.2	2136	37	0	37
smooth remapping	12.5	1957	37	0	37
monotone values	15.2	2136	37	0	37
added noise	16.5	11241	493	450	43
extra smoothing	15.9	1648	17	0	17

The remapping changed the area by 2.8 percentage points and the boundary length by about 8 per cent, but not a single piece or hole was created or destroyed. The monotone change left everything identical, pixel for pixel (the script checks that the two coloured sets coincide: yes). The noise multiplied the number of pieces by more than 13, because it sprinkled tiny one- and two-pixel islands and lakes over the map (Figure 12.3). The extra smoothing merged and erased the smallest hot spots and lost 54% of the pieces. So the counts are protected against exactly the first two operators, and badly broken by the last two, as the argument above predicted.

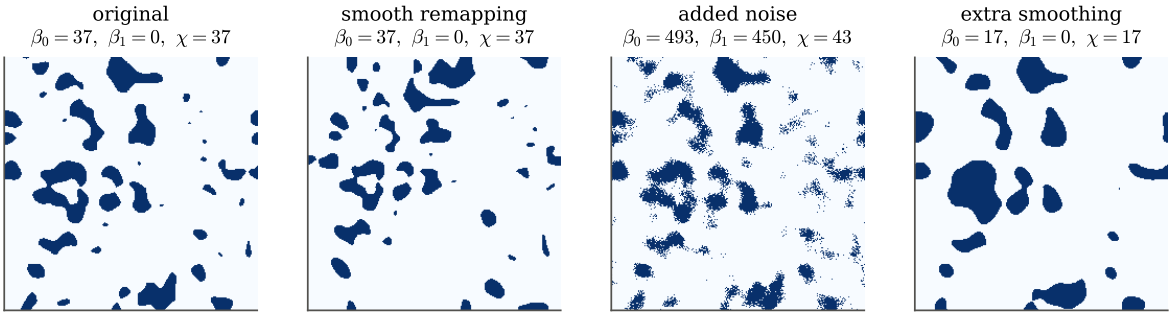


Figure 12.3: The pixels above one standard deviation (dark) for the original field and three of the transformed ones. The smoothly remapped set looks distorted but has the same pieces and holes. Added noise covers the map with tiny spurious pieces; extra smoothing merges or removes the small ones.

Example: lensing of the microwave sky. Section T3.4 carried out this experiment with real lensing deflections on flat patches of the simulated CMB sky. When the deflection was applied as a pure remapping, the Euler characteristic (the pieces minus the holes, defined properly in Section 12.2) changed only at the level of pixel ambiguities; when the beam was applied after the lensing, the change grew to a few tenths of a per cent. That is the same pattern as the table above, with the real nuisance in place of a toy one: the remapping is harmless, the smoothing that follows it is not.

Example: polarization singularities. The polarization of the microwave background has points where it vanishes and its direction is undefined, and around each the direction turns by a

definite half-integer number of turns, its index (Section T3.3). A smooth one-to-one remapping can move such a point and distort its surroundings, but it cannot change its index continuously, so the index is protected [1, 2]. It fails exactly where the conditions fail: when two singularities of opposite index annihilate, when one crosses the edge of the region, or when noise and finite resolution blur it [3]. The method of thought is the same in every example: *identify the nuisance first, then look for information it cannot easily erase*.

12.1.3 Two different questions: what survives, and what changes

The invariance argument is easy to misuse. Topology is unchanged by smooth one-to-one deformations of a *region*. That does not mean the topology of a physical field is unchanged by the *dynamics* that evolves it. Gravitational collapse is not a remapping of the initial density: matter flows together, peaks merge, filaments form and voids empty. The connectivity of the overdense regions at late times is genuinely different from that at early times, and simulations of structure formation track exactly this change with the counts of pieces, loops and voids and their persistence [4, 5]. There the change *is* the signal.

So topology answers two different questions, and an analysis must say which one it asks. For a nuisance that acts like a smooth remapping, the question is *what survives*, and the protected counts are the answer. For a physical process that creates, destroys or reconnects structures, the question is *what changes because of the physics*, and the same counts are the record of it. In both cases the counts are summaries of the data that the power spectrum does not expose: two fields can have identical spectra and very different connectivity, like the rings and their phase-randomised twin of Section 7.7. That is the question behind the last row of the taxonomy of summaries in Section 7.7.1: is there physically meaningful information in the way regions of the field are connected that the statistic we are already using does not expose?

12.2 The topology of a field: excursion sets and what we count in them

Consider a weather map that gives the temperature at every point in a country. To ask a question about topology, we first use those temperatures to define a region. For example, colour every point warmer than 25°C. The coloured set might consist of three disconnected warm regions, one of which surrounds a cooler lake. The threshold has turned a collection of temperatures into a geometric object whose connected components and holes we can count.

12.2.1 Excursion sets

Let the field be a map $\phi : \mathcal{M} \to \mathbb{R}$ from a domain $\mathcal{M}$ (a square patch, a periodic box, the sky) to the real numbers. Throughout, ϕ is measured in units of its own standard deviation σ, so a threshold $\nu = 1$ means one standard deviation above the mean.

Working definition 12.1 (excursion set, Betti numbers, Euler characteristic)

The **excursion set** at threshold ν is

$$X_\nu = \{\boldsymbol{x} \in \mathcal{M} : \ \phi(\boldsymbol{x}) > \nu\}, \tag{12.4}$$

the part of the domain where the field exceeds ν [on a CMB map, the directions hotter than the threshold]. In two dimensions $\beta_0(X_\nu)$, the zeroth **Betti number**, is the number of connected pieces of X_ν [pieces that cannot be joined by a path staying inside X_ν], and $\beta_1(X_\nu)$ is the number of its holes [bounded regions of the complement, each fully enclosed

by X_ν]. The **Euler characteristic** of a set in the plane is

$$\chi(X_\nu) = \beta_0(X_\nu) - \beta_1(X_\nu). \tag{12.5}$$

Operationally: colour the pixels above ν, count the coloured pieces, count the uncoloured regions that do not reach the edge, subtract.

Read (12.4) as a machine. The input is a field, which carries a number at every point; the output is a set, which carries a shape. Everything in this chapter measures that set. For a CMB map $T(\hat{\boldsymbol{n}})$ on the sphere, X_ν is literally the set of directions $\hat{\boldsymbol{n}}$ with $T(\hat{\boldsymbol{n}}) > \nu\sigma$, the hot spots at that threshold. "Studying the topology of a field" means studying the topology of its excursion sets, threshold by threshold.

Why pieces minus holes? Each Betti number on its own is a perfectly good count, and parts 12b and 12c use them separately. The combination (12.5) is singled out for two reasons. First, it can be computed by purely local bookkeeping, counting corners, edges and squares, with no need to find which pixels are connected to which (Section 12.2.3). Second, its average over a Gaussian random field can be calculated exactly (Section 12.5), which neither β_0 nor β_1 alone allows. In three dimensions a solid region can also enclose sealed cavities, counted by β_2, and the Euler characteristic becomes $\beta_0 - \beta_1 + \beta_2$ (Section T3.6.2).

Example: three shapes. A disc has one piece and no hole: $\chi = 1 - 0 = 1$. The ring of Figure 12.2 has one piece and one hole: $\chi = 1 - 1 = 0$. A slice of Swiss cheese with three holes has $\chi = 1 - 3 = -2$. Three separate discs have $\chi = 3$. So χ is positive when the set is made of isolated islands and negative when it is a connected sea full of lakes. A threshold scan will move between these two extremes (Section 12.3).

Why the counts survive deformations. A continuous deformation that neither tears nor glues sends paths to paths, so two points that could be joined inside the set can still be joined afterwards, and two points that could not still cannot; the number of pieces is unchanged. A hole is a piece of the complement fully surrounded by the set, and the same argument applied to the complement keeps the number of holes. In fact more is true: the counts are unchanged even by continuous shrinkings, such as squeezing a ring down to a thin circle, as long as nothing is cut (homotopy invariance; see the topology notes). That is why the counts are robust to the smooth remapping of Section 12.1.1, and why they are blind to how big or how elongated the pieces are.

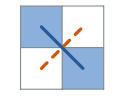
12.2.2 Pixels: which pixels touch?

A real map is a grid of pixels, and the excursion set is a set of coloured squares. To count pieces we must decide when two coloured squares belong to the same piece. Two squares that share an edge clearly touch. Two squares that meet only at a corner are the problem. Under the **4-neighbour** rule only the four squares sharing an edge are neighbours; under the **8-neighbour** rule the four diagonal squares are neighbours too.

Does the choice matter? Look at the smallest troublesome pattern, a 2×2 block coloured like a chessboard (Figure 12.4, left). Under the 8-neighbour rule the two dark squares form one piece, connected through their shared corner. If we also used the 8-neighbour rule for the light squares, they too would form one piece, through the very same corner. Two paths, one dark and one light, would cross at a point without either being interrupted. That cannot happen for regions in the plane: a closed curve always separates its inside from its outside (the Jordan curve theorem; see the topology notes). So if the dark set is 8-connected, its complement must

be 4-connected, and vice versa. With this **dual** convention every dark corner contact cuts the light squares apart there, and the counts of pieces and of holes are consistent with each other.

There is a natural way to see which pairing we are using. Take each coloured pixel to be a *closed* unit square, edges and corners included. Two diagonal squares then share a corner point, so the union is 8-connected. The uncoloured part of the plane is what is left, the union of *open* squares minus every point that belongs to a coloured square; two uncoloured diagonal squares are separated by the coloured corner between them, so the complement is 4-connected. Throughout this chapter the excursion set is the union of closed squares of the pixels above threshold: 8-neighbours for the set, 4-neighbours for the complement. The opposite choice, 4 for the set and 8 for the complement, is equally consistent; the two choices give different counts only where the field crosses the threshold at a saddle inside a single 2×2 block, which becomes rare when the field is smooth on the scale of a pixel (the last example of Section 12.2.3 measures how rare).



8 for both:
paths cross

2	2	2	0	0
2	0	2	0	1
2	2	3	0	0
0	0	0	1	0
1	0	0	0	0

$\nu = 0.5$

2	2	2	0	0
2	0	2	0	1
2	2	3	0	0
0	0	0	1	0
1	0	0	0	0

$\nu = 1.5$

2	2	2	0	0
2	0	2	0	1
2	2	3	0	0
0	0	0	1	0
1	0	0	0	0

$\nu = 2.5$

Figure 12.4: Left: the chessboard pattern. If dark squares touching at a corner were connected and light ones too, the dark path (solid) and the light path (dashed) would cross without either being broken; so the complement of an 8-connected set must be 4-connected. Right: the 5×5 hand example. Each pixel shows its field value; the pixels above the threshold are shaded.

Example: a 5×5 field by hand. Take the field values of Figure 12.4 (right), with row 1 at the top. At $\nu = 0.5$ the coloured pixels are the seven 2s and the 3 that together surround the central 0, and three 1s. Count the pieces with 8-neighbours. The ring of 2s and the 3 is one piece. The 1 in row 4, column 4 touches the 3 at a corner, so it joins the ring. The 1 at the right end of row 2 has no coloured neighbour at all, nor does the 1 at the bottom left. So $\beta_0 = 3$. Count the holes with 4-neighbours for the uncoloured pixels: the central 0 of the ring is surrounded on all four sides and cannot reach the edge, so it is a hole; every other uncoloured pixel reaches the edge. So $\beta_1 = 1$ and $\chi = 3 - 1 = 2$. With the opposite convention, the 1 that touched the 3 only at a corner is a separate piece, $\beta_0 = 4$, while the hole remains, so $\chi = 3$: the one diagonal contact is where the two conventions disagree. At $\nu = 1.5$ only the ring of 2s and the 3 survive: one piece with one hole, $\chi = 0$ under both conventions. At $\nu = 2.5$ only the single 3 survives, $\chi = 1$. The script `code/ch12/02_grid_counting.py` repeats these counts in code and finds 3, 1, 2 (and 3 with the opposite convention) at $\nu = 0.5$, 0 at 1.5 and 1 at 2.5.

Counting pieces and holes in code. Counting the pieces of a binary image is a standard task, connected-component labelling, and `scipy.ndimage.label` does it for either neighbour rule. For the holes we use the definition directly: a hole is a piece of the complement that does not reach the outside. Surround the image with a one-pixel frame of uncoloured pixels, so that everything outside the image becomes one single uncoloured piece; then label the uncoloured pixels with the dual rule and subtract that one outside piece:

Listing 12.1: Betti numbers of a pixel set in the plane (`code/ch12/lib_topo.py`)

```
def betti_plane(mask, conn: int = 8):
    """(beta0, beta1) of a pixel set in the plane.
```

```
    beta0: components of the set with 'conn'-connectivity.
    beta1: holes = components of the complement that cannot reach the outside, counted
           with the DUAL connectivity (4 if conn == 8, else 8).  Padding the image with a
           frame of background makes everything outside the grid one single component, so
           beta1 = (components of the padded complement) - 1.
    """
    mask = np.asarray(mask, dtype=bool)
    b0 = n_components(mask, conn)
    comp = ~np.pad(mask, 1, constant_values=False)
    b1 = n_components(comp, 4 if conn == 8 else 8) - 1
    return b0, b1
```

That a hole of a region in the plane is the same thing as a bounded piece of its complement is a special property of the plane (and of the sphere), known as Alexander duality (see the topology notes). In three dimensions holes are tunnels, and they cannot be found this way.

12.2.3 The Euler characteristic as vertices minus edges plus faces

There is a second way to compute χ that never asks which pixels are connected. Think of the union of closed squares as built from three kinds of cells: the lattice points at the corners of coloured pixels (vertices, V of them), the unit segments on the sides of coloured pixels (edges, E), and the coloured pixels themselves (faces, F). Each vertex and each edge is counted once, however many coloured pixels share it. Then

$$\chi = V - E + F. \tag{12.6}$$

This is the formula Euler found for polyhedra, and it holds for any set built of cells (see the topology notes for the general statement). We can see why it agrees with $\beta_0 - \beta_1$ for pixel sets by building the set one pixel at a time and checking that both sides change by the same amount at every step.

One square on its own has $V = 4$, $E = 4$, $F = 1$: $\chi = 4 - 4 + 1 = 1$, one piece and no hole.

Adding one more square. Suppose the set is already built and we add one coloured square. The new square brings one face. Some of its four vertices and four edges may already be present, namely those where it touches the existing set. Call the part of its boundary where it touches the set the *contact*. The contact is made of c separate arcs, each a path along the square's boundary made of k_i edges and $k_i + 1$ vertices (an arc can be a single corner, $k_i = 0$). Then

$$\begin{aligned} \Delta\chi &\overset{(1)}{=} \Delta V - \Delta E + 1 \\ &\overset{(2)}{=} \Big[4 - \sum_{i=1}^{c} (k_i + 1)\Big] - \Big[4 - \sum_{i=1}^{c} k_i\Big] + 1 \\ &\overset{(3)}{=} 1 - c. \end{aligned} \tag{12.7}$$

(1) One new face; ΔV and ΔE are the vertices and edges of the new square that were not there before.

(2) The square has 4 vertices and 4 edges; the arcs of the contact already contain $k_i + 1$ vertices and k_i edges each.

(3) The 4s cancel and each arc contributes $-(k_i + 1) + k_i = -1$.

Now check $\beta_0 - \beta_1$ in each case. If $c = 0$ the square is isolated: a new piece, $\Delta\beta_0 = +1$, and (12.7) gives $+1$. If $c = 1$ the square is glued to one piece along one arc, which changes neither the number of pieces nor the number of holes, and (12.7) gives 0. If $c = 2$ the square touches the set

in two separate places. Either those two places belonged to different pieces, which the square now joins ($\Delta\beta_0 = -1$), or to the same piece, in which case the square closes a loop around some uncoloured region and creates a hole ($\Delta\beta_1 = +1$). Either way $\Delta(\beta_0 - \beta_1) = -1 = 1 - c$. Each further arc merges one more piece or closes one more hole, and costs one more unit. The last case is a square whose whole boundary is already present, a one-pixel hole being filled: then the contact is a closed loop with 4 vertices and 4 edges, $\Delta V = \Delta E = 0$, $\Delta\chi = +1$, and indeed one hole disappears, $\Delta\beta_1 = -1$. Since both sides start at 0 for the empty set and change identically at every step, $V - E + F = \beta_0 - \beta_1$ for every pixel set.

Example: the ring by cells. The 3×3 block of the hand example without its centre has the $4 \times 4 = 16$ lattice points as vertices, 12 horizontal and 12 vertical unit edges (every side of the central hole is also a side of a coloured neighbour), and 8 faces: $\chi = 16 - 24 + 8 = 0$, one piece and one hole, as counted by hand.

The cell count is a sum of local terms, so it vectorises in a few lines of `numpy`. A lattice vertex is present if any of the four pixels around it is coloured, an edge if either of the two pixels beside it is:

Listing 12.2: Euler characteristic as $V - E + F$ of the pixel complex (`code/ch12/lib_topo.py`)

```
def euler_cubical(mask, conn: int = 8, periodic: bool = False) -> int:
    """chi = V - E + F of the cubical complex of a binary image.

    conn = 8: union of closed pixel squares (pixels are faces).  A lattice vertex or edge
              belongs to the set if ANY pixel touching it does.
    conn = 4: the dual construction (pixels are vertices).  An edge joins two 4-neighbour
              pixels that are both in the set, a face fills a 2x2 block that is entirely in.
    periodic: the grid is a torus (opposite sides glued), as for an FFT-generated field.
    """
    m = np.asarray(mask, dtype=bool)
    if conn == 8:
        if periodic:
            r0 = np.roll(m, 1, 0)
            r1 = np.roll(m, 1, 1)
            V = (m | r0 | r1 | np.roll(r0, 1, 1)).sum()
            E = (m | r0).sum() + (m | r1).sum()
        else:
            P = np.pad(m, 1)
            V = (P[:-1, :-1] | P[1:, :-1] | P[:-1, 1:] | P[1:, 1:]).sum()
            E = (P[:-1, 1:-1] | P[1:, 1:-1]).sum() + (P[1:-1, :-1] | P[1:-1, 1:]).sum()
        F = m.sum()
    else:
        if periodic:
            r0 = np.roll(m, -1, 0)
            r1 = np.roll(m, -1, 1)
            E = (m & r0).sum() + (m & r1).sum()
            F = (m & r0 & r1 & np.roll(r0, -1, 1)).sum()
        else:
            E = (m[:-1, :] & m[1:, :]).sum() + (m[:, :-1] & m[:, 1:]).sum()
            F = (m[:-1, :-1] & m[1:, :-1] & m[:-1, 1:] & m[1:, 1:]).sum()
        V = m.sum()
    return int(V) - int(E) + int(F)
```

The other branch builds the dual complex (4-neighbour convention), in which pixels are vertices, edges join two coloured 4-neighbours and a face fills a fully coloured 2×2 block. The `periodic` option treats the grid as a torus, which is how a field made with the fast Fourier transform really lives (Section 7.4.1). On a torus a set could in principle wrap all the way around and have holes that are not bounded lakes; at the thresholds and smoothing scales used here this never happens, and $\beta_1 = \beta_0 - \chi$.

Example: three methods on many fields. The script `code/ch12/02_grid_counting.py` draws 200 Gaussian fields on a 128×128 grid, smoothed on 4 pixels, and at each of 9 thresholds between -2 and 2 computes χ four ways: by labelling ($\beta_0 - \beta_1$), by cells ($V - E + F$), by `skimage.measure.euler_number` with 8-connectivity, and by the Betti numbers that the cubical-complex code of `gudhi` assigns to the same union of closed squares. In all 1800 cases the four answers agree exactly: 0, 0 and 0 mismatches. The cell count takes 0.1 ms per image, labelling 0.29 ms; `gudhi` takes 17.51 ms per field, but for that price it computes the counts at *every* threshold at once, which is what persistence will need in part 12b. Figure 12.5 shows what is being counted on one small map.

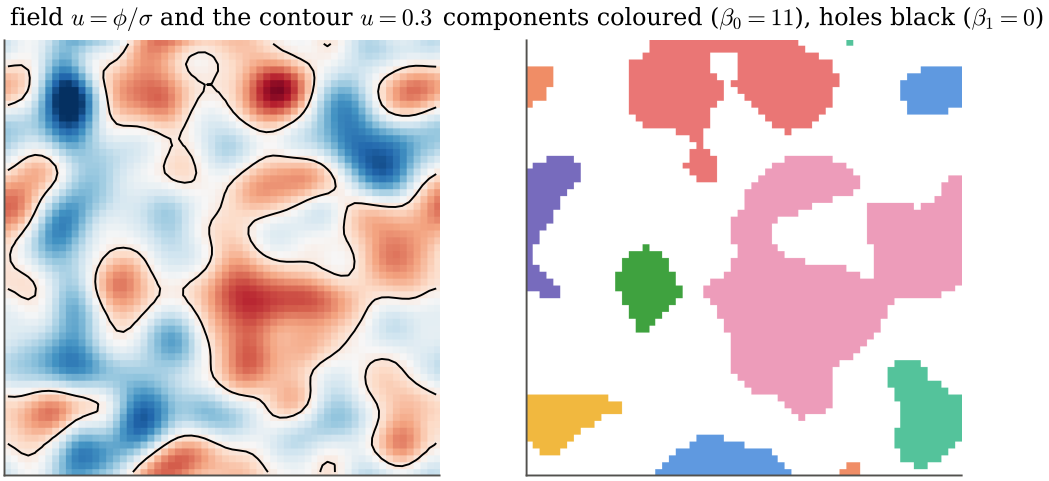


Figure 12.5: One 64×64 Gaussian field (left) with its contour at $\nu = 0.3$, and the excursion set above that contour (right). Each 8-connected piece has its own colour ($\beta_0 = 11$); an uncoloured region that cannot reach the edge is drawn black; this field has $\beta_1 = 0$ of them, so $\chi = 11$.

Example: when does the convention matter? The same script compares the two conventions on fields smoothed on $R = 1$, 2, 4 and 8 pixels. The average difference $|\chi_8 - \chi_4|$, as a fraction of the average $|\chi_8|$, is 15.7%, 4.1%, 1% and 0.1%. A field that changes from pixel to pixel produces many diagonal contacts and the convention matters a great deal; a field smoothed over several pixels produces few, and the two conventions converge. The practical rule: smooth the map on a scale of a few pixels or more before counting, and always state the convention.

Watch out: A count without a convention is not a measurement

The same pixels give different Betti numbers under the 4- and the 8-neighbour rules, and an analysis that uses 8-neighbours for both the set and its complement (or 4 for both) can report a hole count that is inconsistent with its piece count. Fix the pairing (8 for the set, 4 for the complement, or the reverse), use the same one for data and simulations, and smooth on several pixels so that the residual dependence is small.

12.3 Scanning the threshold

Imagine flooding a hilly landscape. With the water level very low, only the deepest valleys are lakes and the dry land is one continent full of holes. As the water rises the lakes grow and join, the continent breaks into islands, and finally only a few mountain tops stick out. The dry land is the excursion set X_ν of the height field, and the water level is the threshold ν, lowered or

raised. A single threshold catches one moment of this story; scanning the threshold tells the whole of it.

12.3.1 One map, all thresholds

The script `code/ch12/03_threshold_scan.py` makes one Gaussian field on a 256×256 grid, smoothed on 4 pixels (the recipe of Section 7.4), scales it to zero mean and unit variance, and computes β_0, β_1 and χ of the excursion set at thresholds from -2.5 to 2.5 in steps of 0.25, with the functions of Section 12.2.2. Three rows of its output:

threshold ν	pieces β_0	holes β_1	$\chi = \beta_0 - \beta_1$
-1	5	65	-60
0	26	21	5
1	77	0	77

At $\nu = -1$ about 84% of the map is above threshold (a standard normal value exceeds -1 with probability 0.84). It forms one large connected sea with a few small islands beside it, and the cold spots below -1 are its holes: χ is large and negative. At $\nu = 0$ half the map is coloured and the set is a tangle of large pieces and holes that nearly cancel. At $\nu = 1$ only 16% is coloured, in separate islands around the hot spots, almost none of them with a hole: χ is large and positive. The full scan is Figure 12.6. The number of pieces rises to a maximum of 78 near $\nu \approx 1$ and then falls, because at still higher thresholds the islands shrink and disappear one by one. The number of holes does the mirror image at negative thresholds.

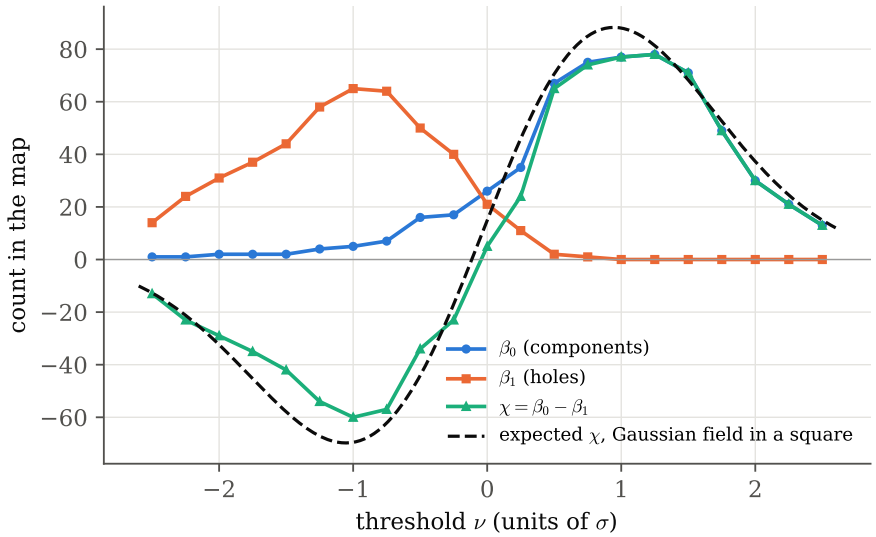


Figure 12.6: Pieces, holes and their difference for one 256×256 Gaussian map as the threshold is scanned. As the threshold rises, the excursion set changes from a connected region with many holes to a collection of isolated peaks. This change in which count dominates takes χ from negative to positive. The dashed curve is the expected χ for a Gaussian field in a square, derived in Sections 12.5 and 12.5.5.

Example: the mirror symmetry. The holes at $\nu = -1$ (65 of them) and the pieces at $\nu = +1$ (77) are similar numbers. That is not a coincidence. A hole of X_{-1} is a bounded region where $\phi < -1$, which is a piece of the excursion set of the field $-\phi$ above $+1$. A Gaussian field with zero mean is exactly as probable as its negative, so the cold spots of ϕ below -1 are statistically identical

to the hot spots of ϕ above +1. On average, then, $\beta_1(-\nu)$ equals $\beta_0(+\nu)$ and $\chi(-\nu) = -\chi(\nu)$ (up to small corrections from the edges of the map, which Section 12.5.5 computes). On one map the two counts differ by the realisation's own fluctuations. A field whose hot spots and cold spots differ, for example one with a skewed distribution of values, breaks this symmetry, and the asymmetry of the χ curve is one of the simplest tests of Gaussianity.

These numbers belong to one random map. They are not yet a measurement of anything: another map from the same theory gives other numbers. They become a statistic only when we know how they scatter from map to map, which is what the Monte Carlo of Section 12.6 and the whole of part 12c are about. The scan raises a second question too: if one threshold is an arbitrary choice, why not keep track of every piece and every hole as the threshold moves, recording at which level each appears and at which it disappears? That bookkeeping is persistent homology, the subject of part 12b.

12.3.2 A change of the values that changes nothing

Suppose the field we observe is not ϕ but a pointwise, increasing function of it,

$$\phi'(\boldsymbol{x}) = f\big(\phi(\boldsymbol{x})\big), \qquad f'(\phi) > 0. \tag{12.8}$$

A change of units is such a function; so is the map from the initial density to a lognormal model of the evolved density, $\phi' = e^{a\phi}$; so is the response of a detector that saturates smoothly. The function changes every value but moves no point. *What does it do to the excursion sets?* Move the threshold along with the values:

$$\boldsymbol{x} \in X'_{f(\nu)} \overset{(1)}{\iff} f\big(\phi(\boldsymbol{x})\big) > f(\nu) \overset{(2)}{\iff} \phi(\boldsymbol{x}) > \nu \overset{(3)}{\iff} \boldsymbol{x} \in X_\nu. \tag{12.9}$$

(1) The definition (12.4) of the excursion set of ϕ' at threshold $f(\nu)$.
(2) A strictly increasing function preserves order in both directions: $a > b$ if and only if $f(a) > f(b)$.
(3) The definition of the excursion set of ϕ at threshold ν.

So $X'_{f(\nu)}$ and X_ν are the same subset of the domain, point for point. Their pieces, holes, area and boundary are identical. A monotone change of the values is a nuisance against which *every* property of the excursion sets is protected, provided the threshold is relabelled.

But we do not know f. In practice nobody tells us the function; we only see ϕ'. If we threshold ϕ' in units of its own standard deviation, as we did for ϕ, the thresholds no longer correspond and the curves look different. There is one label that every monotone f respects: the *fraction of the area* above threshold. By (12.9) the set $X'_{f(\nu)}$ has the same area as X_ν. So instead of labelling a threshold by its value, label it by the area it encloses, and convert that area back to a Gaussian-equivalent threshold ν_A through

$$\Psi(\nu_A) = \text{fraction of the map above the threshold}, \qquad \Psi(\nu) \equiv \int_\nu^\infty \frac{e^{-t^2/2}}{\sqrt{2\pi}}\,\mathrm{d}t, \tag{12.10}$$

the probability that a standard normal variable exceeds ν. For a Gaussian field $\nu_A = \nu$ on average, because the expected area fraction above ν is exactly $\Psi(\nu)$ (Section 12.4). For any monotone function of a Gaussian field the excursion set at area-fraction threshold ν_A is the same set as the Gaussian one at ν_A. In code, the threshold with a given area fraction is a quantile of the map's values:

Listing 12.3: The threshold that encloses a given area fraction (`code/ch12/03_threshold_scan.py`)

```python
def thresh_area(x, nua):
    """Threshold t with area fraction P(x > t) equal to the Gaussian tail Psi(nua)."""
    frac = 0.5 * (1 - erf(np.asarray(nua) / np.sqrt(2)))
    return np.quantile(x, 1 - frac)
```

Example: a lognormal field. The same script takes 60 Gaussian maps ϕ and their lognormal versions $e^{0.8\phi}$, each standardised to zero mean and unit variance. Thresholded in units of their own standard deviation (Figure 12.7, left), the two χ curves disagree badly. At $\nu = -1$ the Gaussian maps give a mean χ of -80.4, the lognormal ones -9.6. The reason is simple once found: the lognormal field is positive and skewed, so its standardised values cannot go below about -1.06, and at $\nu = -1$ almost the whole map is above threshold, with very few holes. Relabelled by area fraction (right panel), the two curves are identical; the script checks that the two excursion sets coincide pixel for pixel at every threshold of every map: yes.

Example: a field that is not a monotone function of a Gaussian one. Now take two independent Gaussian fields ϕ and ψ with the same spectrum and form $\phi^2 + \psi^2$, a χ^2 field with two degrees of freedom, standardised. Its values are skewed like the lognormal's, but it is not a pointwise function of a single Gaussian field: its value at a point depends on two independent fields, and its lowest values sit only near the isolated points where both happen to vanish. Relabelling by area fraction cannot undo that. At $\nu_A = 1$ its mean χ is 139.4 against the Gaussian 80.4, and at $\nu_A = -1$ it is -196.2 against -80.4. The remaining difference is information about how the field is arranged in space, which no change of units can create or remove.

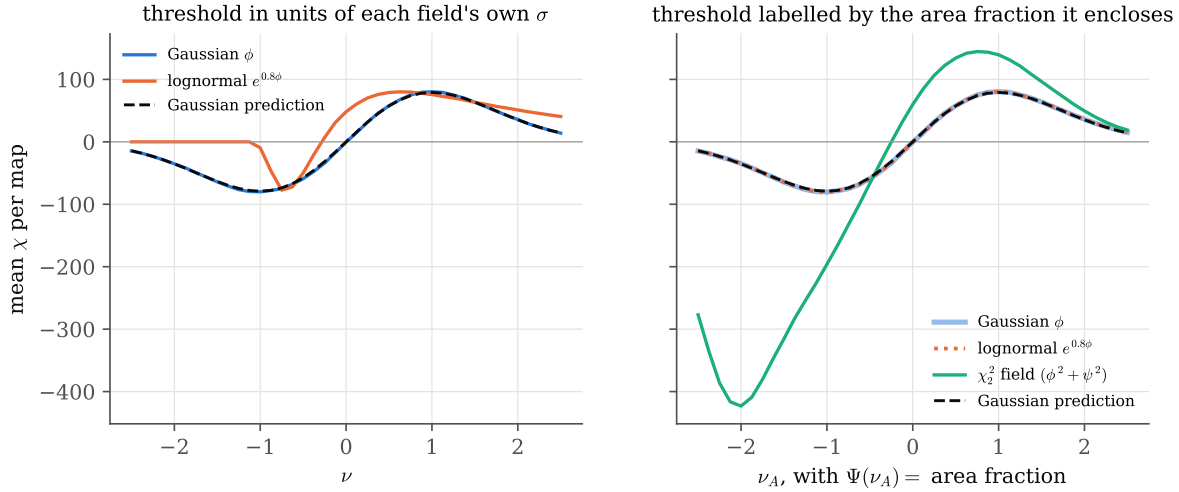


Figure 12.7: Mean χ per 256×256 map. Left: thresholds in units of each field's standard deviation; the lognormal field (orange) looks very different from the Gaussian one (blue). Right: thresholds labelled by the area fraction they enclose; the lognormal and Gaussian curves now lie exactly on top of each other, while the χ^2 field (green), which is not a monotone function of one Gaussian field, still differs.

Section T3.2.2 reached the same conclusion from the other side, for the slightly skewed skies of weakly non-Gaussian models: the part of the change in the Minkowski functionals that comes from a pointwise distortion of the values is only a sideways shift of the curves, and it disappears at fixed area fraction. What the two examples here add is the exact, non-perturbative version: for *any* increasing function, however strongly non-Gaussian the values, the curves at fixed area fraction are untouched. Two lessons follow. First, labelling by area fraction separates the information carried by the one-point distribution of values (which a histogram of the map

already gives) from the information about spatial arrangement, which is what topology is for. Second, the same argument marks the limit of the invariance discussed in Section 12.1.3: gravitational evolution is not a pointwise function of the initial density, because it moves matter, so its effect on the excursion sets does not disappear under relabelling. That is why the topology of the evolved density field carries information about the dynamics.

12.4 Area, boundary length and Euler characteristic

The microwave-sky chapter already measured the shapes of hot spots with three numbers (Section T3.1): the area fraction of the excursion set, a quarter of its boundary length per steradian, and its Euler characteristic per steradian, called v_0, v_1 and v_2. There we derived the expected area fraction and boundary length of a Gaussian sky, (T3.2) and (T3.6), and stated the expected Euler characteristic (T3.7) without proof, promising the derivation for this chapter. Two things are different now. The field lives on a flat square grid, so lengths are in pixels and there is no curvature of the domain to worry about until Section 12.5.6. And we need the Euler characteristic not as a number to be quoted but as a quantity we can compute, which means writing it in a form that the counting argument of Section 12.5 can attack.

12.4.1 Three numbers per threshold

Why exactly these three? Suppose we want a number $S(X)$ that describes the shape of a region X of the plane and has three properties. It is *additive*: for two regions, $S(X \cup Y) = S(X) + S(Y) - S(X \cap Y)$, so the shape number of a set of separate islands is the sum over the islands. It is *motion invariant*: moving or rotating the region does not change it, which is what we want on a statistically homogeneous and isotropic sky. And it changes *continuously* when the region is changed a little. Hadwiger's theorem says that every such S is a linear combination of three numbers: the area A, the boundary length L, and the Euler characteristic χ [110]. In this sense the three are a complete set of shape measures of this kind; any other such measure contains no new information. They are the **Minkowski functionals** of the region.

Working definition 12.2 (Minkowski functionals of an excursion set)

For a field $u = \phi/\sigma$ on a region of area A_{tot}, and a threshold ν:

- area fraction $a(\nu) = \text{area}(X_\nu)/A_{\text{tot}}$ [the share of pixels above ν];
- boundary length per unit area $\ell(\nu) = \text{length}(\partial X_\nu)/A_{\text{tot}}$ [the total length of the contour $u = \nu$ per unit area];
- Euler characteristic per unit area $c(\nu) = \chi(X_\nu)/A_{\text{tot}}$ [pieces minus holes per unit area].

On the sky, $v_0 = a$, $v_1 = \ell/4$ and $v_2 = c$ (with the sphere's own curvature removed, Section 12.5.6).

The area fraction and the boundary length are the "metric" functionals: a smooth remapping changes them (Section 12.1.1). The Euler characteristic is the topological one. Yet Hadwiger's theorem puts it on the same footing, as the third member of a family. To see how a count of pieces can be a functional of the same kind as a length, we need one more way of writing it.

12.4.2 The Euler characteristic as the total turning of the boundary

Walk once around the boundary of a disc, keeping the disc on your left. Your direction of travel turns steadily, and by the time you are back at the start it has turned through one full circle, 2π. Now walk around a lake inside an island, again keeping the land on your left: you go around the lake clockwise, and your direction turns through -2π. Deform the island and the lake however you like: the walk can wiggle, turning left and right, but every left turn that is not undone by a right turn would leave you facing a new direction at the end, and you must end facing the way you started, having gone once around. So the total turning of a closed boundary curve without self-crossings is exactly $\pm 2\pi$, whatever its shape (the turning-tangent theorem; see the topology notes).

The rate of turning per unit length is the **curvature** of the boundary, κ: on a circle of radius r it is $1/r$, so the total turning is $\kappa\times$length $= (1/r)(2\pi r) = 2\pi$. Add up the total turning over all the boundary curves of a region, each walked with the region on the left. Each piece has one outer boundary and contributes $+2\pi$; each hole has one boundary walked clockwise and contributes -2π. Therefore

$$\frac{1}{2\pi}\oint_{\partial X} \kappa \, \mathrm{d}l = \beta_0(X) - \beta_1(X) = \chi(X). \tag{12.11}$$

This is the bridge. The Euler characteristic is the integral, along the boundary, of a local geometric quantity, the curvature, exactly as the boundary length is the integral of 1 along the boundary. That is why it belongs with the area and the length, and why it is the quantity V_2 of [110] (their eq. 6, $V_2 = \frac{1}{2\pi}\int \kappa \, \mathrm{d}l$).

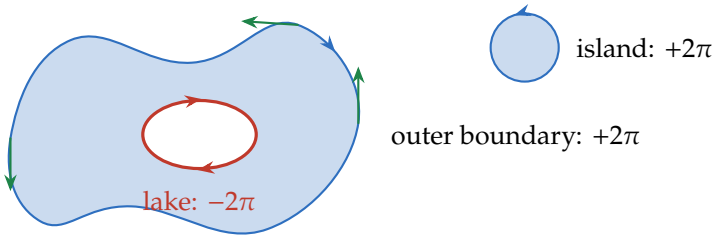


Figure 12.8: Walking each boundary with the set (blue) on the left. The outer boundary of a piece turns in total by $+2\pi$, however wiggly it is; the shore of a lake is walked clockwise and turns by -2π. The total turning divided by 2π is pieces minus holes, here $2-1=1$. The green arrows show the direction of travel at three places.

Example: a disc and a ring. A disc of radius r: curvature $1/r$ along a length $2\pi r$, so (12.11) gives $\frac{1}{2\pi}\cdot 2\pi = 1$. A ring with radii $r_1 < r_2$: the outer circle gives $+1$, the inner circle, walked clockwise, has curvature $-1/r_1$ along $2\pi r_1$ and gives -1; total 0. Neither result depends on the radii, as it must not.

Example: a cap on the sphere. On a curved surface the walk also turns because the surface itself curves, and (12.11) acquires a term from the area. Take the polar cap of angular radius θ on the unit sphere, the set of points within angle θ of the north pole. Its boundary is a circle of latitude of length $2\pi\sin\theta$, whose curvature measured within the sphere (the geodesic curvature) is $\cot\theta$. The surface has Gaussian curvature $K = 1$, and the cap has area $2\pi(1-\cos\theta)$. The Gauss–Bonnet theorem (see the topology notes) adds the integrated surface curvature to

the turning of the boundary:

$$\begin{aligned}\chi(\text{cap}) &\overset{(1)}{=} \frac{1}{2\pi}\Big[\oint \kappa_g \,\mathrm{d}l + \int_{\text{cap}} K \,\mathrm{d}A\Big] \\ &\overset{(2)}{=} \frac{1}{2\pi}\Big[\cot\theta \cdot 2\pi \sin\theta + 2\pi(1-\cos\theta)\Big] \\ &\overset{(3)}{=} \cos\theta + 1 - \cos\theta = 1. \end{aligned} \tag{12.12}$$

(1) Gauss–Bonnet for a region of a curved surface: boundary turning plus enclosed curvature, over 2π.
(2) Geodesic curvature $\cot\theta$ times the length $2\pi\sin\theta$ of the circle; $K = 1$ times the area of the cap.
(3) $\cot\theta \sin\theta = \cos\theta$.

The cap is one piece without holes at every size, and the two terms trade off exactly as θ grows. When $\theta \to \pi$ the boundary shrinks to a point near the south pole and the area term alone gives $\frac{1}{2\pi} \cdot 4\pi = 2$, the Euler characteristic of the whole sphere. This is the origin of the extra term in the full-sky formula of [110] (their eq. 7), $\chi = V_2 + V_0/(2\pi R^2)$ on a sphere of radius R, which Section 12.5.6 needs.

12.4.3 Area and length of a Gaussian field, recalled in the plane

Let $u = \phi/\sigma$ be a statistically homogeneous and isotropic Gaussian field with zero mean and unit variance in the plane. The quantity that sets the scale of everything below is the variance of one component of its gradient,

$$\lambda \equiv \big\langle (\partial_x u)^2 \big\rangle = \tfrac{1}{2}\big\langle |\nabla u|^2 \big\rangle = \frac{\sigma_1^2}{2\sigma_0^2}, \tag{12.13}$$

where σ_0^2 is the variance of ϕ and $\sigma_1^2 = \langle |\nabla\phi|^2 \rangle$ the variance of its gradient, the spectral moments of (7.35). The two components contribute equally by isotropy, hence the $\frac{1}{2}$. On the sky, (T3.5) defined the correlation angle $\theta_c = \sigma_0/\sigma_1$, so $\lambda = 1/(2\theta_c^2)$. A large λ means a field that changes quickly, with many small spots.

The area. The expected fraction of the plane above ν is the probability that the field at one point exceeds ν (the plane average equals the ensemble average at a point, by homogeneity and ergodicity, Section 7.5):

$$\mathbb{E}[a(\nu)] = \mathrm{P}(u > \nu) = \Psi(\nu), \tag{12.14}$$

with Ψ the Gaussian tail of (12.10).

The length. Thicken the contour $u = \nu$ into the thin band $\nu < u < \nu + \mathrm{d}\nu$. Where the gradient is $|\nabla u|$, the band has width $\mathrm{d}\nu/|\nabla u|$, so its area is the contour length times $\mathrm{d}\nu/|\nabla u|$, averaged along the contour. Turned around, the length per unit area is the area fraction of the band, weighted by $|\nabla u|$ and divided by $\mathrm{d}\nu$, which is the expectation of $|\nabla u|\,\delta(u-\nu)$:

$$\mathbb{E}[\ell(\nu)] \overset{(1)}{=} \mathbb{E}\left[|\nabla u|\,\delta(u-\nu)\right] \overset{(2)}{=} \frac{e^{-\nu^2/2}}{\sqrt{2\pi}}\,\mathbb{E}\,|\nabla u| \overset{(3)}{=} \frac{e^{-\nu^2/2}}{\sqrt{2\pi}}\sqrt{\frac{\pi\lambda}{2}} \overset{(4)}{=} \frac{\sqrt{\lambda}}{2}\,e^{-\nu^2/2}. \tag{12.15}$$

(1) The band argument above (the coarea formula).
(2) At one point u and ∇u are uncorrelated, $\langle u\,\partial_i u\rangle = \frac{1}{2}\partial_i\langle u^2\rangle = 0$ for a homogeneous field, and uncorrelated jointly Gaussian variables are independent; the average of $\delta(u-\nu)$ is the standard normal density at ν.

(3) $|\nabla u|$ is the length of a two-component vector with independent $\mathcal{N}(0,\lambda)$ components, a Rayleigh variable with mean $\sqrt{\lambda}\sqrt{\pi/2}$ (Section T3.1.2).

(4) $\frac{1}{\sqrt{2\pi}}\sqrt{\frac{\pi}{2}} = \frac{1}{2}$.

On the sky (T3.6) is the same result with $v_1 = \ell/4$ and $\lambda = 1/(2\theta_c^2)$: $\frac{1}{4}\cdot\frac{1}{2\sqrt{2}\,\theta_c} = \frac{1}{8\sqrt{2}\,\theta_c}$. The Euler characteristic is the one functional that needs the second derivatives of the field, because curvature is a second derivative. Its derivation is the next section.

Example: the numbers on simulated maps. The library function `mfs_flat` measures all three on a periodic map: the area fraction by counting pixels, the length from (12.15) with the delta function replaced by a narrow top hat of width 0.1 and the gradient computed exactly by the fast Fourier transform, and χ by $V - E + F$. Over 400 maps of 512×512 pixels with $\lambda = 0.01391$ per pixel^2 (Section 12.6), the mean area fraction at $\nu = 1$ is 0.1589 against $\Psi(1) = 0.1587$, and the mean length per unit area at $\nu = 0$ is 0.059 against 0.059 per pixel from (12.15), a difference of 0.01% (Figure 12.9).

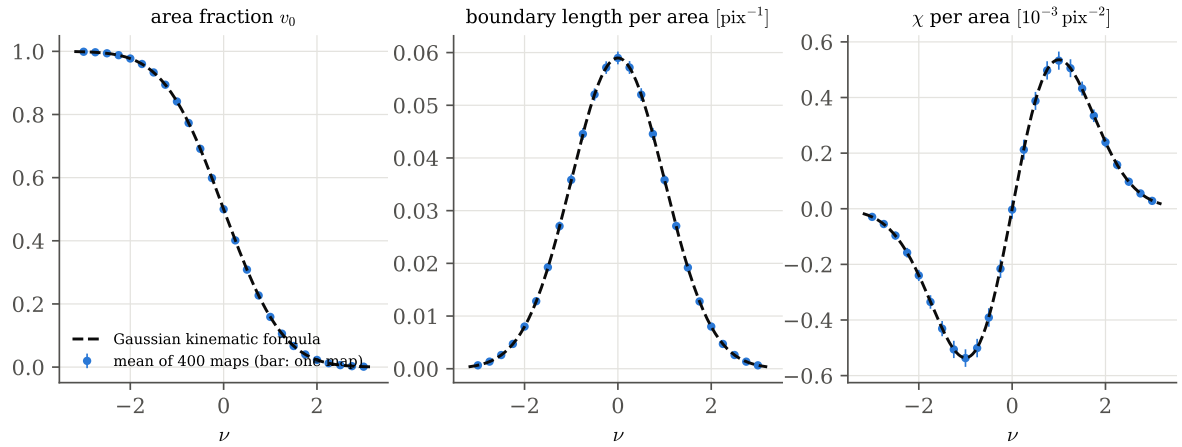


Figure 12.9: The three Minkowski functionals of the excursion sets of 400 Gaussian maps (512×512 pixels, white noise smoothed on 6 pixels). Points: the mean over the maps; bars: the scatter of a single map. Dashed: the Gaussian predictions (12.14), (12.15) and, for χ, (12.23).

12.5 The expected Euler characteristic of a Gaussian field

Counting pieces needs the whole map: to know whether two hot pixels at opposite ends of a long filament belong to one piece we must follow the filament. An average over random fields of such a global count looks hopeless. Yet for a Gaussian field the average Euler characteristic at every threshold has a short closed formula. The trick that makes it possible is to trade the global count for a sum of local events, each of which happens at a single point, and then to ask how often such points occur. That is a question about derivatives of the field at one point, and for a Gaussian field the joint distribution of the value and the derivatives at one point is known exactly. Once the formula is in hand, it tells us what $\chi(\nu)$ can and cannot know about the power spectrum.

The question. For a homogeneous, isotropic Gaussian field u with zero mean and unit variance, what is the expected Euler characteristic of $X_\nu = \{u > \nu\}$ per unit area?

12.5.1 From a global count to local events

Go back to the flooding picture of Section 12.3, and lower the water level ν from far above the highest peak. The dry land X_ν changes its pieces and holes only at particular levels, and each change happens at one point. Which points? A point where the land is sloped is boring: lowering the water a little just moves the shore there. Something can happen only where the ground is flat, at a **critical point** of the height, where the gradient vanishes. In two dimensions there are three kinds, told apart by the second derivatives (the Hessian matrix $H_{ij} = \partial_i \partial_j u$):

- *A peak* (both eigenvalues of H negative). When the water drops below it, a new island appears: β_0 goes up by one, so χ goes up by one.
- *A saddle* (one eigenvalue of each sign), a mountain pass. When the water drops below it, the land on its two sides joins through the pass. Either the two sides were different islands, which merge (β_0 down by one), or they were already connected the long way round, and the new connection closes a ring around a lake (β_1 up by one). Either way χ goes down by one.
- *A pit* (both eigenvalues positive), the bottom of a lake. When the water drops below it, the lake dries up: β_1 goes down by one and χ goes up by one.

Starting from the empty set ($\chi = 0$) at very high water and adding up the changes,

$$\chi(X_\nu) = \#\{\text{peaks above } \nu\} - \#\{\text{saddles above } \nu\} + \#\{\text{pits above } \nu\}, \tag{12.16}$$

for a field on a domain without edges (a periodic box; edges are added in Section 12.5.5). This is the counting rule of Morse theory, and it is the second definition of the Euler characteristic quoted by [115] (their eq. 7). Its sign pattern can be written with the Hessian alone: the determinant $\det H$ is positive at peaks and pits (two eigenvalues of the same sign) and negative at saddles. So

$$\chi(X_\nu) = \sum_{\text{critical points } \boldsymbol{x}_c \text{ with } u(\boldsymbol{x}_c) > \nu} \operatorname{sign} \det H(\boldsymbol{x}_c). \tag{12.17}$$

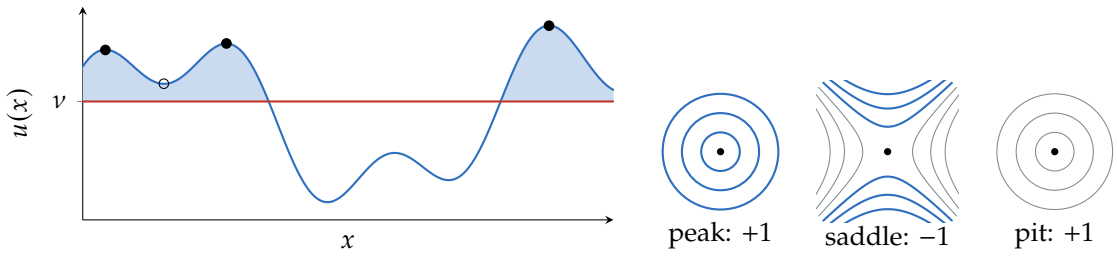


Figure 12.10: Left: in one dimension the excursion set above ν (red line) is a union of intervals (shaded). Each interval contains one more peak (filled dots) than pits (open dot) above the level: two peaks and one pit in the first, one peak in the second. Right: the three kinds of critical point in two dimensions, drawn as contour lines (blue rising towards the centre, grey falling). Lowering the threshold past each changes χ by the amount shown.

12.5.2 The simplest case first: a line

On a line the excursion set is a union of intervals, and its Euler characteristic is the number of intervals. Lowering the level, a new interval appears at a peak and two intervals merge at a pit, so the counting rule reads $\chi = \#\{\text{peaks above } \nu\} - \#\{\text{pits above } \nu\}$ (Figure 12.10, left). On a long line each interval begins where the field crosses ν going up, so the number of intervals per

unit length is also the rate of *up-crossings* of the level. We compute it both ways; the agreement is the first check of the method.

The ingredients. At one point x we need the joint distribution of the value u, the slope u' and the curvature u''. They are jointly Gaussian with zero mean (derivatives of a Gaussian field are linear combinations of its values, so they are Gaussian too). Write $\lambda = \langle u'^2 \rangle$, the one-dimensional version of (12.13). Every covariance we need follows from one identity. For a homogeneous field the average of any product of the field and its derivatives is the same at every point, so its derivative vanishes: $\frac{\mathrm{d}}{\mathrm{d}x}\langle f\, g\rangle = 0$, hence

$$\langle f' g\rangle = -\langle f g'\rangle \qquad \text{(moving a derivative costs a sign).} \tag{12.18}$$

With $f = g = u$ it gives $\langle u'u\rangle = -\langle uu'\rangle$, so $\langle uu'\rangle = 0$. With $f = u'$, $g = u'$ it gives $\langle u''u'\rangle = 0$. With $f = u'$, $g = u$ it gives $\langle u''u\rangle = -\langle u'u'\rangle = -\lambda$. So the slope is uncorrelated with both the value and the curvature, and being jointly Gaussian it is independent of them; the value and the curvature are correlated, with covariance $-\lambda$.

How many up-crossings in a short step $\mathrm{d}x$*?* An up-crossing happens between x and $x + \mathrm{d}x$ if the field starts below the level and ends above it, $u < v < u + u'\mathrm{d}x$, which needs $u' > 0$ and $v - u'\mathrm{d}x < u < v$. The probability of that is

$$\begin{aligned}\mathrm{P}(\text{up-crossing in } \mathrm{d}x) &\overset{(1)}{=} \int_0^\infty \mathrm{d}u' \int_{v-u'\mathrm{d}x}^{v} p(u,u')\,\mathrm{d}u \overset{(2)}{=} \mathrm{d}x \int_0^\infty u'\, p(v,u')\,\mathrm{d}u' \\ &\overset{(3)}{=} \mathrm{d}x\, \frac{e^{-v^2/2}}{\sqrt{2\pi}} \int_0^\infty u' \,\frac{e^{-u'^2/2\lambda}}{\sqrt{2\pi\lambda}}\,\mathrm{d}u' \overset{(4)}{=} \mathrm{d}x\, \frac{e^{-v^2/2}}{\sqrt{2\pi}}\,\sqrt{\frac{\lambda}{2\pi}} = \mathrm{d}x\,\frac{\sqrt{\lambda}}{2\pi}\, e^{-v^2/2}.\end{aligned} \tag{12.19}$$

(1) Integrate the joint density of value and slope over the region just described.
(2) The inner range has length $u'\mathrm{d}x$, so for small $\mathrm{d}x$ the inner integral is the integrand at $u = v$ times $u'\mathrm{d}x$.
(3) The value and the slope are independent, so the joint density factorises into two Gaussians, of variance 1 and λ.
(4) $\int_0^\infty s\, e^{-s^2/2\lambda}\mathrm{d}s = \lambda$, divided by $\sqrt{2\pi\lambda}$.

This is Rice's formula for the rate of level crossings [264]: the expected number of intervals of $\{u > v\}$ per unit length is $\frac{\sqrt{\lambda}}{2\pi}e^{-v^2/2}$.

The same number by counting critical points. Now use the counting rule instead. A critical point is a zero of the slope u'. Near a zero x_c the slope is $u'(x) \approx u''(x_c)(x - x_c)$, so the delta function $\delta(u'(x))$ integrated over a small neighbourhood gives $1/|u''(x_c)|$; multiplied by $|u''|$ it counts the zero exactly once. Peaks have $u'' < 0$ and pits $u'' > 0$, so $\delta(u')\,(-u'')$ counts peaks as $+1$ and pits as -1, which is exactly the sign rule. Averaging:

$$\begin{aligned}\frac{\mathbb{E}[\chi]}{\text{length}} &\overset{(1)}{=} \mathbb{E}\left[\delta(u')\,(-u'')\,\mathbb{1}\{u > v\}\right] \overset{(2)}{=} p_{u'}(0)\;\mathbb{E}\left[(-u'')\,\mathbb{1}\{u > v\}\right] \\ &\overset{(3)}{=} \frac{1}{\sqrt{2\pi\lambda}}\int_v^\infty \lambda u\,\frac{e^{-u^2/2}}{\sqrt{2\pi}}\,\mathrm{d}u \overset{(4)}{=} \frac{\lambda}{\sqrt{2\pi\lambda}}\,\frac{e^{-v^2/2}}{\sqrt{2\pi}} = \frac{\sqrt{\lambda}}{2\pi}\,e^{-v^2/2}.\end{aligned} \tag{12.20}$$

(1) The counting rule, with each critical point counted by $\delta(u')|u''|$ and signed by $-\operatorname{sign} u''$; the indicator $\mathbb{1}\{u > v\}$ keeps only critical points above the level. On a homogeneous line the average per unit length is the average at one point.
(2) The slope is independent of the value and the curvature, so the average splits; the average of $\delta(u')$ is the density of u' at zero.

(3) The density of u' at zero is $1/\sqrt{2\pi\lambda}$. For the curvature we only need its average at a given value u. For jointly Gaussian variables with zero mean the conditional mean is linear, $\mathbb{E}[Y \mid X] = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}\,X} X$ (Chapter 2); with $\mathrm{Cov}(u, u'') = -\lambda$ and $\mathrm{Var}\,u = 1$, $\mathbb{E}[-u'' \mid u] = \lambda u$.

(4) $\int_v^\infty u\, e^{-u^2/2} \mathrm{d}u = e^{-v^2/2}$.

The two routes agree. The second one is the one that generalises, because it never needed to know what an interval is, only how to recognise and sign a critical point.

12.5.3 The plane

Counting critical points in two dimensions. The critical points are the zeros of the gradient ∇u, a map from the plane to the plane. Near a zero, $\nabla u(\boldsymbol{x}) \approx H(\boldsymbol{x}_c)(\boldsymbol{x} - \boldsymbol{x}_c)$, and the two-dimensional delta function $\delta^2(\nabla u)$ integrated over a neighbourhood gives $1/|\det H|$, the inverse Jacobian of that linear map. So $\delta^2(\nabla u)\,|\det H|$ counts each critical point once, and $\delta^2(\nabla u)\det H$ counts it with the sign of (12.17). Averaging over the field, per unit area,

$$\mathbb{E}[c(v)] = \mathbb{E}\left[\delta^2(\nabla u)\,\det H\;\mathbb{1}\{u > v\}\right]. \tag{12.21}$$

This is the two-dimensional version of the first line of (12.20), the Kac–Rice counting formula. The absolute value that would make the average of $|\det H|$ hard has disappeared, because the sign rule wants $\det H$ itself, a polynomial in the second derivatives.

The ingredients. The identity (12.18) holds for derivatives in any direction. It gives, exactly as on the line:

- $\langle u\, \partial_i u\rangle = 0$ and $\langle \partial_i u\, \partial_j \partial_k u\rangle = 0$: the gradient is independent of the value and of the Hessian at the same point.
- $\langle u\, \partial_i \partial_j u\rangle = -\langle \partial_i u\, \partial_j u\rangle = -\lambda \delta_{ij}$: by isotropy the two gradient components have the same variance λ and are uncorrelated.
- $\langle \partial_x^2 u\, \partial_y^2 u\rangle = -\langle \partial_x u\, \partial_x \partial_y^2 u\rangle = \langle \partial_x \partial_y u\, \partial_x \partial_y u\rangle$: moving one ∂_x across and then one ∂_y back. Call this common value m.

The last line is the one non-obvious fact, and the whole calculation turns on it.

The average determinant at a given value. Write $u_{xx} = \partial_x^2 u$ and so on. Given the value u, the conditional means are $\mathbb{E}[u_{xx} \mid u] = \mathbb{E}[u_{yy} \mid u] = -\lambda u$ and $\mathbb{E}[u_{xy} \mid u] = 0$ (the linear rule of Section 12.5.2 with the covariances just listed). The conditional covariances are the unconditional ones minus the part explained by u: $\mathrm{Cov}(u_{xx}, u_{yy} \mid u) = m - \lambda^2$ and $\mathrm{Var}(u_{xy} \mid u) = m - 0^2 = m$. Then

$$\begin{aligned} \mathbb{E}[\det H \mid u] &\overset{(1)}{=} \mathbb{E}[u_{xx}u_{yy} \mid u] - \mathbb{E}[u_{xy}^2 \mid u] \\ &\overset{(2)}{=} \left[(-\lambda u)^2 + m - \lambda^2\right] - \left[0^2 + m\right] \\ &\overset{(3)}{=} \lambda^2(u^2 - 1). \end{aligned} \tag{12.22}$$

(1) $\det H = u_{xx}u_{yy} - u_{xy}^2$.

(2) For any two variables, $\mathbb{E}[YZ \mid u] = \mathbb{E}[Y \mid u]\,\mathbb{E}[Z \mid u] + \mathrm{Cov}(Y, Z \mid u)$.

(3) The unknown fourth moment m cancels, by the identity $\langle u_{xx}u_{yy}\rangle = \langle u_{xy}^2\rangle$.

Everything about the spectrum beyond λ has dropped out.

Putting it together.

$$\mathbb{E}[c(\nu)] \overset{(1)}{=} p_{\nabla u}(\mathbf{0})\ \mathbb{E}\left[\det H\ \mathbb{1}\{u>\nu\}\right] \overset{(2)}{=} \frac{1}{2\pi\lambda}\int_{\nu}^{\infty}\lambda^2(u^2-1)\,\frac{e^{-u^2/2}}{\sqrt{2\pi}}\,\mathrm{d}u$$
$$\overset{(3)}{=} \frac{\lambda}{2\pi}\,\frac{\nu\,e^{-\nu^2/2}}{\sqrt{2\pi}} = \frac{\lambda}{(2\pi)^{3/2}}\,\nu\,e^{-\nu^2/2}. \tag{12.23}$$

(1) The gradient is independent of the value and the Hessian, so the average of $\delta^2(\nabla u)$ factors out as the density of ∇u at zero.
(2) That density is $(2\pi\lambda)^{-1}$, two independent $\mathcal{N}(0,\lambda)$ components at zero; then average $\det H$ first at fixed u with (12.22) and integrate over u above ν.
(3) $\frac{\mathrm{d}}{\mathrm{d}u}\left[-u\,e^{-u^2/2}\right] = (u^2-1)e^{-u^2/2}$, so the integral is $\nu e^{-\nu^2/2}$.

Key idea: Expected Euler characteristic of a Gaussian excursion set

For a homogeneous, isotropic Gaussian field in units of its standard deviation, with $\lambda = \langle(\partial_x u)^2\rangle = \sigma_1^2/2\sigma_0^2$, the expected Euler characteristic of $\{u>\nu\}$ per unit area is

$$\mathbb{E}[c(\nu)] = \frac{\lambda}{(2\pi)^{3/2}}\,\nu\,e^{-\nu^2/2}.$$

The shape $\nu e^{-\nu^2/2}$ is the same for every spectrum; the spectrum enters only through the single number λ.

What the formula knows about the spectrum. The whole power spectrum $P(k)$ enters through one number. Write $\sigma_j^2 = \int \frac{\mathrm{d}^2k}{(2\pi)^2}k^{2j}P(k)$ for the spectral moments of (7.35); then $\lambda = \sigma_1^2/(2\sigma_0^2)$, half the mean of k^2 weighted by the power. Two spectra with the same λ give the same expected χ curve, however different they look. For example, the family $P \propto k^n e^{-k^2R^2}$ has

$$\lambda \overset{(1)}{=} \frac{1}{2}\,\frac{\int_0^\infty k^{3+n}e^{-k^2R^2}\mathrm{d}k}{\int_0^\infty k^{1+n}e^{-k^2R^2}\mathrm{d}k} \overset{(2)}{=} \frac{1}{2R^2}\,\frac{\Gamma\left(\frac{n}{2}+2\right)}{\Gamma\left(\frac{n}{2}+1\right)} \overset{(3)}{=} \frac{n+2}{4R^2}. \tag{12.24}$$

(1) $\mathrm{d}^2k = 2\pi k\,\mathrm{d}k$ for an isotropic integrand; the 2π's cancel in the ratio.
(2) Substitute $t = k^2R^2$: $\int_0^\infty k^{2s-1}e^{-k^2R^2}\mathrm{d}k = \Gamma(s)/(2R^{2s})$, with $s = \frac{n}{2}+2$ and $s = \frac{n}{2}+1$.
(3) $\Gamma(s+1) = s\,\Gamma(s)$.

So $n=0$ with $R=6$, $n=-1$ with $R=\sqrt{18}$ and $n=2$ with $R=\sqrt{72}$ all have $\lambda = 1/72$ and the same expected χ curve, although the first is smooth blobs, the second has much more small-scale structure, and the third has a missing long-wavelength part (Section 12.6 tests this).

Two consequences follow, and they are the reason the formula matters for the rest of the chapter. For a Gaussian field, the Euler characteristic curve carries no information beyond the power spectrum: it is fixed by one of its moments. And the *shape* of the curve is universal, $\nu e^{-\nu^2/2}$, so any measured departure from that shape is a model-independent sign that the field is not Gaussian. Topology is therefore most useful exactly where the power spectrum is not the whole story, which is the situation of Section 7.7 and of part 12c.

Example: a 512×512 map. With $\lambda = 1/72$ per pixel^2 and $512^2 = 262144$ pixels, the expected χ at $\nu = 1$ is

$$262144 \times \frac{1/72}{(2\pi)^{3/2}} \times 1 \times e^{-1/2} = 262144 \times \frac{0.01389}{15.75} \times 0.6065 \approx 140.$$

The map has about 140 more pieces than holes at one standard deviation. At $\nu = 0$ the expectation is zero: pieces and holes balance exactly, on average, at the median.

12.5.4 Three dimensions

In a box the excursion set is a solid region. Lowering the level through a critical point now changes χ by $+1$ at a peak (three negative eigenvalues, a new piece), by -1 at a saddle with two negative eigenvalues (two pieces merge, or a loop closes around a tunnel), by $+1$ at a saddle with one negative eigenvalue (a tunnel is closed off or a cavity is sealed), and by -1 at a pit (a cavity fills). In each case the change is $-\operatorname{sign}\det H$, because $\det H$ is the product of three eigenvalues. So $\mathbb{E}[c_3(\nu)] = \mathbb{E}[\delta^3(\nabla u)(-\det H)\mathbb{1}\{u > \nu\}]$ per unit volume, with $\lambda = \langle(\partial_x u)^2\rangle = \sigma_1^2/3\sigma_0^2$ now a third of the mean of k^2.

The average determinant. Given u, write $H = -\lambda u\,\mathbb{1} + Z$, with Z the fluctuating part of the Hessian, which has zero conditional mean. The determinant of $a\mathbb{1} + Z$ expands as $a^3 + a^2 \operatorname{tr} Z + a\, M_2(Z) + \det Z$, where M_2 is the sum of the three principal 2×2 minors $Z_{xx}Z_{yy} - Z_{xy}^2$ and so on. Then

$$\mathbb{E}[\det H \mid u] \overset{(1)}{=} a^3 + a\;\mathbb{E}[M_2(Z) \mid u] \overset{(2)}{=} a^3 + 3a\big[(m - \lambda^2) - m\big] \overset{(3)}{=} -\lambda^3 u^3 + 3\lambda^3 u = -\lambda^3(u^3 - 3u). \tag{12.25}$$

(1) $\mathbb{E}[\operatorname{tr} Z \mid u] = 0$, and $\mathbb{E}[\det Z \mid u] = 0$ because $\det Z$ is a sum of products of three zero-mean jointly Gaussian variables, whose average vanishes.
(2) Each of the three minors averages to $\operatorname{Cov}(u_{xx}, u_{yy} \mid u) - \operatorname{Var}(u_{xy} \mid u)$, exactly as in (12.22).
(3) $a = -\lambda u$.

In two dimensions the same expansion, $\det(a\mathbb{1} + Z) = a^2 + a \operatorname{tr} Z + \det Z$, reproduces (12.22). Now

$$\mathbb{E}[c_3(\nu)] \overset{(1)}{=} \frac{1}{(2\pi\lambda)^{3/2}} \int_\nu^\infty \lambda^3(u^3 - 3u)\frac{e^{-u^2/2}}{\sqrt{2\pi}}\,\mathrm{d}u \overset{(2)}{=} \frac{\lambda^{3/2}}{(2\pi)^2}\,(\nu^2 - 1)\,e^{-\nu^2/2}. \tag{12.26}$$

(1) The density of the three gradient components at zero, and $\mathbb{E}[-\det H \mid u]$ from (12.25).
(2) $\frac{\mathrm{d}}{\mathrm{d}u}\big[-(u^2-1)e^{-u^2/2}\big] = (u^3 - 3u)e^{-u^2/2}$.

At the median, $\nu = 0$, the Euler characteristic is negative: a Gaussian field in three dimensions is a sponge, with more tunnels than pieces, the property that Section T3.6.1 recalled and Section 12.7.1 reproduces.

12.5.5 Edges

A real map is a square with edges, not a torus. The counting rule still works if we count the critical points of the field restricted to the closed square, and these include points on the edges. Take the bottom edge, with the square above it, and let x run along the edge. A point of the edge behaves like a peak of the field on the closed square if the field has a maximum *along* the edge there ($u_x = 0$, $u_{xx} < 0$) and decreases *into* the square ($u_y < 0$): lowering the water past it creates a new piece that touches the edge, $+1$. If instead it is a minimum along the edge and decreases into the square, it behaves like a saddle, -1. Points where the field increases into the square are not critical points of the restricted field. So the edge contributes, per unit length,

$$\mathbb{E}\left[\delta(u_x)(-u_{xx})\,\mathbb{1}\{u_y < 0\}\,\mathbb{1}\{u > \nu\}\right] \overset{(1)}{=} \mathrm{P}(u_y < 0)\;\mathbb{E}\left[\delta(u_x)(-u_{xx})\,\mathbb{1}\{u > \nu\}\right] \overset{(2)}{=} \frac{1}{2}\,\frac{\sqrt{\lambda}}{2\pi}\,e^{-\nu^2/2}. \tag{12.27}$$

(1) The normal derivative u_y is uncorrelated with u, u_x and u_{xx} (by (12.18) and isotropy), hence independent of them.
(2) u_y is symmetric about zero; the remaining average is the one-dimensional count (12.20).

A corner is a critical point of the restricted field when the field decreases into the square along both edges, which happens with probability $\frac{1}{4}$, and it then counts +1 if it is above ν. For an $L \times L$ square, with four edges and four corners,

$$\mathbb{E}[\chi_{\text{square}}(\nu)] = L^2 \frac{\lambda\, \nu\, e^{-\nu^2/2}}{(2\pi)^{3/2}} + 4L \cdot \frac{1}{2} \frac{\sqrt{\lambda}\, e^{-\nu^2/2}}{2\pi} + 4 \cdot \frac{1}{4} \Psi(\nu). \tag{12.28}$$

The edge term is largest at $\nu = 0$, where the area term vanishes. It is the reason the threshold scan of Figure 12.6 does not pass through zero at $\nu = 0$, and it is why masks matter: every edge of a mask adds such a term, and a mask with a long, ragged boundary adds a lot.

12.5.6 The sphere, and the general formula

On the full sky there are no edges, but the domain is curved. The Gauss–Bonnet form of χ from Section 12.4.2 splits the count into two parts: the turning of the boundary of X_ν, and the curvature of the sphere enclosed by X_ν, $\chi = \frac{1}{2\pi}\big[\oint \kappa_g \, \mathrm{d}l + \int_{X_\nu} K \, \mathrm{d}A\big]$, with $K = 1/R^2$. The first part is a sum of local contributions along the contours and has, per unit area, the same expectation $\lambda \nu e^{-\nu^2/2}/(2\pi)^{3/2}$ as in the plane, with λ now measured per steradian. The second part is $\frac{1}{2\pi R^2}$ times the area of X_ν, whose expectation is $4\pi R^2 \Psi(\nu)$. So, on the unit sphere,

$$\mathbb{E}[\chi_{\text{sky}}(\nu)] = 4\pi \frac{\lambda\, \nu\, e^{-\nu^2/2}}{(2\pi)^{3/2}} + 2\,\Psi(\nu), \qquad \lambda = \frac{\sum_\ell (2\ell+1) C_\ell\, \ell(\ell+1)/2}{\sum_\ell (2\ell+1) C_\ell}, \tag{12.29}$$

where $\lambda = \sigma_1^2/2\sigma_0^2$ follows from the sums (T3.3) and (T3.4). At very low thresholds the excursion set covers the whole sky and $\chi \to 2$, the Euler characteristic of a sphere; that is the 2Ψ term. That the plane density carries over exactly to the sphere, with no correction from the curvature of the sky, is a theorem [265]; Section 12.7.2 checks it numerically.

The formulas (12.14), (12.20), (12.23), (12.26), (12.28) and (12.29) are all instances of a single statement, the **Gaussian kinematic formula** [265, 266]:

$$\mathbb{E}[\chi(X_\nu \cap \mathcal{M})] = \sum_{j=0}^{d} \mathcal{L}_j(\mathcal{M})\, \rho_j(\nu), \qquad \begin{aligned} \rho_0 &= \Psi(\nu), & \rho_1 &= \tfrac{\lambda^{1/2}}{2\pi} e^{-\nu^2/2}, \\ \rho_2 &= \tfrac{\lambda}{(2\pi)^{3/2}}\, \nu\, e^{-\nu^2/2}, & \rho_3 &= \tfrac{\lambda^{3/2}}{(2\pi)^2} (\nu^2 - 1) e^{-\nu^2/2}. \end{aligned} \tag{12.30}$$

The densities ρ_j come from the field: ρ_j is the expected χ per unit j-dimensional volume, computed by the counting argument in j dimensions. The numbers $\mathcal{L}_j(\mathcal{M})$ come from the domain alone (they are its Lipschitz–Killing curvatures): $\mathcal{L}_0$ is its Euler characteristic, $\mathcal{L}_1$ half its boundary length (in two dimensions), $\mathcal{L}_2$ its area. For a square of side L, $(\mathcal{L}_0, \mathcal{L}_1, \mathcal{L}_2) = (1, 2L, L^2)$, which is (12.28); for the unit sphere $(2, 0, 4\pi)$, which is (12.29); for a torus $(0, 0, A)$. A mask enters only through the $\mathcal{L}_j$ of the unmasked region. The same structure, with the ρ_j replaced by other Hermite polynomials, gives the expected area and boundary length; [267] derived all the Minkowski functionals of Gaussian fields in any dimension, and Section 12.7.3 matches our results to Matsubara's general expression.

12.6 Testing the formula on simulated maps

The derivation of Section 12.5 made several moves that deserve to be checked: the counting rule that turned a global count into critical points, the independence of the gradient from

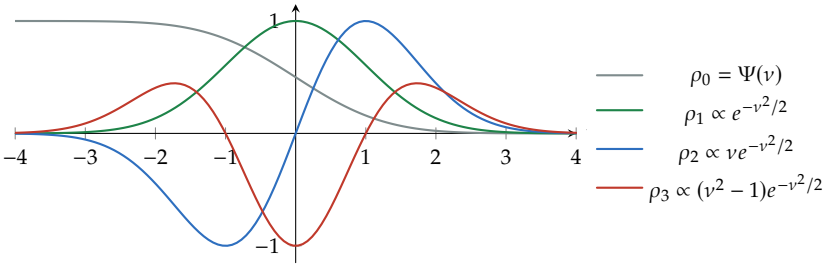


Figure 12.11: The shapes of the four Euler-characteristic densities, each scaled to a maximum of one in absolute value. In one dimension χ counts intervals, positive everywhere; in two dimensions it changes sign at the median; in three it is negative (sponge-like) between $\nu = -1$ and $\nu = 1$. (Ψ is drawn with a standard smooth approximation to the normal tail.)

everything else, the cancellation of the fourth moment m, and the claim that nothing but λ survives. A formula derived by hand and code written by hand can both be wrong; when they agree on thousands of simulated maps, both are probably right. And the simulations give something the formula does not: how much χ scatters from one map to the next, which is what any measurement on a single sky needs.

The experiment. The script `code/ch12/04_gkf_plane.py` draws 400 periodic Gaussian maps of 512×512 pixels for each of the three spectra $P \propto k^n e^{-k^2R^2}$ of (12.24) with $n = 0, -1, 2$ and $R = 6, \sqrt{18}, \sqrt{72}$ pixels, which all have $\lambda = 1/72 = 0.01389$ per pixel^2 in the continuum. On the finite grid the sums over the box modes give slightly different values, 0.01391, 0.01441 and 0.01389 (the $n = -1$ spectrum has the most power near the grid scale, where the lattice sum departs most from the integral), and the script uses these exact lattice values in the predictions. For each map it computes the three Minkowski functionals at 25 thresholds from -3 to 3, with χ from $V - E + F$ on the torus, and the same maps are also treated as squares with edges. The run takes a few minutes and is cached in `data/ch12/`.

The mean. For $n = 0$ the mean of χ at $\nu = 1$ over the 400 maps is 139.6, with a Monte Carlo error of 0.4, against 140.5 from (12.23). At $\nu = -1$ the maps give -140.8 against -140.5, and at $\nu = 0$ they give -0.82 ± 0.43 against exactly zero. Over all 25 thresholds the largest deviation of the mean from the formula is 2.49 times its Monte Carlo error, what one expects from chance among 25 comparisons. The area and length agree equally well (Figure 12.9).

One number from the spectrum. Figure 12.12 shows small corners of one map of each spectrum, which look clearly different: the $n = -1$ map is rough, the $n = 2$ map is dominated by a single scale. Yet their mean χ curves lie on the same line. At $\nu = 1$ the $n = -1$ maps give 144.9 (prediction 145.5 with its lattice λ), the $n = 2$ maps give 140.1 (prediction 140.2). This is the content of the key idea of Section 12.5.3 seen on data: the mean Euler characteristic of a Gaussian field cannot tell these three spectra apart.

But not the same scatter. The standard deviation of χ at $\nu = 1$ for a single map is 8.6 for $n = 0$, 11 for $n = -1$ and 7 for $n = 2$. The means are equal, the spreads are not. The variance of a count of critical points depends on how critical points are correlated with one another, which involves the whole correlation function and not only its curvature at the origin. So for a Gaussian field the mean χ curve carries one number of the spectrum, while its covariance carries more. Part 12c, which builds likelihoods for topological summaries, has to take the covariance from simulations for exactly this reason: no formula as simple as (12.23) exists for it.

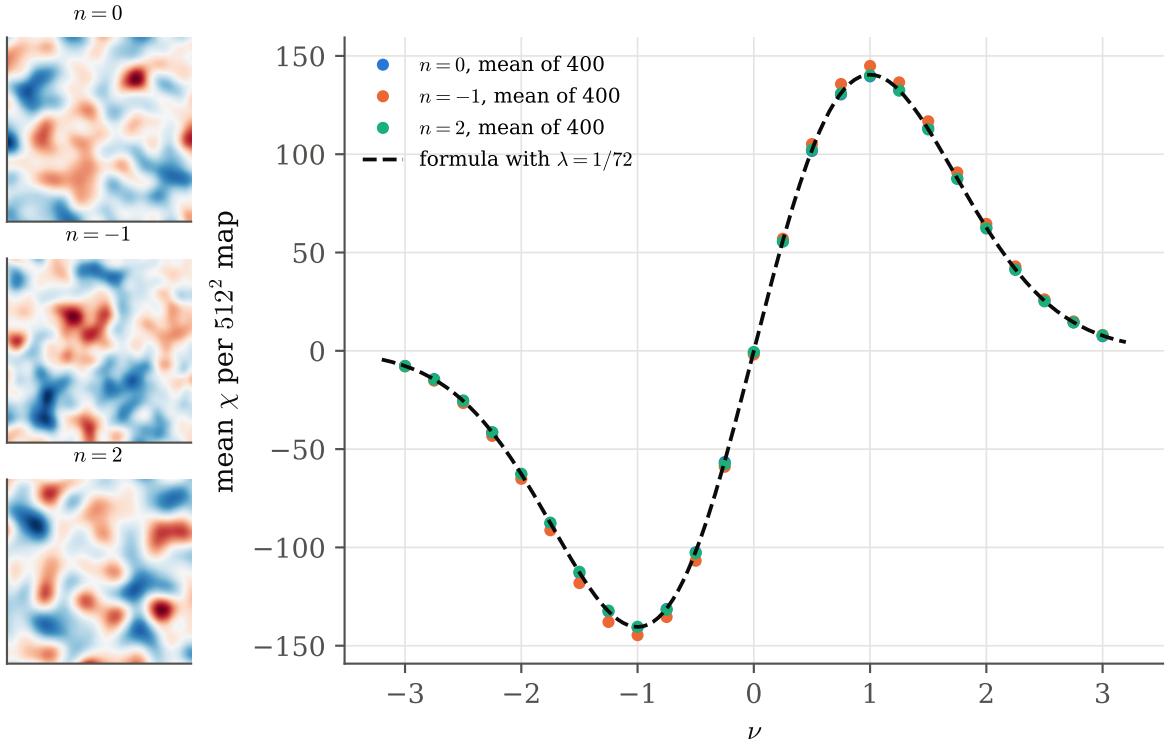


Figure 12.12: Three Gaussian fields with different spectra, $P \propto k^n e^{-k^2R^2}$ with $n = 0, -1, 2$, chosen to have the same λ. Left: a 128×128 corner of one map of each. Right: their mean χ over 400 maps of 512×512 pixels, all on the single prediction (dashed).

Example: what the edges add. The same $n = 0$ maps, cut open into 512×512 squares (no periodic gluing), have more pieces than the torus: pieces that wrapped across the seam are now counted twice, and pieces touching an edge are no longer balanced by holes. The edge term of (12.28), $2L\sqrt{\lambda}\, e^{-\nu^2/2}/2\pi$, is 19.2 at $\nu = 0$, and the full square formula predicts 19.7; the squares give 18.9 ± 0.5. At $\nu = 1$ the prediction is 152.3 and the squares give 151.2 (Figure 12.13, left). The deficit at $\nu = 0$, about one unit or one and a half Monte Carlo errors, is of the size expected from the pixel grid along the edges, where the counting convention matters most; we do not pursue it further. The lesson is quantitative: on a 512×512 map the edges contribute as much to χ at the median as all the topology inside, and a masked sky, with its long mask boundary, is affected even more.

Example: the sampling distribution of χ. The right panel of Figure 12.13 is the histogram of χ at $\nu = 1$ over the 400 torus maps. It is centred on the formula and close to a Gaussian with standard deviation 8.6, about 6% of the mean. That is the sampling distribution of the statistic, the thing a likelihood needs (Chapter 3): a single 512×512 map measures $\chi(1)$ to about 6%, and a measured value 25 units away from the Gaussian expectation would be a three-standard-deviation departure. The near-Gaussian shape is no accident: χ is a sum of signed counts over many weakly correlated regions of the map, and the central limit theorem (Chapter 2) applies to such sums.

How this is used in real research: Verification here, research instrument elsewhere

Everything in this section is a *verification*: the expected Euler characteristic of a Gaussian field is known exactly, and the simulations test the derivation and the code against it.

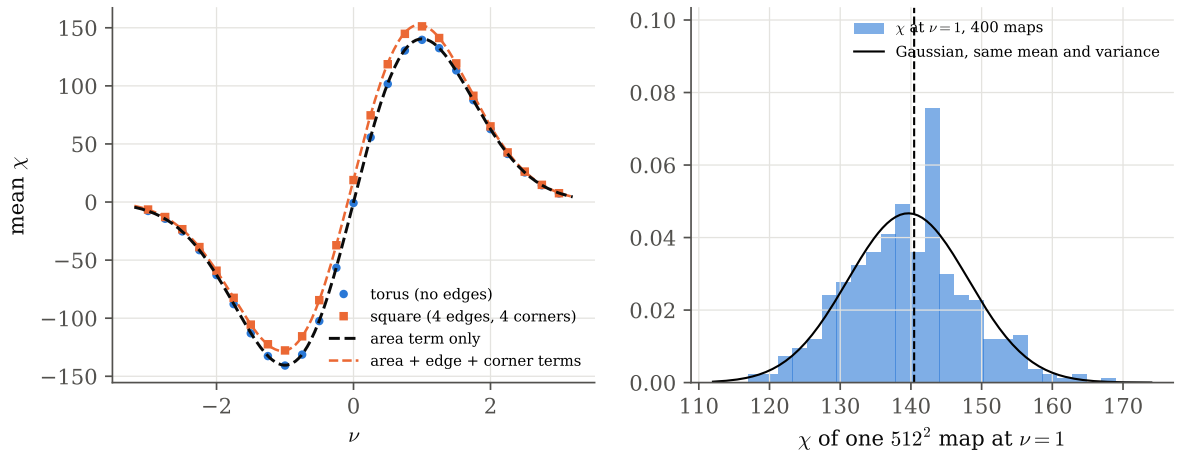


Figure 12.13: Left: mean χ of the same maps on a torus (circles) and cut into squares with four edges and four corners (squares); the black dashed line is the area term alone, the orange dashed line adds the edge and corner terms of (12.28). Right: the distribution of χ at $\nu = 1$ over 400 maps on the torus, with a Gaussian of the same mean and variance and the predicted mean (dashed).

The same machinery becomes the research instrument wherever no formula exists. The covariance of χ at different thresholds, the effect of a real mask with point-source holes, of anisotropic noise, of a non-Gaussian model, of lensing followed by a beam: none has a closed form, and analyses such as *Planck*'s compare the data with large sets of simulated skies processed by exactly the same pipeline [114]. In that use the formula (12.30) is still valuable, as the check that the pipeline reproduces the Gaussian limit before it is trusted on anything else.

12.7 Reproducing the classic results

The Gaussian formula for the Euler characteristic has been derived, tested and used for forty years, in three dimensions for the galaxy distribution, on the sphere for the microwave sky, and in a general form that also handles weak non-Gaussianity. Four papers mark the way. Each asked a different question, set its problem up in its own notation, and checked the answer against its own simulations. Each can be redone with the machinery of Section 12.5 and our own code. Their conventions (which variance, which normalisation, which units for the threshold) differ from paper to paper, and are the main source of confusion when numbers are compared, so each derivation below ends by mapping their symbols to ours.

12.7.1 Hamilton, Gott and Weinberg (1986): is the universe a sponge?

Their question. In the mid-1980s the first redshift surveys showed bubbles and walls, and three pictures of the large-scale structure competed: isolated clusters in an empty background ("meatballs"), empty bubbles in a connected background ("Swiss cheese"), or a "sponge" in which high- and low-density regions are both connected and interlock. A Gaussian random field, the prediction of inflation, is symmetric between overdense and underdense regions, so at the median density it must be a sponge. Hamilton, Gott and Weinberg proposed to measure this with the curvature of the surfaces of constant density, and derived what a Gaussian field predicts [115].

Their setup. They smooth the density contrast $\delta(\boldsymbol{r})$ with a window and take the surface $\delta = \nu\,\xi(0)^{1/2}$, where $\xi(0)$ is the variance of the smoothed field. On that surface they integrate the Gaussian curvature $K = 1/(a_1 a_2)$, with a_1, a_2 the two principal radii of curvature at a point. By the Gauss–Bonnet theorem the integral over one closed surface of genus g (the number of handles) is $\int K\,\mathrm{d}A = 4\pi(1-g)$ (their eq. 2), so a ball contributes $+4\pi$ and a doughnut 0. Their statistic is $C(\nu)$, the total curvature of all the surfaces per unit volume. They test it on eight simulations, each 5×32^3 points laid at random in a periodic box of 32^3 cells (so pure shot noise with no clustering), smoothed with $W(r) = e^{-r^2/r_0^2}$ with r_0 twice the cell size, and they measure the curvature by tiling space with cubes, the same idea as our voxel counting.

Their equation, derived. Two facts connect C to our Euler characteristic. Gauss–Bonnet for a closed surface reads $\int K\,\mathrm{d}A = 2\pi\chi(\text{surface})$, and the surface bounding a solid region has twice the Euler characteristic of the solid, $\chi(\partial X) = 2\chi(X)$ (a ball: $\chi = 1$, its sphere $\chi = 2$; see the topology notes). With $\mathbb{E}[c_3(\nu)]$ from (12.26),

$$C(\nu) \overset{(1)}{=} \frac{2\pi\,\chi(\partial X_\nu)}{V} \overset{(2)}{=} 4\pi\,\frac{\chi(X_\nu)}{V} \overset{(3)}{=} 4\pi\,\frac{\lambda^{3/2}}{(2\pi)^2}(\nu^2-1)e^{-\nu^2/2} \overset{(4)}{=} -\frac{1}{\pi}\left[\frac{-\xi''(0)}{\xi(0)}\right]^{3/2}(1-\nu^2)e^{-\nu^2/2}. \tag{12.31}$$

(1) Gauss–Bonnet, summed over all the surfaces in the volume V.
(2) The boundary of a solid has twice its Euler characteristic.
(3) Take the expectation with (12.26).
(4) $4\pi/(4\pi^2) = 1/\pi$; and for an isotropic field $-\xi''(0)/\xi(0) = \langle k^2\rangle/3 = \lambda$, the variance of one gradient component in units of the field variance (their eq. 4).

This is their eq. (3). Symbol map: their $\xi(0)$ is our σ_0^2, their $-\xi''(0)$ our $\sigma_0^2\lambda$ in three dimensions, their genus per unit volume is $-\chi/V$. The minimum of C is at $\nu = 0$, where $C = -\lambda^{3/2}/\pi$: a sponge. In their Fig. 4 they plot C in units with $\lambda = 1$, so the curve is $-\pi^{-1}(1-\nu^2)e^{-\nu^2/2}$, with minimum -0.3183 and maxima 0.142 at $\nu = \pm\sqrt{3}$.

What we compute. The script `code/ch12/05_hgw3d.py` reproduces their setup with Poisson counts of mean 5 in each of the 32^3 cells (equivalent, for this purpose, to 5×32^3 points dropped at random), subtracts the mean, smooths with the Fourier transform of their window, and counts $\chi = V-E+F-C$ of the union of closed voxels above each threshold, the three-dimensional version of (12.6); on one non-periodic block the count agrees with `skimage.measure.euler_number` (*yes*). The result is divided by $\lambda^{3/2}$ computed from the lattice modes, $\lambda = 0.251$ (the continuum value is $1/r_0^2 = 0.25$):

Listing 12.4: Hamilton, Gott and Weinberg's simulation, redone (`code/ch12/05_hgw3d.py`)

```
def smoothed_poisson(N, r0, mean_per_cell=5.0):
    """HGW's field: Poisson counts in cells, smoothed with exp(-r^2/r0^2) (FT: exp(-k^2 r0^2/4))."""
    n = rng.poisson(mean_per_cell, size=(N, N, N)).astype(float)
    kx, ky, kz = kvec(N)
    W = np.exp(-(kx ** 2 + ky ** 2 + kz ** 2) * r0 ** 2 / 4)
    return np.fft.irfftn(np.fft.rfftn(n - n.mean()) * W, s=(N, N, N), axes=(0, 1, 2))


def lam3(N, r0):
    """<k^2>/3 of the smoothed white-noise field on the lattice: sum k^2 W^2 / (3 sum W^2)."""
    k = 2 * np.pi * np.fft.fftfreq(N)
    kx, ky, kz = np.meshgrid(k, k, k, indexing="ij")
```

```
    k2 = kx ** 2 + ky ** 2 + kz ** 2
    w2 = np.exp(-k2 * r0 ** 2 / 2)
    w2.flat[0] = 0.0
    return (k2 * w2).sum() / (3 * w2.sum())


def run(N, r0, nsim):
    C = np.zeros((nsim, len(nus)))
    for s in range(nsim):
        f = smoothed_poisson(N, r0)
        u = (f - f.mean()) / f.std()
        C[s] = [4 * np.pi * euler_cubical_3d(u > nu, periodic=True) / N ** 3 for nu in nus]
    return C / lam3(N, r0) ** 1.5          # HGW's normalisation: -xi''(0)/xi(0) = <k^2>/3
    = 1
```

It runs their eight simulations, then 400 more with the same setup, then 200 boxes of 64^3 cells with $r_0 = 4$ cells, which resolve the smoothing scale twice as finely.

Side by side. Figure 12.14 overlays the three runs on their curve, and the table compares the numbers that can be read off their figure.

	formula (12.31)	HGW Fig. 4 (read)	ours, 8 sims	ours, 400 sims
$C(0)$	-0.3183	-0.32	-0.308 ± 0.006	-0.3208 ± 0.0016
$C(+1.75)$	0.142	0.15	0.127	0.1303 ± 0.0008
$C(-1.75)$	0.142	0.14		0.1475 ± 0.0008

Their eight simulations and ours agree with the formula, and with each other, within the scatter of eight boxes. Their error bars are the scatter of a single box, about 0.045 at $\nu = 0$; with eight boxes the mean is known to about 0.016, and our eight boxes give an error of the mean of 0.006.

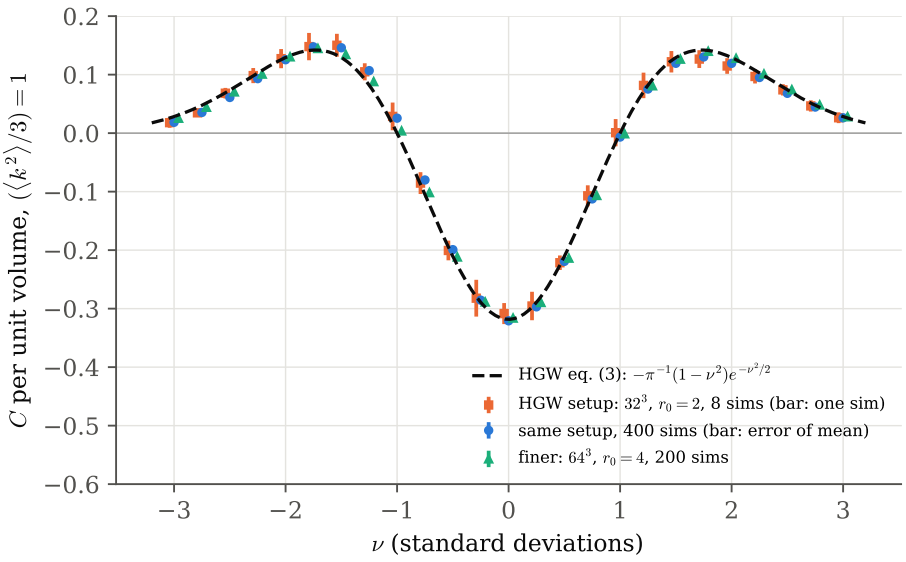


Figure 12.14: The total curvature of isodensity surfaces per unit volume, in units with $\lambda = 1$. Dashed: the Gaussian prediction (12.31). Squares: our redo of the eight simulations of Hamilton, Gott and Weinberg (bars: scatter of one simulation, as in their Fig. 4). Circles: 400 simulations of the same setup (bars: error of the mean). Triangles: 200 finer simulations.

Why they differ. With 400 boxes the statistical error is small enough to look for two effects that eight boxes hide. Hamilton, Gott and Weinberg estimated that tiling with cubes half the

smoothing length biases $C(0)$ by about 1%. At the median our coarse boxes differ from the formula by 0.8% (a positive number means more negative), 1.6 standard errors; the finer boxes differ by −0.7%, 1.1 standard errors. A standard error here is about half a per cent, so a 1% bias sits at the edge of what 400 boxes can see: our runs are consistent with their estimate, but they cannot prove that the tiling is the cause. The residual non-Gaussianity of shot noise shows elsewhere, at $\nu = \pm 1.75$: the coarse boxes give 0.1303 at +1.75 and 0.1475 at −1.75, against 0.142 for both. Shot noise is skewed (a Poisson count has a long tail to high values), so its high peaks are fewer and broader than a Gaussian field's and its low troughs more numerous; the curve is distorted asymmetrically. In the finer boxes each smoothing volume holds about eight times more points, the field is closer to Gaussian, and the asymmetry shrinks to 0.1398 and 0.1446. The size and sign of this distortion is what Matsubara's formula of Section 12.7.3 predicts from the skewness.

Lesson. Eight simulations were enough in 1986 to show that a Gaussian field is a sponge; four hundred show that "Gaussian" smoothed shot noise is not quite Gaussian, and that the genus curve is sensitive enough to notice.

12.7.2 Schmalzing and Górski (1998): Minkowski functionals on the sky

Their question. The COBE satellite had mapped the microwave sky with a 7° beam, and the question was whether the fluctuations were Gaussian. Schmalzing and Górski set out to measure all three Minkowski functionals on the sphere, directly from pixelised maps, and asked whether their estimators reproduce the Gaussian expectations on simulated skies [110].

Their setup. The sky is the unit sphere, the field is the temperature u in mK, and the excursion set is $Q_\nu = \{u > \nu\}$ with the threshold ν in mK, not in units of σ. Their power spectrum is built to look like COBE's:

ingredient	value (their eqs. 17–20)
primordial spectrum	Harrison–Zel'dovich, $n = 1$: $C_\ell \propto 1/\ell(\ell+1)$, with $C_{10} = (7\,\mu\mathrm{K})^2$
beam	Gaussian, 7° FWHM: $g_\ell = e^{-\sigma_b^2 \ell(\ell+1)/2}$, $\sigma_b = 7^\circ/\sqrt{8\ln 2}$
noise	white, $C_\ell^{\mathrm{noise}} = (3\,\mu\mathrm{K})^2$
extra smoothing	Gaussian, $e^{-\omega^2\ell(\ell+1)/2}$ with $\omega = 3^\circ$
simulations	1000 skies on the 6144 pixels of the COBE map

They call the width of the extra smoothing "3°" without saying whether it is ω or a FWHM. Their Fig. 1 settles it: the smoothed noise spectrum $\ell(\ell+1)(3\,\mu\mathrm{K})^2 e^{-\omega^2\ell(\ell+1)}$ peaks where $\ell(\ell+1) = 1/\omega^2$, at height $(3\,\mu\mathrm{K})^2/(e\,\omega^2)$. For $\omega = 3^\circ = 0.0524$ rad this is $\ell \approx 18.6$ and $1.2 \times 10^{-3}\,\mathrm{mK}^2$, which is where their dotted curve peaks.

Their equation, derived. Their eq. (14) gives the expected functionals in terms of three numbers measured from the map: the mean μ, the variance (which they call σ, not σ^2) and $\tau = \frac{1}{2}\langle u_{;i} u_{;i}\rangle$, half the mean squared gradient. In our language τ/σ is exactly $\lambda = \sigma_1^2/2\sigma_0^2$, and their threshold in standard units is $x = (\nu-\mu)/\sqrt{\sigma}$. Their Euler-characteristic density is

$$v_2(\nu) \overset{(1)}{=} \frac{\tau}{2\pi^{3/2}\sigma}\,\frac{\nu-\mu}{\sqrt{2\sigma}}\,\exp\Big(-\frac{(\nu-\mu)^2}{2\sigma}\Big) \overset{(2)}{=} \frac{\lambda}{2^{3/2}\pi^{3/2}}\,x\,e^{-x^2/2} \overset{(3)}{=} \frac{\lambda}{(2\pi)^{3/2}}\,x\,e^{-x^2/2}, \tag{12.32}$$

(1) Their eq. (14), third line.
(2) $\tau/\sigma = \lambda$ and $(\nu-\mu)/\sqrt{2\sigma} = x/\sqrt{2}$.

(3) $2^{3/2}\pi^{3/2} = (2\pi)^{3/2}$.

which is our (12.23). Their $v_1 = \frac{\sqrt{\tau}}{8\sqrt{\sigma}}e^{-x^2/2} = \frac{\sqrt{\lambda}}{8}e^{-x^2/2}$ is a quarter of our length density (12.15), and v_0 is $\Psi(x)$. Their eq. (7), $\chi = V_2 + V_0/(2\pi R^2)$, is the Gauss–Bonnet split of Section 12.4.2: on the unit sphere with $v_j = V_j/4\pi$, $\chi = 4\pi v_2 + 2v_0$, which is (12.29).

A misprint. Their eq. (16) gives $\sigma = \sum_\ell (2\ell+1)C_\ell$ and $\tau = \sum_\ell (2\ell+1)C_\ell\,\ell(\ell+1)/2$. Both sums lack the factor $1/4\pi$ of the addition theorem: the variance of a map is $\sum_\ell \frac{2\ell+1}{4\pi}C_\ell$ (T3.3). The slip cancels in $\lambda = \tau/\sigma$, so the shapes and heights of v_1, v_2 and χ are unaffected. But it makes the standard deviation $\sqrt{4\pi} \approx 3.5$ times too large (145.3 instead of 41 μK for their spectrum) and would stretch every curve plotted against the threshold in mK by that factor. Their Fig. 2 runs from -0.25 to $+0.25$ mK and its curves have died out by about ± 0.13 mK, three times 41 μK: they used the correct variance and only the printed equation is wrong.

What we compute. The script `code/ch12/06_sphere_sg98.py` builds their spectrum,

Listing 12.5: The COBE-like spectrum of Schmalzing and Górski (`code/ch12/06_sphere_sg98.py`)

```
# ------------------------------------------------------------------ SG98's spectrum, eqs. (17)-(20), in mK^2
ell = np.arange(LMAX + 1)
cl_hz = np.zeros(LMAX + 1)
cl_hz[2:] = 6.0 / (ell[2:] * (ell[2:] + 1))          # eq. (17) with n = 1: Gamma ratios -> 6 / l(l+1)
cl_hz *= (7e-3) ** 2 / cl_hz[10]                      # normalised to C_10 = (7 microK)^2
sig_b = np.radians(7.0) / np.sqrt(8 * np.log(2))      # 7 deg FWHM beam
g_beam = np.exp(-0.5 * sig_b ** 2 * ell * (ell + 1))
g_smooth = np.exp(-0.5 * np.radians(3.0) ** 2 * ell * (ell + 1))   # eq. (19), omega = 3 deg
cl_noise = (3e-3) ** 2 * np.ones(LMAX + 1)            # eq. (18)
cl_noise[:2] = 0.0
cl = g_smooth ** 2 * (cl_hz * g_beam ** 2 + cl_noise)  # eq. (20)

s2, tau = sigma_tau(cl)                               # correct eq. (16): sums divided by 4 pi
lam = tau / s2
s2_bad, tau_bad = 4 * np.pi * s2, 4 * np.pi * tau     # eq. (16) as printed
sig = np.sqrt(s2)
```

draws 1000 skies at HEALPix $N_{\rm side} = 128$ (196608 pixels of 0.46°) and again at $N_{\rm side} = 32$ (12288 pixels of 1.8°, close to the 2.6° COBE pixels), and measures the functionals with the library `lib_sphere.py`: area by counting pixels; length by the band argument of (12.15) with the gradient computed from the $a_{\ell m}$; χ by $V - E + F$ of the HEALPix pixel mesh (which pixels share each corner and each edge is worked out once from the pixel corners). As a check, χ is also computed as pieces of the hot set (pixels touching at an edge or a corner) minus pieces of the cold set (pixels sharing an edge) plus one, the spherical form of Alexander duality: β_1 of the hot set is the number of cold pieces minus one, because on the sphere one cold piece is the "outside". In 4100 comparisons the two counts disagree 0 times. For this spectrum $\lambda = 51.4\,\mathrm{sr}^{-1}$, a correlation angle $\theta_c = 1/\sqrt{2\lambda} = 5.65°$.

Side by side.

	formula	SG98 Fig. 2 (read)	ours, $N_{\text{side}} = 128$
peak of v_1 (at $\nu = 0$)	0.896	≈ 0.9	0.899
peak of v_2 [sr^{-1}]	1.98	≈ 2.1	2
χ at its peak ($\nu \approx \sigma$)	25.1	≈ 26	25.5 ± 5.9 (one sky)
threshold of the χ peak [mK]	0.041	≈ 0.04	0.043
χ at $\nu = -3\sigma$	0.63	$\to 2$ at low ν	0.64

Figure 12.15 is our version of their Fig. 2. The mean curves of our skies lie on the formula, the values read from their figure sit on top of both, and the full-sky χ rises to the sphere's own value 2 at low thresholds, the $2v_0$ term without which the curve would go to zero (dotted).

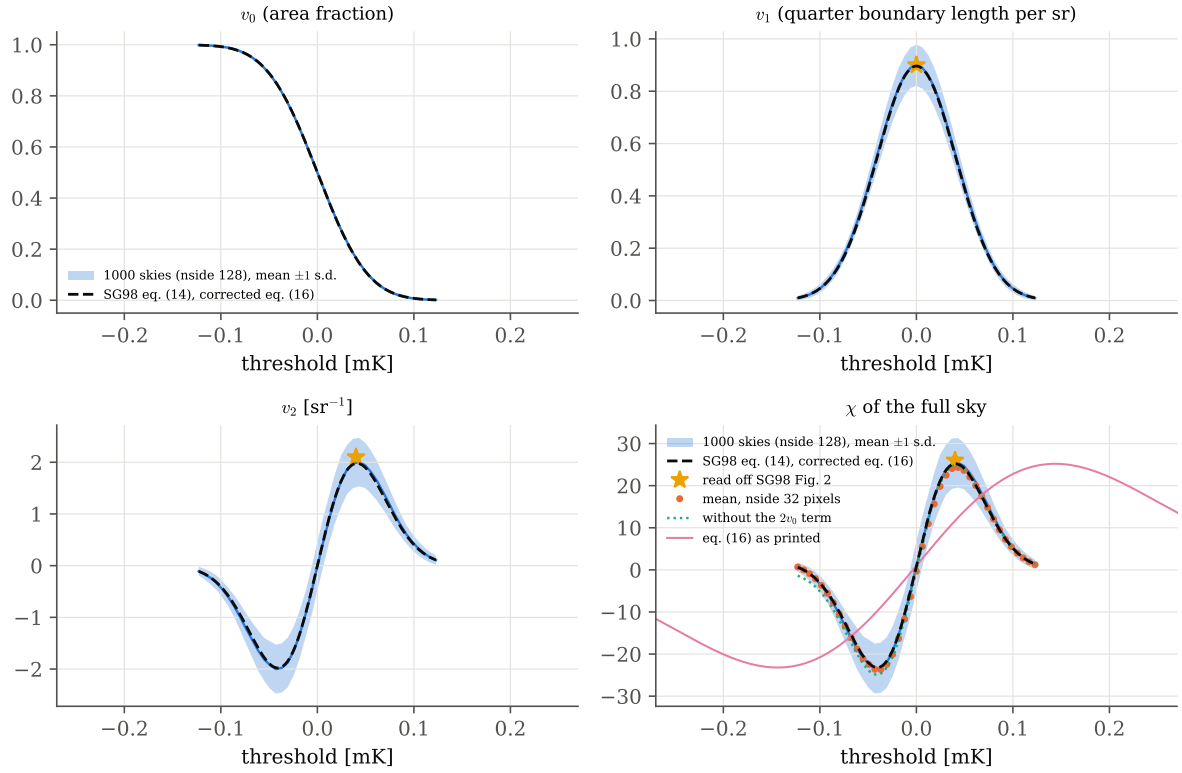


Figure 12.15: The Minkowski functionals of 1000 COBE-like Gaussian skies against threshold in mK, as in Schmalzing and Górski's Fig. 2. Band: mean ± one standard deviation of a single sky ($N_{\text{side}} = 128$). Dashed: their eq. (14) with the variance computed correctly. Stars: values read off their figure. In the χ panel, orange dots are the mean at $N_{\text{side}} = 32$, the dotted line omits the $2v_0$ term, and the pink line shows what the printed eq. (16) would give.

Why they differ. Three effects, each small. First, pixel size: at $N_{\text{side}} = 32$ the pixels are a third of the correlation angle, and counting χ on them loses $4.5 \pm 0.2\%$ at the peak (24.3 instead of 25.5, on the same skies), because small spots and narrow necks are not resolved. Schmalzing and Górski avoided this by not counting at all: their estimator integrates the curvature formula of their eq. (12) over pixels, using derivatives of the field, which does not care whether a spot is resolved by the grid. Second, our thresholds are in units of each sky's own mean and standard deviation, as theirs are (their eq. 15). Those two numbers fluctuate from sky to sky, because a sky with a 5.65° correlation angle has only a few hundred independent spots. Each sky's threshold scale is therefore shifted and stretched a little at random, and averaging the curve over those random rescalings bends it slightly where it is most curved. The largest deviation of our 1000-sky mean from the formula is 6.5 Monte Carlo errors, but only 0.79 in χ, at $\nu = 1.8$,

against a scatter of several units for one sky; it would disappear with thresholds fixed in units of the ensemble standard deviation, which a real sky does not offer. Third, the numbers read from their figure are good to about five per cent.

Lesson. The formula carries over to the sphere with one extra term, the sphere's own Euler characteristic times the area fraction; and a misprinted normalisation that cancels in every shape still matters as soon as a curve is plotted against a physical threshold.

12.7.3 Matsubara (2003): one formula for every functional

His question. The Gaussian predictions are only the starting point: the late universe is not Gaussian, and the question is how the Minkowski functionals change when it departs slightly. Matsubara derived the Gaussian functionals and their first non-Gaussian corrections, in any dimension and for every functional, in a single expression [113].

His equation, and ours. For a field with variance σ_0^2 and gradient variance σ_1^2 in d dimensions, his eq. (3.54) reads, at zeroth order in the non-Gaussianity,

$$V_k^{(d)}(\nu) = \frac{1}{(2\pi)^{(k+1)/2}} \frac{\omega_d}{\omega_{d-k}\,\omega_k} \left(\frac{\sigma_1}{\sqrt{d}\,\sigma_0}\right)^k e^{-\nu^2/2} H_{k-1}(\nu), \qquad \omega_k = \frac{\pi^{k/2}}{\Gamma(k/2+1)}, \tag{12.33}$$

where ω_k is the volume of the unit ball in k dimensions ($\omega_0 = 1$, $\omega_1 = 2$, $\omega_2 = \pi$, $\omega_3 = 4\pi/3$) and H_n are the Hermite polynomials $H_0 = 1$, $H_1 = \nu$, $H_2 = \nu^2 - 1$, with $e^{-\nu^2/2} H_{-1}(\nu) = \int_\nu^\infty e^{-t^2/2}\mathrm{d}t$. Symbol map: $\sigma_1/(\sqrt{d}\,\sigma_0) = \sqrt{\lambda}$, because σ_1^2/σ_0^2 is the sum of the d equal gradient-component variances. Then, in the plane,

$$V_2^{(2)} \overset{(1)}{=} \frac{1}{(2\pi)^{3/2}} \frac{\pi}{\omega_0\,\pi} \lambda\, e^{-\nu^2/2}\nu = \frac{\lambda}{(2\pi)^{3/2}}\, \nu e^{-\nu^2/2}, \qquad V_1^{(2)} \overset{(2)}{=} \frac{1}{2\pi} \frac{\pi}{2\cdot 2} \sqrt{\lambda}\, e^{-\nu^2/2} = \frac{\sqrt{\lambda}}{8} e^{-\nu^2/2},$$
$$V_0^{(2)} \overset{(3)}{=} \frac{1}{\sqrt{2\pi}} \frac{\pi}{\pi\cdot 1} \int_\nu^\infty e^{-t^2/2}\mathrm{d}t = \Psi(\nu), \qquad V_3^{(3)} \overset{(4)}{=} \frac{1}{(2\pi)^2} \frac{4\pi/3}{1\cdot 4\pi/3} \lambda^{3/2} e^{-\nu^2/2}(\nu^2 - 1). \tag{12.34}$$

(1) $k = d = 2$: $\omega_2/(\omega_0\omega_2) = 1$, $H_1 = \nu$. This is the Euler characteristic density (12.23); his eq. (3.18) for the two-dimensional genus is the same expression.
(2) $k = 1$, $d = 2$: $\omega_2/(\omega_1\omega_1) = \pi/4$, $H_0 = 1$. This is a quarter of the length density (12.15), the normalisation used for v_1 on the sky.
(3) $k = 0$: the area fraction (12.14).
(4) $k = d = 3$: the three-dimensional Euler characteristic density (12.26), ρ_3 of (12.30); minus it is his three-dimensional genus.

Every Gaussian result of this part is one line of (12.33), and every Monte Carlo of Sections 12.6, 12.7.1 and 12.7.2 is therefore also a check of it: in two dimensions to a few tenths of a per cent (Section 12.6), in three to about one per cent (Section 12.7.1), on the sphere to the accuracy of the pixel counting (Section 12.7.2).

The non-Gaussian part. Matsubara's formula continues with a term proportional to σ_0 and to three skewness parameters $S^{(0)}$, $S^{(1)}$, $S^{(2)}$, averages of u^3, $u^2\nabla^2 u$ and $|\nabla u|^2\nabla^2 u$ [113]; Section T3.2.1 wrote it out for the sky and tested it on skewed maps. Two features matter here. The correction is built from the Hermite polynomials H_{k+2}, H_k and H_{k-2}, which are odd where the Gaussian term is even and vice versa, so it distorts the curves asymmetrically, as the shot noise did in Section 12.7.1. And when the threshold is labelled by area fraction (Section 12.3.2)

the term with $S^{(0)}$, the one-point skewness, disappears and only the differences $S^{(1)} - S^{(0)}$ and $S^{(2)} - S^{(0)}$ survive (his eq. 3.54 in the variable ν_A). That is the perturbative version of the exact statement of Section 12.3.2: relabelling removes what is in the histogram of the values and keeps what is in their arrangement.

Lesson. The four Gaussian densities are one formula. What the field contributes is a single number per dimension, $\lambda^{k/2}$; everything else is geometry (ω_k) and the Hermite polynomials.

12.7.4 *Planck* 2018: are the microwave sky's shapes Gaussian?

Their question. Section T3.5.1 described how the *Planck* collaboration used the three functionals, at many resolutions, to test whether the temperature and polarization maps are consistent with Gaussian, isotropic skies [114]. Here we reproduce the Gaussian side of that test, the curves against which the data are compared (their Fig. 10), and work out in which normalisation those curves are drawn.

Their setup. The four component-separated maps are degraded to a chosen N_{side} and smoothed with a Gaussian beam whose FWHM grows with the pixel size (40′ at $N_{\text{side}} = 256$, their Table 1). A common mask leaves 75% of the sky. The thresholds are 12 values from -3 to 3 in units of the map's own standard deviation σ_0. At each threshold they compute the area fraction, the boundary length and the genus (their eqs. 13–15), divide out the part of each that depends on the power spectrum to get what they call "normalised" functionals, and compare the 36 numbers with 999 FFP10 simulations, which contain the CMB, the noise and the component-separation pipeline, through a χ^2 with the simulation covariance (their eqs. 20–22).

Their equations, derived. Their eqs. (16)–(19) are (12.33) with $d = 2$: $V_k = A_k v_k$, with $v_k = e^{-\nu^2/2} H_{k-1}(\nu)$ and $A_k = \frac{1}{(2\pi)^{(k+1)/2}} \frac{\omega_2}{\omega_{2-k}\omega_k} \left(\frac{\sigma_1}{\sqrt{2}\sigma_0}\right)^k$. All the dependence on the spectrum is in the factor $(\sigma_1/\sigma_0)^k$. Its value for their maps follows from the sums (T3.3) and (T3.4) with the fiducial spectrum, the 40′ beam and the $N_{\text{side}} = 256$ pixel window: $\sigma_1/\sigma_0 = 108.4$ per radian, a correlation angle $\theta_c = 31.7'$, compared with a pixel of 13.7′. The genus amplitude per steradian is then $\lambda/(2\pi)^{3/2} \approx 108.4^2/(2 \times 15.75) \approx 370$.

The normalisation of their figure. Their Fig. 10 shows the boundary length peaking at about 0.081 and the genus between about -0.020 and $+0.021$, at every resolution from $N_{\text{side}} = 32$ to 1024. These cannot be the functionals per steradian (the genus amplitude above is in the hundreds and grows with resolution), nor the normalised v_k of their eq. (17) (which peak at 1 and $e^{-1/2} = 0.61$). They are reproduced exactly by $V_k/(\sigma_1/\sigma_0)^k$, the Gaussian functionals with the spectral factor divided out but the numerical prefactors of A_k kept:

$$\begin{aligned} \frac{V_1}{\sigma_1/\sigma_0} &\overset{(1)}{=} \frac{1}{2\pi}\,\frac{\pi}{2\cdot 2}\,\frac{1}{\sqrt{2}}\,e^{-\nu^2/2} = \frac{e^{-\nu^2/2}}{8\sqrt{2}} \quad\Rightarrow\quad \text{peak } 0.0884, \\ \frac{V_2}{(\sigma_1/\sigma_0)^2} &\overset{(2)}{=} \frac{1}{(2\pi)^{3/2}}\,\frac{1}{2}\,\nu\,e^{-\nu^2/2} \quad\Rightarrow\quad \text{extremes } \pm\frac{e^{-1/2}}{2(2\pi)^{3/2}} = \pm 0.0193. \end{aligned} \tag{12.35}$$

(1) A_1 with σ_1/σ_0 set to one: $\omega_2/(\omega_1\omega_1) = \pi/4$ and $(1/\sqrt{2})^1$.
(2) A_2 with σ_1/σ_0 set to one: $\omega_2/(\omega_0\omega_2) = 1$ and $(1/\sqrt{2})^2 = \frac{1}{2}$; $\nu e^{-\nu^2/2}$ is extreme at $\nu = \pm 1$.

At their nearest thresholds to the extremes, $\nu = \pm 0.27$ for the length and ± 0.82 for the genus, these give 0.0852 and ± 0.0186. Why divide out only $(\sigma_1/\sigma_0)^k$? Because it is the one factor that

changes with resolution (the beam sets σ_1) and from map to map (cosmic variance); dividing it out makes the curves at different $N_{\rm side}$ comparable and removes most of the fluctuation of the overall amplitude, leaving the shape, which is where non-Gaussianity would show. Their eq. (15), read literally, would add a further $1/2\pi$ to the genus; the plotted curves do not have it.

What we compute. The script `code/ch12/07_planck_mf.py` draws 300 full-sky Gaussian maps with the fiducial spectrum (Chapter T1), the 40′ beam and the pixel window at $N_{\rm side} = 256$, $\ell \le 767$. For each map it measures σ_0 and σ_1 (from the exact derivative maps) and the functionals at their 12 thresholds and on a finer grid:

Listing 12.6: Planck-style normalised functionals of one map (`code/ch12/07_planck_mf.py`)

```
def measure(alm, nside, nus, msh, lmax, dnu=0.1):
    """V0, V1/(s1/s0), V2/(s1/s0)^2 per steradian, and s1/s0 of the map (rad^-1)."""
    t, dth, dph = hp.alm2map_der1(alm, nside, lmax=lmax)
    t = t - t.mean()
    s0 = t.std()
    g = np.hypot(dth, dph)
    r = np.sqrt((g ** 2).mean()) / s0                  # sigma1 / sigma0, measured on this map
    u, gu = t / s0, g / s0
    v0 = np.array([(u > nu).mean() for nu in nus])
    v1 = np.array([0.25 * (gu * (np.abs(u - nu) < dnu / 2)).mean() / dnu for nu in nus])
    chi = np.array([chi_sphere(u > nu, msh) for nu in nus])
    v2 = (chi - 2 * v0) / (4 * np.pi)                  # remove the sphere's own curvature term
    return v0, v1 / r, v2 / r ** 2, r
```

The measured σ_1/σ_0 is 108.1 ± 2.9 per radian against 108.4 from the spectrum. Then, as *Planck* does, it forms the 36-number vector of each map, estimates the covariance from the other $300-1$ maps (with the Hartlap factor 0.876 [6]), and computes each map's χ^2.

Side by side.

	formula (12.35)	*Planck* Fig. 10 (read)	ours, mean	ours, one map
length at $\nu = 0.27$	0.0852	0.081	0.0851	$\pm 8.000 \times 10^{-4}$
genus at $\nu = 0.82$	0.0186	0.021	0.0173	$\pm 4.000 \times 10^{-4}$
$\chi^2/N_{\rm dof}$, Gaussian sky	1	peak near 1	1.01 (median 0.97)	

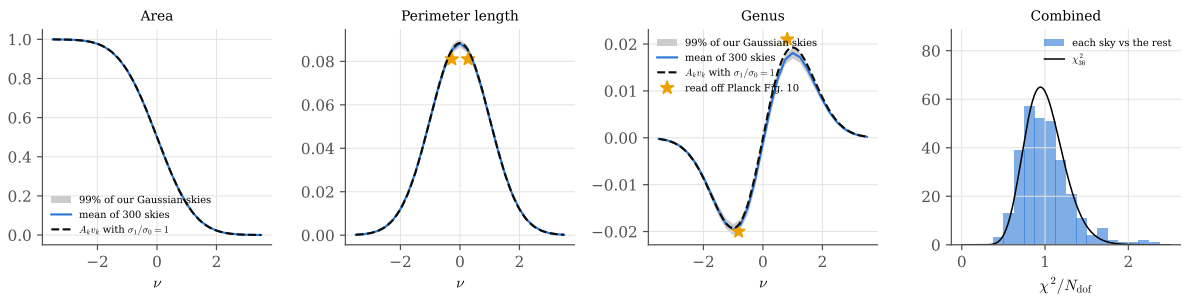


Figure 12.16: Our version of the $N_{\rm side} = 256$ row of *Planck*'s Fig. 10. Grey: the central 99% of 300 Gaussian skies; blue: their mean; dashed: (12.35); stars: values read off the published figure. Right: $\chi^2/N_{\rm dof}$ of each sky against the other skies, using all 36 numbers, with the χ^2_{36} distribution.

Why they differ. The length agrees with the formula to the third digit and with the published curve to within the accuracy of reading it. The genus of our maps is 0.0173, about 7% below the formula. That is the pixel effect of Section 12.2.2, now large because the correlation angle is only 2.3 pixels. Its signature is clearest at the median: the full-sky χ at $\nu = 0$ should be $2\Psi(0) = 1$, and our skies give -172 ± 6. Near $\nu = 0$ the expected χ rises with slope $4\pi\lambda/(2\pi)^{3/2} = 4688$ per unit of ν, so this is the whole curve shifted by 0.037 standard deviations towards higher thresholds. The 8-neighbour rule joins hot pixels that touch at a corner, which makes the hot set behave as if the threshold were slightly lower at every saddle that falls inside a 2×2 block of pixels; the effect is as large as the fraction of saddles that are not resolved by the grid. The same skies sampled at $N_{\text{side}} = 512$ give a genus of 0.0183 instead of 0.0174 at $\nu = 0.82$, a deficit of 4.8% that finer pixels mostly remove. The published curve, read at ≈ 0.021, lies about 10% above the formula. We cannot settle the reason from the paper: their genus is computed from counts of hot and cold spots on the masked sky with their own pixel treatment, and the reading of a small figure is uncertain at this level. It does not matter for their test, and this is the point of their method: data and simulations pass through the identical estimator, so any such offset is common to both and cancels in the comparison. Finally, the χ^2 of Gaussian skies against Gaussian skies follows the χ^2_{36} distribution, mean 1.01 per degree of freedom, which is what the histograms of their right-hand column show for the FFP10 simulations.

Lesson. A published curve is only reproducible once its normalisation is reverse-engineered; here the curves are the Gaussian functionals with σ_1/σ_0 divided out. And the reason the *Planck* test is robust to such details is that it never compares the data with a formula, only with simulations processed the same way.

12.8 How it is done in research

The calculations so far used clean, full-sky or periodic Gaussian maps, because that is where the answer is known and the code can be checked. A real analysis of the microwave sky starts from maps that have been through an instrument and a foreground-cleaning pipeline, covers only part of the sky, and asks whether the measured curves are consistent with a model whose only complete description is a set of simulations. The pipeline of the *Planck* temperature and polarization analysis [114] runs as follows, each step with its purpose.

The data products. Four maps of the CMB, each produced by a different component-separation method (Commander, NILC, SEVEM, SMICA), at $N_{\text{side}} = 2048$; for each, a common mask of the Galaxy and point sources; and 999 FFP10 simulations, which are Gaussian CMB skies from the best-fit spectrum, lensed, observed with the real scanning, beams and noise, and passed through the same component separation. The simulations are the null hypothesis: a sky that is Gaussian and isotropic apart from everything the instrument and the pipeline do to it.

The pipeline.

1. *Choose a resolution.* Degrade each map to N_{side} between 16 and 2048 and smooth it with a Gaussian beam whose FWHM is about three pixels (their Table 1). The smoothing makes the field smooth on the pixel scale, which the counting needs (Section 12.2.2), and fixes σ_1/σ_0. Each resolution probes a different range of scales.
2. *Apply the mask.* Only unmasked pixels enter. The mask's boundary adds pieces to the excursion set and changes χ (Section 12.5.5); because the simulations are masked identically, the effect is carried by the null distribution rather than modelled.

3. *Set the thresholds.* Twelve thresholds from -3 to 3, in units of the standard deviation measured on the unmasked part of the map itself.
4. *Measure.* The area fraction by counting pixels, the boundary length from the contour lengths, the genus from the numbers of hot and cold spots; then divide by $(\sigma_1/\sigma_0)^k$ (Section 12.7.4).
5. *Repeat on every simulation*, with the identical code, to obtain the sampling distribution of the 36 numbers.
6. *Compare.* Form the χ^2 of the data vector with the mean and covariance of the simulations; the p-value is the fraction of simulations with a larger χ^2 (their Table 5). Repeat for each resolution, for each component-separation method, and for E-mode polarization.

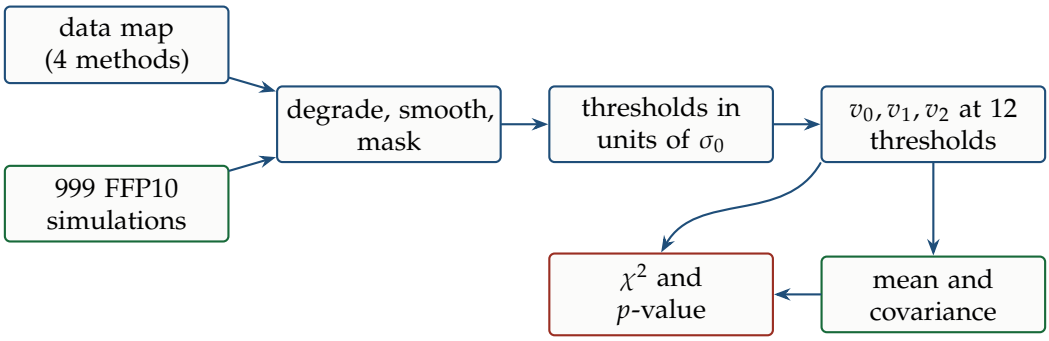


Figure 12.17: The Minkowski-functional test of a sky map. Data and simulations follow the same path; the simulations (green) provide the mean and covariance of the 36 numbers, against which the data vector is compared through a χ^2 (red).

The checks. Each step has a failure mode, and a real analysis tests for it before looking at the data. The Gaussian limit: on Gaussian simulations without mask the pipeline must reproduce the formulas of Section 12.5, which is what Sections 12.6, 12.7.2 and 12.7.4 did. Convergence: the covariance of 36 numbers from 999 simulations is well determined, but its inverse is biased by the factor $(n-p-2)/(n-1)$ [6], 0.96 for $n = 999$ and $p = 36$; with fewer simulations, or more thresholds, the correction grows quickly. Resolution and convention: the counts must be stable when the pixel size is halved at fixed smoothing (our genus moved by 4.8% at $N_{\rm side} = 256$, Section 12.7.4), and the connectivity convention must be the same for data and simulations. The mask: the counts near the mask boundary are the most fragile, and their effect is checked by changing the mask. Agreement between the four component-separation methods is itself a test of foreground residuals.

12.8.1 What a mask does, and how the formula absorbs it

The general formula (12.30) says that a mask enters only through three numbers of the unmasked region: its Euler characteristic $\mathcal{L}_0$, half its boundary length $\mathcal{L}_1$ and its area $\mathcal{L}_2$. A tempting shortcut, "the masked sky is a fraction $f_{\rm sky}$ of the full sky, so multiply the full-sky prediction by $f_{\rm sky}$", misses the boundary term entirely. How much does that matter?

Example: a Galactic band. The script `code/ch12/08_mask_gkf.py` masks the band $|b| < 20^\circ$ around the Galactic equator, leaving two polar caps of angular radius 70° ($f_{\rm sky} = 65.8\%$). For two caps, $\mathcal{L}_0 = 2$ (two discs), $\mathcal{L}_1$ is half of the two boundary circles, $\tfrac{1}{2} \cdot 2 \cdot 2\pi \sin 70^\circ = 5.9$, and $\mathcal{L}_2$ is their area, $2 \cdot 2\pi(1 - \cos 70^\circ) = 8.27\,\mathrm{sr}$. On 300 Gaussian skies with the fiducial spectrum, a 160′ beam and $N_{\rm side} = 256$ ($\lambda = 306\,\mathrm{sr}^{-1}$, a correlation angle of about nine pixels, so that the

pixel shift found in Section 12.7.4 is negligible), standardised on the unmasked pixels, the mean χ of the unmasked excursion set at $\nu = 0$ is 14.5 ± 0.9. The formula with the three $\mathcal{L}_j$ predicts 17.4, of which the boundary term $\mathcal{L}_1\rho_1(0)$ is 16.4; the f_{sky} shortcut predicts 0.66. At $\nu = 1$ the skies give 109, the formula 108, the shortcut 97 (Figure 12.18). The largest deviation of the simulated mean from the formula, over all thresholds, is 3.7 Monte Carlo errors, a few per cent of the peak value of χ: the residual of the pixel counting and of standardising each sky on its own unmasked pixels. So the boundary of a simple mask adds a known, sizeable amount to χ, and the formula accounts for it exactly; a real mask, with hundreds of point-source holes each adding its own boundary, makes $\mathcal{L}_1$ large and is one reason real analyses rely on simulations masked in the same way.

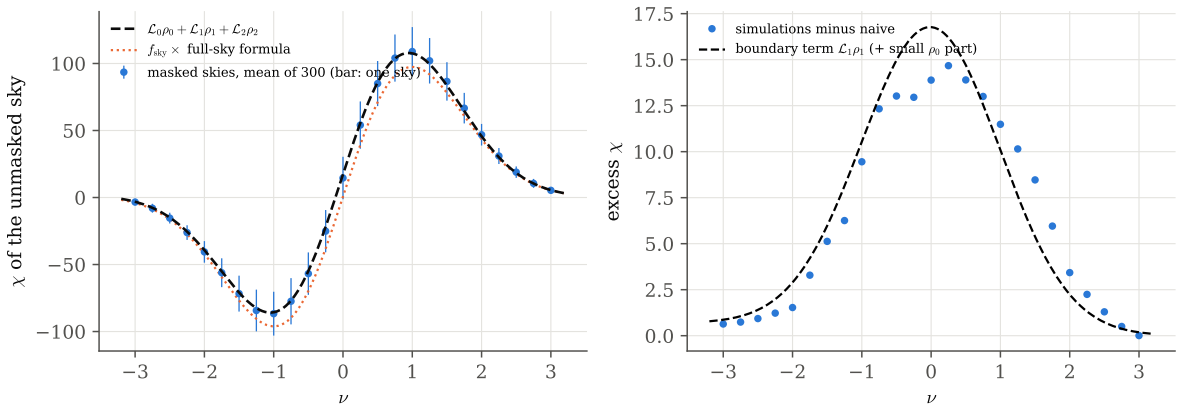


Figure 12.18: Euler characteristic of the sky outside a $|b| < 20°$ band. Left: the mean over 300 Gaussian skies (bars: one sky), the formula with the Lipschitz–Killing numbers of the two caps (dashed) and the full-sky formula scaled by f_{sky} (dotted). Right: the difference between the simulations and the f_{sky} shortcut is the boundary term.

Laptop scale, and what it costs. Every computation in this part was reduced to run in minutes on a laptop. Each reduction, and what the full analysis does instead:

- *Planck* analyses maps up to $N_{\text{side}} = 2048$ with 999 end-to-end simulations including noise, lensing and component separation; we used 300 Gaussian skies at $N_{\text{side}} = 256$ without noise, which reproduces the Gaussian curves and the χ^2 distribution but cannot test foreground residuals or noise inhomogeneity.
- *Planck* masks 25% of the sky with a mask full of point-source holes; we used the full sky for the reproduction and a simple band for the mask test, which shows the boundary term but understates its size for a real mask.
- Our flat maps are 512×512 pixels with 400 realisations per spectrum; this resolves the mean χ to about 0.3%, enough to see the edge term and the lattice value of λ, not enough to study the far tails beyond $|\nu| \approx 3$.
- Our three-dimensional boxes are 32^3 and 64^3 cells, where cosmic-web studies use 512^3 or more; the small boxes reproduce the Gaussian genus curve but resolve only a few correlation lengths per side, so the residual shot-noise skewness is larger than in a large survey.

Recap

- A nuisance is an operator $\phi \to \mathcal{N}[\phi]$; a statistic T is protected if $T[\mathcal{N}[\phi]] = T[\phi]$. Counts of pieces and holes are protected against smooth one-to-one remappings

(lensing, to first order) and monotone changes of the values; not against noise, smoothing, tearing, or a feature crossing a mask edge.

- The topology of a field is the topology of its excursion sets $X_\nu = \{\phi > \nu\}$. In two dimensions β_0 counts pieces, β_1 holes, and $\chi = \beta_0 - \beta_1$.
- On pixels, use dual connectivity (8 for the set, 4 for its complement, or the reverse); $\chi = V - E + F$ of the closed-pixel complex equals $\beta_0 - \beta_1$; labelling, cell counting, `scikit-image` and `gudhi` agree exactly.
- Under a monotone transform the excursion sets are the same sets with relabelled thresholds; labelling thresholds by area fraction removes all pointwise (one-point) non-Gaussianity and keeps only spatial information.
- The Minkowski functionals (area, length, χ) are the complete set of additive, motion-invariant shape measures; χ is the total turning of the boundary over 2π, plus enclosed curvature on a curved sky.
- Counting critical points with the sign of $\det H$ gives, for a Gaussian field, $\mathbb{E}[\chi]/A = \lambda\nu e^{-\nu^2/2}/(2\pi)^{3/2}$ with $\lambda = \sigma_1^2/2\sigma_0^2$; in 3-D $\lambda^{3/2}(\nu^2-1)e^{-\nu^2/2}/(2\pi)^2$; edges, masks and the sphere enter through $\mathbb{E}\,\chi = \sum_j \mathcal{L}_j \rho_j$.
- For a Gaussian field the mean χ curve knows only λ; its covariance knows more. A departure from the shape $\nu e^{-\nu^2/2}$ is a model-independent sign of non-Gaussianity.
- Reproduced: Hamilton–Gott–Weinberg's sponge (to 1%), Schmalzing–Górski's sky functionals (and their misprinted eq. 16), Matsubara's general formula, and *Planck*'s Gaussian curves, whose plotted normalisation is $V_k/(\sigma_1/\sigma_0)^k$.

12.9 One threshold is not enough: births, deaths and persistence

The threshold scan of Section 12.3 counted the pieces and holes of the excursion set of one Gaussian map at one threshold after another. At $\nu = 1$ it found a certain number of pieces, and that number is all the scan remembers about $\nu = 1$. But the pieces are not all alike. One may be the top of a broad mountain that stands far above the water; another may be a pebble that pokes out by a hair and would be gone at $\nu = 1.05$. The count gives both the same weight, and a physical conclusion drawn from it can change when we move the threshold by a small amount that nobody could defend as special. The cure is to stop treating the thresholds one at a time. We follow every piece and every hole through all thresholds, and record the level at which it appears and the level at which it disappears. A pebble then shows up as something that lived over a tiny range of levels, and a mountain as something that lived over a wide range, and the difference is no longer hidden.

The picture to keep in mind is the flooded landscape of Section 12.3, run in one direction only. The field $\phi(\boldsymbol{x})$ is the height of the ground at the point $\boldsymbol{x}$, measured as before in units of its standard deviation. The water level is ν. The dry land at level ν is the set of points with height at least ν,

$$X_\nu = \{\boldsymbol{x} :\ \phi(\boldsymbol{x}) \geq \nu\}, \tag{12.36}$$

the excursion set of Eq. (12.4) with the boundary points included (for a continuous field the two versions differ only at levels equal to the height of some pixel, and that changes no count we will make). We start with the water above the highest summit, so nothing is dry, and we let the water drain slowly away. The dry land only ever grows: if a point is dry at level ν, it is dry

at every lower level, so $X_\nu \subseteq X_{\nu'}$ whenever $\nu \geq \nu'$.

The question. As the water level drops from above the highest peak to below the deepest valley, at which level does each island first appear, and at which level does it stop being a separate island? At which level does each lake (a pool of water completely surrounded by land) appear, and at which level does it drain away?

12.9.1 The simplest landscape: a line

Take a one-dimensional landscape, nine heights at equally spaced points joined by straight lines,

$$(h_0, \ldots, h_8) = (0.2,\ 1.8,\ 0.9,\ 2.6,\ 0.5,\ 1.4,\ 1.1,\ 2.0,\ 0.3), \tag{12.37}$$

drawn in Figure 12.19. It has four summits, at heights 2.6 (call it A), 2.0 (D), 1.8 (B) and 1.4 (C), and three valleys between them, at 0.9, 0.5 and 1.1. In one dimension a piece of dry land is an interval, and nothing can enclose a lake, so islands are the only things to follow. We lower the water and stop at every level where something happens.

Level 2.6. The top of A breaks the surface. One island exists, born at 2.6.

Level 2.0. The top of D breaks the surface, far from A. A second island is born, at 2.0.

Level 1.8 *and level* 1.4. The same happens at B and at C. Four islands are now dry: A, D, B and C, born at 2.6, 2.0, 1.8 and 1.4.

Level 1.1. The water drops to the bottom of the valley between C and D (the point with height 1.1). Below this level the two islands are joined by dry ground. Two islands have become one. Which one stopped existing? Neither has disappeared physically; they have merged. But the number of islands has dropped by one, so for the record we must decide which island ends here and which continues.

The elder rule. The rule we adopt is that the island that was born later, the one with the lower summit, dies at the merger, and the older one carries on with the merged land. Here C (born at 1.4) is younger than D (born at 2.0), so C dies at 1.1, and its record is the pair (1.4, 1.1). Mountaineers know this number: the height of a summit above the highest pass on any route from it to higher ground (the pass the best route crosses) is its *prominence*, and C's prominence is $1.4 - 1.1 = 0.3$. The rule is not arbitrary. Suppose the older island died instead. The highest summit of the whole map, which is the oldest island there is, would then die at the first pass where it meets any other island, and the diagram would say that the tallest mountain lives only as long as the smallest pebble beside it. With the elder rule the older island carries on, a summit dies at the first pass, going down from the summit, that reaches higher ground, and every summit except the highest is paired with exactly one pass, whatever order the pixels are visited in (equal heights are ties, broken in any fixed order, which changes no persistence).

Level 0.9. The valley between B and A floods dry. B (born at 1.8) is younger than A (born at 2.6), so B dies: the pair (1.8, 0.9).

Level 0.5. The valley between A and the merged C–D island dries. That island carries the birth level of D, 2.0, which is younger than A's 2.6, so it dies: the pair (2.0, 0.5).

Below 0.5. Nothing else merges. The edges at 0.2 and 0.3 join A as the water sinks further. A never dies; we give it the death level $-\infty$, the pair $(2.6, -\infty)$.

The whole history of the landscape is therefore four pairs,

$$\{(2.6, -\infty),\ (2.0, 0.5),\ (1.8, 0.9),\ (1.4, 1.1)\}, \tag{12.38}$$

one for each summit. The differences between birth and death, ∞, 1.5, 0.9 and 0.3, are the

prominences, and they rank the summits by how much of the water's descent each survives as a separate island.

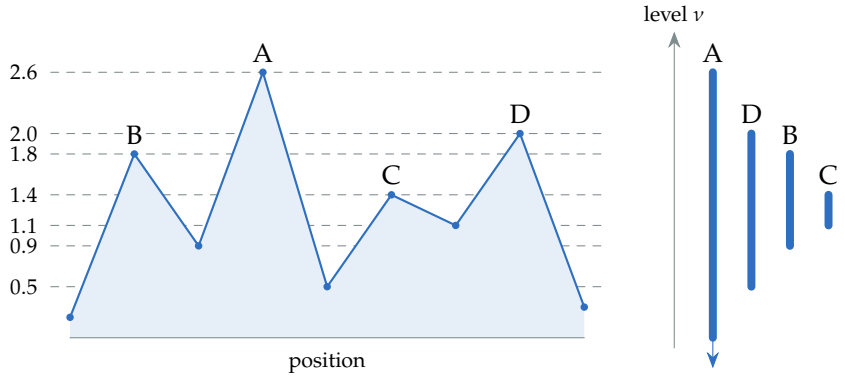


Figure 12.19: The landscape (12.37) and its barcode. The dashed lines are the water levels at which something happens. On the right each island is a vertical bar drawn against the same height axis: it starts at the height of its summit, where the island appears, and ends at the height of the pass where it merges into an older island. The bar of A, the highest summit, never ends.

Working definition 12.3 (filtration, persistence pair, diagram, barcode)

A **filtration** is a family of sets X_ν that only grows as the parameter moves in one direction [here the dry land (12.36), which grows as the level ν drops]. A feature (an island, a lake, a void) is **born** at the level b where it first appears and **dies** at the level d where it merges into an older feature or fills up [for superlevel sets $b \geq d$, because the level decreases]. The pair (b, d) is a **persistence pair** and $b - d$ is its **persistence**. The **persistence diagram** is the set of all pairs, drawn as points in the (b, d) plane, one diagram for islands (H_0) and one for lakes (H_1) [H_k is the standard name for the k-dimensional features: $k = 0$ pieces, $k = 1$ loops, $k = 2$ enclosed voids]. The **barcode** draws the same pairs as intervals. *Operationally:* sort the pixels by height, lower the level through them, and write down a pair whenever a feature dies.

The name of the subject comes from the H_k: they are the homology groups of algebraic topology, whose dimensions are the Betti numbers β_k of Section 12.2. Following them through a filtration is **persistent homology** [268]. We need no more of the algebra than the counts (see the topology notes for the groups themselves), but one fact explains why the record is trustworthy. Inside a computer the dry land is built from cells: the corners, edges and squares of the pixels. Adding one cell changes exactly one Betti number, by exactly one. A new corner is a new island. A new edge either joins two different islands, so one island dies, or joins two corners of the same island, which closes a loop around some water, so a lake is born. A new square either fills a loop, so a lake dies, or changes nothing. (A dry pixel brings its square together with any of its edges and corners that are not yet dry, all at the same level, so several such events can share one level.) Every change of a Betti number is therefore a single event that either creates a feature or destroys one. Pairing each destruction with the creation it undoes, by the elder rule, is what makes the history a list of non-overlapping lifetimes rather than a tangle [268, 269].

12.9.2 A grid by hand: islands and a lake

In two dimensions dry land can enclose water, and a second kind of feature appears. Take the 5×5 grid of heights

$$\phi = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 9 & 4 & 8 & 0 \\ 0 & 5 & 1 & 6 & 0 \\ 0 & 3 & 7 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \tag{12.39}$$

row 0 at the top. As in Section 12.2.2, a pixel is dry when its height is at least the water level, dry pixels that touch along an edge or at a corner belong to the same island (8-neighbours), and wet pixels belong to the same body of water only if they share an edge (4-neighbours). A body of water that does not reach the edge of the map is a lake. We lower the level through the distinct heights $9, 8, 7, \ldots$ and watch (Figure 12.20).

Level 9. The pixel in row 1, column 1 dries: island A is born at 9.

Level 8. The pixel $(1, 3)$ dries. Its eight neighbours are all still wet, so it is a new island, B, born at 8.

Level 7. The pixel $(3, 2)$ dries. Its neighbours $(2, 1)$, $(2, 2)$, $(2, 3)$ have heights 5, 1, 6, all still wet. A third island, C, is born at 7.

Level 6. The pixel $(2, 3)$ dries. It touches B at $(1, 3)$ through an edge and C at $(3, 2)$ through a corner, so B and C merge. C is younger (born at 7 against 8), so C dies: the pair $(7, 6)$.

Level 5. The pixel $(2, 1)$ dries. It touches A at $(1, 1)$ through an edge and the B–C island at $(3, 2)$ through a corner. B is younger than A, so B dies: the pair $(8, 5)$. All the dry land is now one island.

Level 4. The pixel $(1, 2)$ dries and joins the island. Nothing merges, but look at the centre pixel $(2, 2)$, height 1, still wet. Its four edge-neighbours $(1, 2)$, $(2, 1)$, $(2, 3)$, $(3, 2)$ are now all dry, so the water at the centre is cut off from the open water outside: a lake is born at 4.

Levels 3 and 2. The pixels $(3, 1)$ and $(3, 3)$ dry and join the island. The lake is unchanged.

Level 1. The centre itself dries. The lake has drained: it dies at 1, the pair $(4, 1)$.

Level 0. The border dries and joins the island. A never dies: $(9, -\infty)$.

The diagrams are

$$\text{islands } (H_0): \ \{(9, -\infty),\ (8, 5),\ (7, 6)\}, \qquad \text{lakes } (H_1): \ \{(4, 1)\}. \tag{12.40}$$

A lake is born at the pass where the ring of land around it closes, and dies at the bottom of the lake, the lowest point it contains; it is the island story turned upside down, which Section 12.11.2 turns into an algorithm. Figure 12.21 shows the same history as a barcode and as a persistence diagram.

12.9.3 From the diagram back to the counts

The diagram contains every threshold count of Section 12.3. *How many islands are dry at level v?* An island is present at level v when it has been born and has not yet died. It is born at b and the level is decreasing, so it has been born if $v \le b$; it dies at d, so it is still alive if $v > d$. Counting the pairs that pass both tests gives the **Betti curve**

$$\beta_k(v) = \#\{(b, d) \in \mathrm{Dgm}_k : \ d < v \le b\}, \tag{12.41}$$

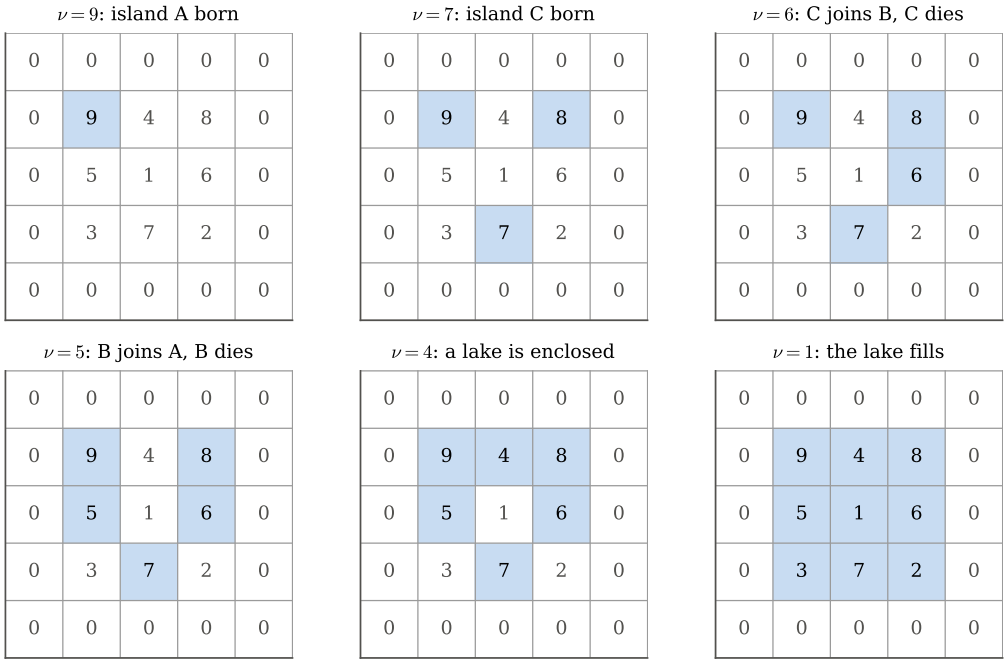


Figure 12.20: The 5×5 landscape (12.39) at six water levels; dry pixels are shaded. Island A appears at 9, B at 8 (not shown) and C at 7; C merges into B at 6 and B into A at 5. At 4 the ring of land around the centre closes and the centre becomes a lake, which drains when the level reaches the centre's height 1.

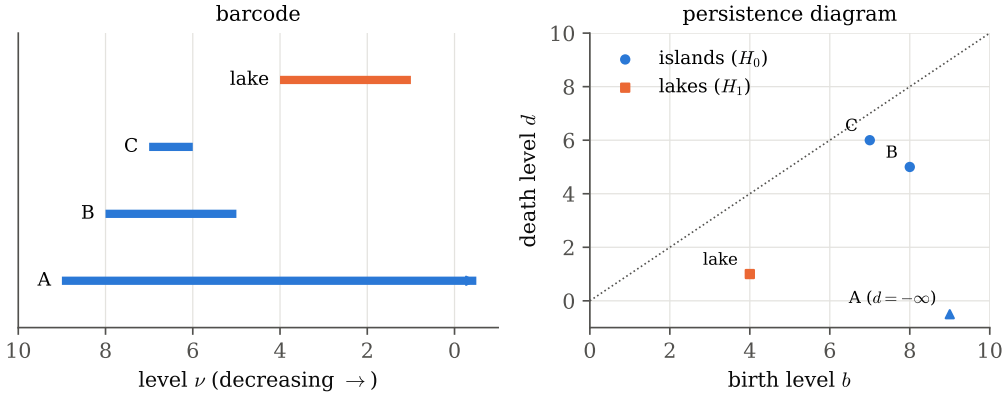


Figure 12.21: The history of Figure 12.20 in the two standard forms. Left: the barcode, one bar per feature, from its birth level to its death level, with the level decreasing to the right. Right: the persistence diagram, one point (b, d) per feature; the dotted line is $d = b$. Island A, which never dies, is drawn as a triangle at the bottom of the panel. The farther a point lies below the diagonal, the longer the feature lived.

where Dgm_k is the diagram of the k-dimensional features. In the diagram plane the condition $d < \nu \le b$ picks out the points to the right of the vertical line $b = \nu$ and below the horizontal line $d = \nu$: a quadrant with its corner on the diagonal at (ν, ν) [5].

Example: the line at $\nu = 1$. With the pairs (12.38),

$$\beta_0(1.0) \overset{(1)}{=} \#\{(2.6,-\infty),(2.0,0.5),(1.8,0.9),(1.4,1.1) \ : \ d < 1.0 \le b\}$$
$$\overset{(2)}{=} \#\{(2.6,-\infty),(2.0,0.5),(1.8,0.9)\} = 3.$$

(1) Definition (12.41) applied to the four pairs of the line.
(2) All four births are at least 1.0. The deaths $-\infty$, 0.5 and 0.9 are below 1.0; the death 1.1 is not, because C has already merged into D when the water is at 1.0.

Directly from (12.37): the heights at least 1.0 are $h_1 = 1.8$, $h_3 = 2.6$ and the run $h_5, h_6, h_7 = 1.4, 1.1, 2.0$, which form three separate intervals. The two counts agree, as they must.

Example: the grid at $\nu = 4.5$ and $\nu = 3$. From (12.40), at $\nu = 4.5$ only the island pair $(9, -\infty)$ satisfies $d < 4.5 \le b$: the pairs $(8, 5)$ and $(7, 6)$ have deaths 5 and 6, not below 4.5, so B and C have already merged into A. Hence $\beta_0(4.5) = 1$. The lake pair $(4, 1)$ fails $4.5 \le b = 4$, so $\beta_1(4.5) = 0$: the ring has not closed yet. At $\nu = 3$ the lake pair satisfies $1 < 3 \le 4$, so $\beta_1(3) = 1$, and still $\beta_0(3) = 1$. The Euler characteristic $\chi = \beta_0 - \beta_1$ of Eq. (12.5) is 1 at $\nu = 4.5$ and 0 at $\nu = 3$, the values of a disc and of a ring.

Example: the diagram knows more than the curve. Two different landscapes can have the same Betti curve. Take one with island pairs $\{(3, 1), (2, 0)\}$ and another with $\{(3, 0), (2, 1)\}$, each with one immortal island besides. In both, β_0 counts one extra island for $2 < \nu \le 3$, two for $1 < \nu \le 2$, and one for $0 < \nu \le 1$, so the curves are identical. The diagrams are not. In the first, the island born at 3 merges at 1, and in the second the island born at 3 survives down to 0 while the one born at 2 merges at 1. The curve counts how many islands are present at each level; the diagram also says which island is which, so it knows how long each one lived. That is why research analyses that use Betti curves [4, 5] also look at the diagrams themselves.

12.9.4 Persistence measures robustness to the choice of threshold

A feature with pair (b, d) is present in X_ν for every level in the interval $d < \nu \le b$, an interval of width

$$\Delta\nu = b - d. \tag{12.42}$$

This is the sense in which persistence answers the complaint at the start of the section. Any analysis that picks one threshold inside that interval sees the feature; any threshold outside it misses it. A feature with large $\Delta\nu$ is seen by a wide range of reasonable choices, and conclusions drawn from it do not hang on one of them. A feature with tiny $\Delta\nu$ is seen only in a narrow window of levels, and whether it is counted at all depends on where the threshold happens to fall. In the diagram, long-lived features are points far below the diagonal $d = b$, and short-lived ones crowd along it.

Key idea: Persistence

Lowering the threshold through every level turns the excursion sets into a filtration. Each island, lake or void has a birth level and a death level; the pairs form the persistence diagram, and the number of pairs alive at any level is the Betti number there (12.41). The persistence $b - d$ is the range of thresholds over which a feature exists: points far from the diagonal are structures that no reasonable choice of threshold can miss.

Whether a short bar is noise or a small real structure is a different question, the question of stability: if the field is known only up to an error of size ϵ, a feature whose persistence is below about 2ϵ may have been created by the error, while a feature that persists much longer cannot have been.

12.10 Small changes of the field, small changes of the diagram

An observed map contains a beam-smoothed signal, instrumental noise and occasionally a bad pixel. In the landscape of Section 12.9, a small noise fluctuation can introduce a summit and a nearby pass, giving the noisy map an island that was absent from the underlying field. What distinguishes such an island is the small difference between its birth and death levels. It appears close to the diagonal of the persistence diagram, whereas a prominent island can survive a small perturbation with only a modest shift in its birth and death levels. We can quantify this distinction and check it on simulated maps.

The question. Suppose every value of the field changes by at most ϵ,

$$\|f - g\|_\infty \equiv \max_{\boldsymbol{x}} |f(\boldsymbol{x}) - g(\boldsymbol{x})| \le \epsilon, \tag{12.43}$$

where f is the true field and g the measured one. How far can the persistence diagram of g be from that of f? To answer we need two things: a way to measure the distance between two diagrams, and an argument for why that distance is small.

12.10.1 The simplest case first: a noise wiggle on a slope

Start where nothing should happen. Let f be a field that only rises along a line, $f(x) \le f(x')$ whenever $x < x'$, so its superlevel sets have no features in this stretch: every dry point is joined to the higher ground on its right. Add noise $\eta(x)$ with $|\eta| \le \epsilon$, so $g = f + \eta$. Suppose the noise has made a small summit at x_p, a separate island of g born at $g(x_p)$. To merge into the higher ground on its right it must be joined through the lowest point $x_s > x_p$ between them, the pass, where it dies at $g(x_s)$. How long can it live?

$$\begin{aligned} g(x_p) - g(x_s) &\overset{(1)}{=} \big[f(x_p) - f(x_s)\big] + \big[\eta(x_p) - \eta(x_s)\big] \\ &\overset{(2)}{\le} 0 + \big[\eta(x_p) - \eta(x_s)\big] \\ &\overset{(3)}{\le} \epsilon + \epsilon = 2\epsilon. \end{aligned} \tag{12.44}$$

(1) Write $g = f + \eta$ at both points and regroup.
(2) The pass lies to the right of the summit, $x_s > x_p$, and f rises to the right, so $f(x_p) \le f(x_s)$.
(3) Each noise value lies between $-\epsilon$ and ϵ, so a difference of two of them is at most 2ϵ.

A feature that the noise creates out of nothing has persistence at most 2ϵ. Its point (b, d) in the diagram therefore lies within a vertical (or horizontal) distance 2ϵ of the diagonal, and within ϵ of the diagonal point $\left(\frac{b+d}{2}, \frac{b+d}{2}\right)$ in each coordinate. The same chain read the other way shows that a genuine island of f is not destroyed by the noise unless its own persistence is below 2ϵ, and that its birth (a summit height) and its death (a pass height) each move by at most ϵ, because every height does.

12.10.2 A distance between diagrams

To compare two diagrams we pair up their points and ask how far the paired points are. Two complications appear at once. The diagrams may have different numbers of points, and a short-lived feature present in one may simply be absent from the other. Both are handled by allowing a point to be paired with the diagonal, which means "this feature has shrunk to zero persistence in the other diagram".

Working definition 12.4 (bottleneck distance)

A **matching** between diagrams D and D' pairs each point of D either with one point of D' or with the diagonal, and likewise each point of D'. The cost of pairing (b, d) with (b', d') is $\max(|b - b'|, |d - d'|)$ [the larger of the shift in birth and the shift in death]; the cost of pairing (b, d) with the diagonal is $(b - d)/2$ [the shift needed to move it onto the diagonal, to the point $((b + d)/2, (b + d)/2)$]. The **bottleneck distance** $d_B(D, D')$ is the smallest possible value of the largest cost in a matching. *Operationally:* try matchings, record the worst pair in each, and keep the matching whose worst pair is best.

Example: a matching by hand. Let $D = \{(3, 0), (2, 1.6)\}$ and $D' = \{(2.8, 0.1)\}$ (Figure 12.22). Only two matchings are worth trying. In the first, $(3, 0)$ goes with $(2.8, 0.1)$ and $(2, 1.6)$ goes to the diagonal:

$$\text{worst cost} \overset{(1)}{=} \max\left(\max(|3 - 2.8|, |0 - 0.1|),\ \tfrac{2-1.6}{2}\right) = \max(0.2,\ 0.2) = 0.2.$$

(1) The first entry is the cost of the paired points, the second the cost of sending $(2, 1.6)$ to the diagonal.

In the second, $(2, 1.6)$ goes with $(2.8, 0.1)$ and $(3, 0)$ goes to the diagonal: the costs are $\max(0.8, 1.5) = 1.5$ and $3/2 = 1.5$, so the worst is 1.5. The bottleneck distance is the better of the two, $d_B = 0.2$. The long-lived feature moved by 0.2, and the short one could have been created or destroyed by a change of size 0.2.

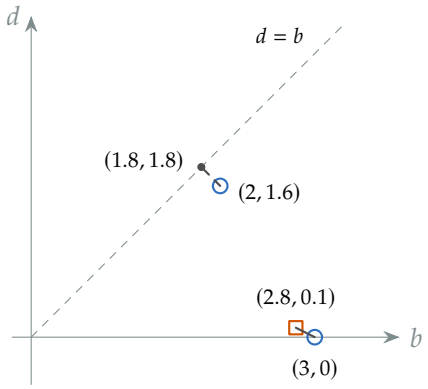


Figure 12.22: The best matching between D (circles) and D' (square). The long-lived point $(3, 0)$ is paired with $(2.8, 0.1)$; the short-lived point $(2, 1.6)$ has no partner and is sent to the nearest point of the diagonal, $(1.8, 1.8)$, at cost 0.2. Both costs are 0.2, so $d_B = 0.2$.

12.10.3 The stability theorem

The wiggle of Section 12.10.1 gave the bound in the simplest case. Cohen-Steiner, Edelsbrunner and Harer proved that it holds in general [270].

Key idea: Stability of persistence diagrams

For two fields f and g on the same domain (continuous, with finitely many critical levels, which covers every pixel map),

$$d_B\big(\mathrm{Dgm}_k(f),\ \mathrm{Dgm}_k(g)\big) \le \|f - g\|_\infty \qquad \text{for every } k.$$

Physically: a perturbation of size ϵ moves every robust feature by at most ϵ in birth and in death, and every feature it creates or destroys has persistence at most 2ϵ, so it sits within

the band $b - d \le 2\epsilon$ along the diagonal.

Why it is true, at the physicist's level. Births and deaths are heights of special points, summits, passes and lake bottoms, and (12.43) says that no height moves by more than ϵ. If the special points of g were simply the special points of f, moved a little, every pair would move by at most ϵ in each coordinate and we would be done. The real work in the proof is the case where g has extra special points (the noise wiggles) or fewer (a feature of f washed out); the slope argument (12.44) shows that such points can only belong to pairs of persistence at most 2ϵ, which the matching sends to the diagonal at cost at most ϵ. The general proof organises this bookkeeping for all levels at once [270]. Note what the theorem does *not* say: it does not bound the number of new points, only where they can be. A noisy map can have a hundred times more islands than the clean one and still be within ϵ of it in the bottleneck distance, because all the extra islands are near the diagonal [5].

Testing it. The script `code/ch12/b03_stability.py` makes a smooth Gaussian field f on a 128×128 grid, adds either white noise or noise smoothed over a pixel and a half, at amplitudes from 0.01 to 0.4 standard deviations of f, and computes $\|f - g\|_\infty$ and the bottleneck distance between the island diagrams and between the lake diagrams of f and g (with gudhi's `bottleneck_distance`, checked once against persim; they agree to 0). Over all 60 trials the largest value of $d_B / \|f - g\|_\infty$ is 0.9132: the bound is never broken, and it comes close to being reached (Figure 12.23, left). It cannot be improved in general, as the simplest perturbation shows: add the same constant ϵ to every pixel. Every summit and every pass rises by exactly ϵ, so every point of the diagram moves by ϵ in birth and in death, no matching does better, and $d_B = \epsilon = \|f - g\|_\infty$.

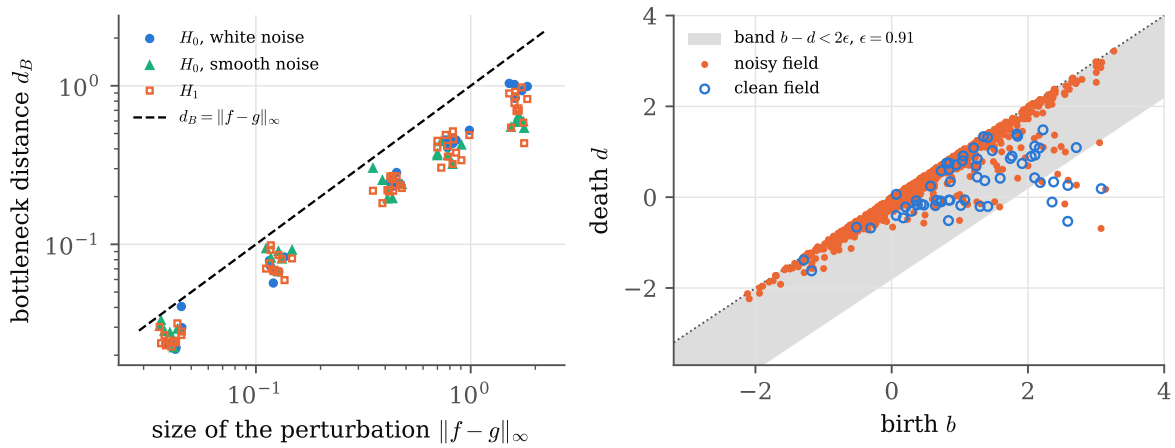


Figure 12.23: Left: bottleneck distance between the diagrams of a clean field and a noisy copy, against the size of the noise, for white and smoothed noise and for islands and lakes. Every point lies below the dashed line $d_B = \|f - g\|_\infty$. Right: the island diagrams of one clean field (open circles) and of the same field with white noise of 0.2 standard deviations added (dots). The grey band is $b - d < 2\epsilon$. The noise adds a crowd of points inside the band; the points of the clean field far from the diagonal have partners close by.

Example: where the noise islands sit. In the right panel of Figure 12.23 the white noise has standard deviation 0.2, but its largest value over the 16 384 pixels is $\epsilon = 0.908$ (the maximum of many Gaussian numbers lies several standard deviations out). The clean field has 60 islands; the noisy one has 871. Of these, 863 have persistence below 2ϵ, inside the band the theorem

allows for created features. The most persistent island lives for 3.117 in the clean field and 3.763 in the noisy one; the tenth most persistent, 1.502 and 1.614. Smoothing the noisy map over a pixel and a half brings the count back to 57 islands: the smoothing removes the wiggles that made the short bars, which is the beam of Section 8.2.1 doing, for topology, what it does for the power spectrum.

What the bound costs us. The theorem controls the worst pixel, not the typical one. For Gaussian noise the worst pixel of a map of 10^4 to 10^6 pixels lies about 4 to 5 standard deviations out, so the band of 2ϵ is wide, and many genuine but small features of the field lie inside it, where the theorem cannot tell them from noise. Worse, one bad pixel (a cosmic-ray hit, an unmasked point source) makes ϵ large on its own. In practice we do not rely on the worst-case band. We simulate the noise, pass the simulations through the same pipeline, and read off where noise features actually fall in the diagram; that is the Monte Carlo of part 12c. The theorem's job is the qualitative guarantee: features far from the diagonal are safe, and the diagram is a continuous function of the data, so it is a sensible input to a statistical analysis.

Watch out: Bottleneck is not the only distance

The bottleneck distance looks only at the worst-matched pair and ignores everything else. The p-**Wasserstein distance** instead adds up the costs, $W_p = \left(\min_{\text{matchings}} \sum \text{cost}^p\right)^{1/p}$, so it feels all the small points near the diagonal too. It is more sensitive to how many noise features there are, and for that reason it is not bounded by $\|f - g\|_\infty$ alone; its stability needs extra conditions on the fields [269]. Use the bottleneck distance when the question is "did any robust feature move?", and a Wasserstein distance when the whole population of features matters.

12.11 Computing persistence: union–find and the elder rule

The 5×5 landscape of Section 12.9.2 took a page by hand. A 256×256 simulated patch has 65 536 pixels, and a HEALPix map at $N_{\text{side}} = 2048$ has fifty million. The procedure itself does not change: sort the pixels from the highest to the lowest, let them become dry one at a time, and whenever a new dry pixel touches dry neighbours, find out which islands those neighbours belong to and merge them, recording a death by the elder rule. The only hard part is the question "which island does this pixel belong to?", asked again and again. If we answered it by relabelling whole islands at every merger, a big island would be relabelled thousands of times. The standard answer is a data structure called **union–find**, which makes each question cost almost nothing.

The question. How can a program keep track of which island every dry pixel belongs to, merge two islands in one step, and remember the summit height of each island, so that it can apply the elder rule?

12.11.1 Union–find

Store each island as a tree of pointers. Every dry pixel p holds one number, parent$[p]$, the pixel it points to. Following the pointers from any pixel of an island always ends at the same pixel, the **root**, which points to itself. The root is the island's name, and we store the island's summit height at the root. Two operations are needed.

- *Find*: to learn which island p belongs to, follow parent from p until reaching a pixel that points to itself.
- *Union*: to merge two islands, make the root of one point to the root of the other. One pointer changes; nothing else does.

Trees made this way can grow long chains, which would make *find* slow. A one-line repair keeps them shallow: while walking up, make each visited pixel point to its grandparent (**path halving**). With it, a long sequence of operations costs, on average, a small constant number of steps each, so the total work is proportional to the number of pixels, apart from the sort that puts them in order of height [269].

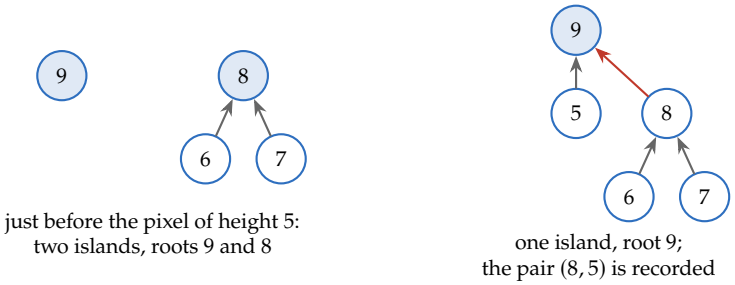


Figure 12.24: Union–find on the island step of the 5×5 example at level 5 (Section 12.9.2). Each circle is a dry pixel labelled by its height, each arrow its parent pointer; the shaded circles are roots, which carry the summit height of their island. When the pixel of height 5 joins both islands, the root of the younger island (summit 8) is made to point to the root of the older one (summit 9, red arrow), and the death of the younger island at level 5 is written down.

The whole algorithm for the islands of a pixel map is then the function `islands` of `code/ch12/lib_persist.py`:

Listing 12.7: Islands (H_0) of the superlevel sets by union–find (`code/ch12/lib_persist.py`)

```
def islands(f, conn: int = 8):
    """Birth-death pairs of the islands of {f >= nu} as the level nu is lowered
    (union-find)."""
    ny, nx = f.shape
    h = f.ravel()
    order = np.argsort(-h, kind="stable")       # pixels from the highest to the lowest
    parent = np.full(h.size, -1)                # -1 means: still under water
    peak = np.empty(h.size)                     # for the root of each island: its peak
    height
    steps = NEIGH8 if conn == 8 else NEIGH4

    def root(p):                                # follow parents to the island's root
        while parent[p] != p:
            parent[p] = parent[parent[p]]       # path halving keeps the trees shallow
            p = parent[p]
        return p

    pairs = []
    for p in order:                             # lower the water level pixel by pixel
        parent[p], peak[p] = p, h[p]            # p emerges as a one-pixel island
        i, j = divmod(p, nx)
        for di, dj in steps:
            a, b = i + di, j + dj
            if not (0 <= a < ny and 0 <= b < nx) or parent[a * nx + b] < 0:
                continue                        # off the grid, or neighbour still under
    water
            r1, r2 = root(p), root(a * nx + b)
            if r1 == r2:
```

```
55                 continue                          # already the same island
56             if peak[r1] < peak[r2]:
57                 r1, r2 = r2, r1                   # r1: the elder island (higher peak)
58             if peak[r2] > h[p]:                   # the younger one dies at this level
59                 pairs.append((peak[r2], h[p]))
60             parent[r2] = r1                       # ... and becomes part of the elder
61     pairs.append((h[order[0]], -np.inf))          # the highest island never dies
62     return np.array(pairs)
```

Line 34 sorts the pixels from the highest to the lowest: this is the falling water. Lines 35–36 hold the two arrays of union–find, `parent` (with -1 for a pixel still under water) and `peak`, the summit height stored at each root. Lines 39–43 are *find* with path halving. The loop of line 46 makes one pixel dry per step. Line 47 makes it a one-pixel island, its own root, with its own height as summit. Lines 49–52 visit the eight neighbours and skip those off the map or still wet. Line 53 finds the roots of the new pixel and of the neighbour. If they differ, the two islands are merging at the level $h[p]$ of the new pixel. Lines 56–57 apply the elder rule: `r1` becomes the root with the higher summit. Lines 58–59 record the death of the younger island, (its summit, $h[p]$), unless the younger "island" is just the new pixel itself, whose summit equals $h[p]$ and whose pair would have zero persistence. Line 60 is the *union*. When every pixel is dry, line 61 adds the island that never dies.

Example: the code on the hand example. Run on the grid (12.39), `islands` returns the pairs $(7,6)$, $(8,5)$ and $(9,-\infty)$, exactly (12.40), in the order in which the deaths happen (`code/ch12/b01_hand_examples.py`). gudhi, given the same grid, returns the same three pairs.

12.11.2 Lakes from islands: duality in the plane

Lakes need a second idea. A lake of the dry land is a body of water that does not reach the open sea. So instead of watching the land, watch the water. The water at level v is the set $\{\phi < v\}$. As the level falls, the water shrinks; read backwards, with the level rising, the water grows, starting at the lowest points of the map and spreading. That is exactly the island story for the upside-down landscape $-\phi$: its summits are the lake bottoms of ϕ, and its passes are the passes of ϕ.

The translation. Run `islands` on $-\phi$, with two changes. First, water pixels are connected only through edges, 4-neighbours, because the land is 8-connected and the two must be dual (Section 12.2.2). Second, surround the map with a frame of open sea that is always wet, by giving the frame the height $+\infty$ in $-\phi$; the frame then becomes the oldest island of $-\phi$, and any water connected to the edge of the map merges into it. Now take a water body of $-\phi$ born at $-\phi = -m$ and merging at $-\phi = -s$. In the language of the land: the water body has its bottom at height m and joins other water through a pass of height $s > m$. With the level falling, it is part of a larger body while $v > s$; when v drops below s it is cut off from the rest, so if the rest is the sea, a lake is born at s; and it drains when v drops below its bottom m. So a pair $(-m,-s)$ of $-\phi$ becomes the lake pair (s,m) of ϕ. On the hand example the centre pixel is a water island of $-\phi$ born at -1. In the reversed picture the water rises; the first of its four edge-neighbours to go under is the pixel of height 4, which is already joined to the border, so the centre's water body merges into the sea at -4. The lake pair is $(4,1)$, as found by hand. That this bookkeeping gives exactly the lakes, the holes of the land, is a case of **Alexander duality** (see the topology notes): in the plane, the holes of a set are the bounded pieces of its complement.

Listing 12.8: Lakes (H_1) by duality: the islands of $-\phi$ with a sea around the map (`code/ch12/lib_persist.py`)

```
65 def lakes(f):
66     """Birth-death pairs of the holes (lakes) of {f >= nu}, by Alexander duality in the
       plane.
67
68     A hole of the land is a piece of the water {f < nu} that does not reach the open sea.
       We
69     surround the map by a frame of sea (+inf in -f), compute the islands of -f with
70     4-neighbours, and translate: a water body born at -f = -m (the lake bottom, height m)
       and
71     merging into another at -f = -s (the pass, height s) is a lake of the land that appears
72     when the level drops to s and fills up when it drops below m.
73     """
74     g = np.pad(-np.asarray(f, float), 1, constant_values=np.inf)
75     p = islands(g, conn=4)
76     p = p[np.isfinite(p[:, 0]) & np.isfinite(p[:, 1])]  # drop the open sea itself
77     return np.column_stack([-p[:, 1], -p[:, 0]])        # (birth, death) = (pass, bottom)
```

Line 74 builds $-\phi$ with the frame of sea, line 75 runs the island code with 4-neighbours, line 76 drops the one immortal water body (the sea), and line 77 translates each pair back.

On the torus the shortcut fails. A periodic map, such as a simulation box, has no edge and so no open sea. Two new things happen. A band of land that wraps around the torus contains a loop that encloses no lake; and the water can also wrap around, so no water body is "outside". The lake count then misses the loops that wind around the torus, and the frame of sea, if we add it, cuts features that should continue across the edges. For periodic maps we therefore use gudhi's `PeriodicCubicalComplex` (Section 12.13), which builds the torus directly.

12.11.3 Checks against the libraries, and timing

`code/ch12/b02_grf_persistence.py` makes one Gaussian field on a 256×256 grid, smoothed on 4 pixels (the recipe of Section 7.4), and runs four independent checks.

- *gudhi.* Its `CubicalComplex`, built from the pixels as squares, computes the persistence of the sublevel sets of $-\phi$, which are our superlevel sets of ϕ with the same 8- and 4-neighbour rules. Both codes find 207 island pairs and 231 lake pairs, and the bottleneck distance between our diagrams and gudhi's (computed by persim) is 0: they are identical.
- *ripser.* Its `lower_star_img` computes the island pairs of an image. The bottleneck distance to ours is 1.115×10^{-7}, the rounding error of the single-precision numbers ripser uses internally.
- *scipy.* At 61 levels from -3 to 3 we count the islands directly with `scipy.ndimage.label` (8-neighbours) and the lakes as the 4-connected water bodies that do not touch the edge. These counts and the Betti curves (12.41) read from our diagrams differ by at most 0 at any level.
- *scikit-image.* `skimage.measure.label` (with `connectivity=2`, its name for 8-neighbours) counts the islands again, and `skimage.measure.euler_number` computes χ of the dry land by counting cells, as in Section 12.2.3. They equal β_0 and $\beta_0 - \beta_1$ from our diagrams at every level.

A fifth check is a physical one. The field $e^{0.8\phi}$ is a strictly increasing function of ϕ, so by the argument of Section 12.3.2 its dry land at level $e^{0.8\nu}$ is the dry land of ϕ at level ν. Its diagram must be the diagram of ϕ with every level relabelled, and it is: taking the logarithm of its pairs and dividing by 0.8 gives back the pairs of ϕ exactly.

Figure 12.25 shows what the diagram of a smooth random field looks like. The islands are born at summits, almost all above $\nu = 0$, and die at passes, spread from about -2 to 2. The lakes

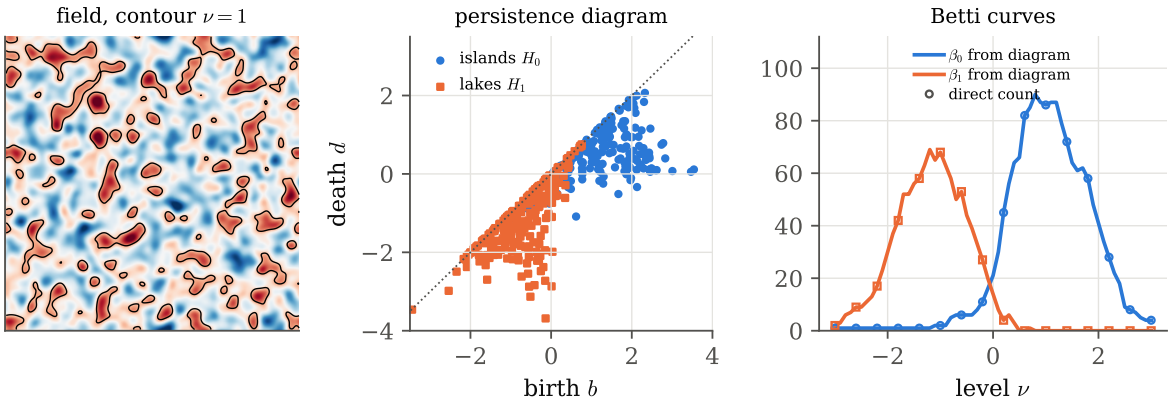


Figure 12.25: One 256×256 Gaussian field and its persistence. Left: the field, with the boundary of the dry land at $\nu = 1$. Middle: the diagrams of islands (circles) and lakes (squares). Islands are born at positive levels, the summits, and lakes die at negative levels, the lake bottoms. Right: the Betti curves read from the diagrams (lines) and counted directly with `scipy.ndimage.label` (open symbols).

are their mirror image: born at passes, dying at lake bottoms, almost all below 0. The mirror symmetry is the one of Section 12.3: a Gaussian field is as likely as its negative, so the lakes of ϕ are statistically the islands of $-\phi$. The longest-lived island lives for 3.427, and a fraction 0.3738 of the islands live for less than 0.25; even without noise, a smooth field has many shallow summits sitting on the shoulders of bigger ones.

Timing. The union–find code is plain Python, one loop over pixels. Its run time grows in proportion to the number of pixels (a fitted power of 1.001), from about a second at 256^2 to 15.83 s for islands and lakes at 1024^2; gudhi, written in C++, takes 1.757 s there (Figure 12.26). Both are fast enough for every map used in this chapter's experiments. Scaled to a full-resolution *Planck* map of fifty million pixels, the measured times become about thirteen minutes for the Python code and about a minute and a half for gudhi (on the sphere the neighbours would come from `healpy.get_all_neighbours` instead of the grid).

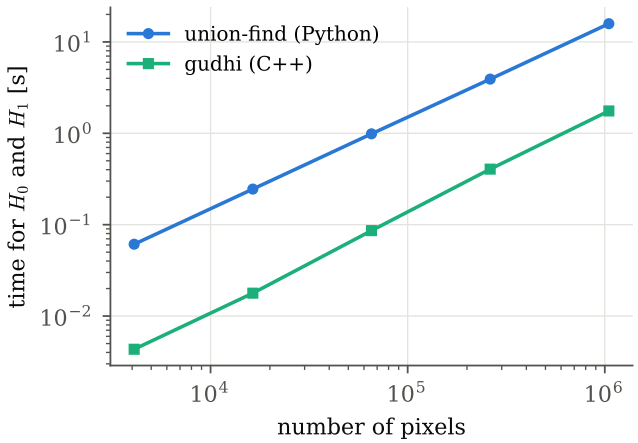


Figure 12.26: Time to compute the island and lake diagrams against the number of pixels, for the union–find code of Section 12.11.1 and for gudhi. Both grow in proportion to the number of pixels; the compiled library is about ten times faster.

How this is used in real research: Persistence software in research

Research codes never write their own union–find. For pixel maps and simulation boxes the standard tools are gudhi's cubical complexes (periodic or not), Cubical Ripser, and DIPHA for maps too large for one machine; for point clouds, ripser and gudhi's Rips and alpha complexes; and for diagrams, persim and gudhi's own distances [269]. Cosmic-web studies compute persistence on a Delaunay tessellation of the particles with the density estimated at each particle, using the PHAT library [4, 5]. The from-scratch code earns its place as a check: on a small map, a disagreement between it and the library almost always means a convention has been mixed up (superlevel against sublevel sets, 4- against 8-neighbours, the sign of the field).

12.12 Point clouds: shapes grown from points

A galaxy survey or an N-body simulation does not hand us a field. It hands us points: positions of galaxies, haloes or dark-matter particles. One way forward is to turn the points into a density field on a grid and use the landscape of the previous sections; Section 12.13 does that. The other is to build shapes directly from the points. Put a small ball around every point and let all the balls grow together. When two balls touch, their points are linked; when three points are pairwise linked, the triangle between them is filled in. At a small radius every point is alone. As the radius grows, close points join into clusters, clusters join each other, rings of points close into loops, and the loops eventually fill in. The radius plays the role the water level played for a field, and the same bookkeeping of births and deaths applies.

The question. For a set of points, at which linking distance does each cluster stop being separate, and at which distances do loops of points appear and fill in? And how do we compute this without drawing balls?

Working definition 12.5 (Vietoris–Rips filtration)

For points $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ and a scale $r \geq 0$, the **Vietoris–Rips complex** VR_r contains every point, an edge between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ whenever $|\boldsymbol{x}_i - \boldsymbol{x}_j| \leq r$, and a filled triangle (and higher simplices) on every set of points that are pairwise within r [a triangle enters exactly when its longest edge does]. As r grows the complex only gains edges and triangles, so it is a filtration. Every point is born at $r = 0$; a cluster dies at the r where it joins another; a loop is born at the length of the edge that closes it and dies at the length of the edge that completes the triangles filling it.

In this filtration the parameter increases, so births come before deaths, $b \leq d$, and the diagram points lie above the diagonal; this is the convention of every point-cloud library. Figure 12.27 shows the complex of a noisy ring of 40 points with a small clump of 12 beside it at three scales.

12.12.1 Clusters: the minimum spanning tree

Connectivity in VR_r is decided by the edges alone: triangles fill area but join nothing that their edges have not already joined. So to follow the clusters we only need to add the edges in order of length and ask, for each edge, whether it joins two different clusters. That is **Kruskal's algorithm**: sort all $n(n-1)/2$ edges by length; take them one by one; if an edge joins two different clusters, merge them (one dies at that length) and keep the edge; if it joins a cluster to itself, skip it. The kept edges form the **minimum spanning tree**, the shortest network of edges

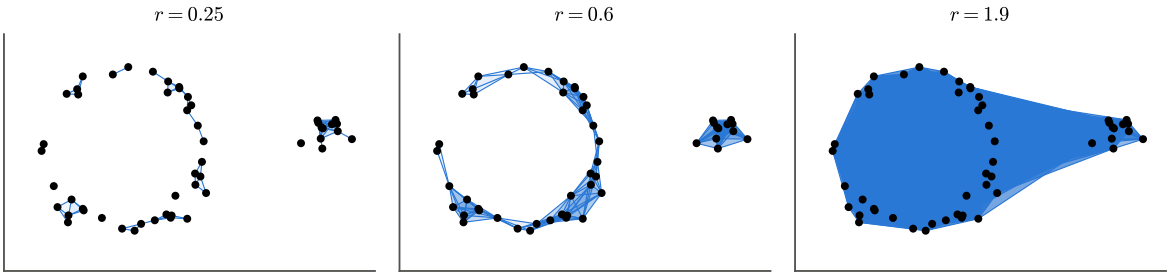


Figure 12.27: The Vietoris–Rips complex of a noisy ring with a clump beside it, at three scales r. At $r = 0.25$ only near neighbours are linked. At $r = 0.6$ the ring is linked all the way round except for one gap on its left, which closes a little later into a loop around an empty middle, and the clump is one cluster. At $r = 1.9$ triangles have filled the middle of the ring and the clump is joined to it: one blob.

that connects all the points, and their lengths are exactly the deaths of the island diagram of the points. All births are at 0, so the elder rule never has to choose by age; any choice of which cluster dies gives the same pairs.

Example: six points by hand. Take $A = (0,0)$, $B = (1,0)$, $C = (0,1)$, $D = (4,0)$, $E = (5,0)$, $F = (4,2)$ (Figure 12.28). The shortest distances, in order, are $AB = AC = DE = 1$, $BC = \sqrt{2}$, $DF = 2$, $EF = \sqrt{5}$, then $BD = 3$. Kruskal takes AB (one cluster dies at 1), AC (another at 1), DE (another at 1). BC joins B and C, already in one cluster through A, so it is skipped; at the same length the triangle ABC enters (its longest edge is BC), so no loop ever exists there. DF joins F to the cluster $\{D, E\}$: a death at 2. EF is skipped. $BD = 3$ joins the two clusters: a death at 3. The island diagram is

$$\{(0,1),\ (0,1),\ (0,1),\ (0,2),\ (0,3),\ (0,\infty)\}, \tag{12.45}$$

and the minimum spanning tree has total length $1 + 1 + 1 + 2 + 3 = 8$. Read the bars: three die at 1, the spacing inside the two groups; one at 2, the loose point F; and one bar, from 0 to 3, stands out. It says that there are two groups of points, separate for all scales between 2 (when the second group is complete) and 3 (when they join). The script `b04_point_cloud.py` returns exactly these deaths, and so does ripser.

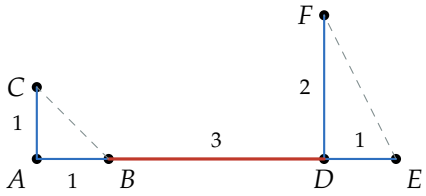


Figure 12.28: The six points of the hand example and their minimum spanning tree. Blue edges join points within a group; the red edge, of length 3, is the last one added and joins the two groups. Dashed edges are skipped by Kruskal's algorithm because their endpoints are already connected.

Two old friends. Hierarchical clustering with "single linkage" joins at each step the two clusters with the closest pair of members; its merge heights are the edge lengths of the minimum spanning tree, and `scipy.cluster.hierarchy.linkage(X, method="single")` returns exactly our deaths. And the friends-of-friends group finder used on N-body simulations, which links any two particles closer than a linking length b and calls each connected group a halo, returns the clusters of VR_b. The island diagram of the particles therefore contains the

friends-of-friends catalogue for every linking length at once: the number of haloes at linking length b is the number of bars still alive at $r = b$.

Listing 12.9: Island diagram of a point cloud: Kruskal's minimum spanning tree (code/ch12/lib_persist.py)

```
98  def kruskal_h0(X):
99      """H0 of the Vietoris-Rips filtration = the edges of the minimum spanning tree
        (Kruskal).
100
101     All points are born at r = 0; a component dies at the length of the edge that merges it
102     into another.  Returns the sorted death values (n - 1 of them; one component never
        dies).
103     """
104     n = len(X)
105     i, j = np.triu_indices(n, k=1)
106     length = np.linalg.norm(X[i] - X[j], axis=1)
107     parent = list(range(n))
108
109     def root(p):
110         while parent[p] != p:
111             parent[p] = parent[parent[p]]
112             p = parent[p]
113         return p
114
115     deaths = []
116     for e in np.argsort(length, kind="stable"):  # shortest edges first
117         r1, r2 = root(i[e]), root(j[e])
118         if r1 != r2:                             # the edge joins two components: one dies
119             parent[r1] = r2
120             deaths.append(length[e])
121             if len(deaths) == n - 1:
122                 break
123     return np.array(deaths)
```

Lines 105–106 list every pair of points and its distance. The loop of line 116 takes the edges from the shortest up; `root` is the *find* of Section 12.11.1. Lines 117–120 keep an edge only when it joins two different clusters, and record its length as a death. After $n-1$ deaths a single cluster is left and the loop can stop.

12.12.2 Loops: reducing the boundary matrix

Loops cannot be found with union–find, which only knows who is connected to whom. The general algorithm keeps, for every edge and triangle, the list of its faces, and does linear algebra on these lists [5, 268].

Example: the unit square by hand. Four points at the corners of a unit square, $v_0 = (0,0)$, $v_1 = (1,0)$, $v_2 = (1,1)$, $v_3 = (0,1)$. At $r = 1$ the four sides enter. Every triangle on these points needs a diagonal, of length $\sqrt{2}$, so no triangle enters yet: the four sides enclose an empty loop, and a loop is born at $r = 1$. At $r = \sqrt{2}$ both diagonals enter and with them all four triangles, which cover the square: the loop dies. The loop diagram is $\{(1, \sqrt{2})\}$.

How the algorithm finds this. List the cells in the order they enter: the four points, then the sides $e_{01}, e_{12}, e_{23}, e_{03}$ (length 1), then the diagonals e_{02}, e_{13} and the triangles (length $\sqrt{2}$). Write each cell as the list of its faces, its **boundary**: $\partial e_{01} = \{v_0, v_1\}$, $\partial t_{023} = \{e_{02}, e_{23}, e_{03}\}$, and so on. Lists are added modulo 2: a face present in both lists cancels. Go through the cells in order. For each, look at the latest-entering face in its list, its *lowest* entry. If an earlier cell already owns

that lowest entry, add that cell's (already reduced) list to this one and look again. Stop when the lowest entry is new, or the list is empty.

- e_{01}: list $\{v_0, v_1\}$, lowest v_1, new. So e_{01} kills the cluster born at v_1: the pair $(0, 1)$. Likewise e_{12} kills v_2 and e_{23} kills v_3.
- e_{03}: list $\{v_0, v_3\}$. Lowest v_3 is owned by e_{23}: add, $\{v_0, v_3\} + \{v_2, v_3\} = \{v_0, v_2\}$. Lowest v_2 is owned by e_{12}: add, giving $\{v_0, v_1\}$; lowest v_1 is owned by e_{01}: add, giving $\{\}$. The list is empty: e_{03} kills nothing, so it creates something. It is the edge that closes the loop, born at $r = 1$.
- The diagonals reduce to empty lists the same way: each closes a new loop at $\sqrt{2}$.
- The triangle t_{012}, list $\{e_{01}, e_{12}, e_{02}\}$, lowest e_{02}, new: it kills the loop that e_{02} created, the pair $(\sqrt{2}, \sqrt{2})$, of zero persistence. The triangle t_{013} kills the loop of e_{13} the same way.
- The triangle t_{023}, list $\{e_{02}, e_{23}, e_{03}\}$. Lowest e_{02} is owned by t_{012}: add, giving $\{e_{01}, e_{12}, e_{23}, e_{03}\}$, the four sides. Lowest e_{03}, new: t_{023} kills the loop born when e_{03} entered, the pair $(1, \sqrt{2})$.

The reduced list of t_{023} is the boundary of the square: the algorithm has discovered that the two triangles t_{012} and t_{023} together fill the loop of the four sides. Every pair of a persistence diagram, in any dimension, is found this way [268].

Listing 12.10: Persistence of a Rips filtration by reducing the boundary matrix (small clouds) (code/ch12/lib_persist.py)

```
def rips_bruteforce(X, maxdim: int = 1):
    """Persistence of the Vietoris-Rips filtration by reducing the boundary matrix over Z_2.

    Simplices up to dimension maxdim + 1.  Filtration value of a simplex = its longest edge.
    Returns {dim: array of (birth, death)}, deaths of essential classes = inf.  Small
    clouds only.
    """
    n = len(X)
    D = np.linalg.norm(X[:, None, :] - X[None, :, :], axis=-1)
    simplices = []
    for k in range(1, maxdim + 3):
        for s in itertools.combinations(range(n), k):
            val = 0.0 if k == 1 else max(D[a, b] for a, b in itertools.combinations(s, 2))
            simplices.append((val, k - 1, s))
    simplices.sort(key=lambda t: (t[0], t[1]))          # by value, faces before cofaces
    index = {s: c for c, (_, _, s) in enumerate(simplices)}
    low_of = {}                                         # pivot row -> column that owns it
    out = {k: [] for k in range(maxdim + 1)}
    paired = set()
    for c, (val, dim, s) in enumerate(simplices):
        col = {index[f] for f in itertools.combinations(s, dim)} if dim > 0 else set()
        while col and max(col) in low_of:               # add earlier columns until the
    pivot is new
            col ^= low_of[max(col)][1]
        if col:
            r = max(col)                                # simplex r (dim-1) is killed by c
            low_of[r] = (c, col)
            paired.update((r, c))
            b = simplices[r][0]
            if dim - 1 <= maxdim and val > b:
                out[dim - 1].append((b, val))
    for c, (val, dim, s) in enumerate(simplices):       # unpaired simplices: essential
    classes
        if c not in paired and dim <= maxdim:
            # a simplex whose column reduced to zero creates a class; if never killed it is
    essential
```

```
            out[dim].append((val, np.inf))
    return {k: np.array(v).reshape(-1, 2) for k, v in out.items()}
```

Lines 133–138 build all points, edges and triangles with their entry scale, the longest edge. Line 139 sorts them, faces before the cells they bound at equal scale. In the loop of line 144, line 145 builds the list of faces, lines 146–147 do the additions, and lines 148–154 record a pair when the reduced list is not empty (pairs of zero persistence are dropped). Lines 155–158 add the features that are never killed.

Example: a ring and a clump. For the 52 points of Figure 12.27, the loop diagram has one point far from the diagonal, (0.7001, 1.668): the ring closes at $r \approx 0.7001$, about the largest gap between neighbouring points on it, and fills at $r \approx 1.668$, when triangles can span its middle; the next loop lives for only 0.00431. In the island diagram, the longest finite bar ends at 1.233, the distance at which the clump joins the ring (Figure 12.29). Our reduction, ripser and gudhi's `RipsComplex` give the same loop diagram: the bottleneck distances between them are 2.129×10^{-8} and 2.129×10^{-8}, single-precision rounding again. Our Kruskal deaths equal ripser's island deaths and scipy's single-linkage heights to machine precision.

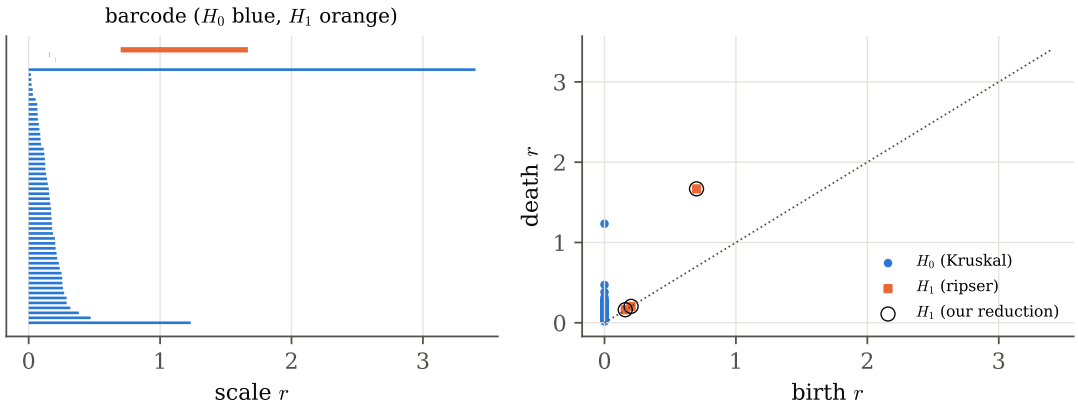


Figure 12.29: Persistence of the ring and clump of Figure 12.27. Left: the barcode; every island bar starts at $r = 0$, and the one long loop bar (orange, top) is the ring. Right: the diagram, with island points on the vertical axis (birth 0) and loop points above the diagonal; the ring's loop, found identically by ripser and by our reduction (open circles), stands far from the diagonal.

Cost. The reduction works on every triangle, and there are $n(n-1)(n-2)/6$ of them: already 23478 cells for 52 points, and our pure-Python reduction takes 0.3344 s on them. It is a teaching code. ripser avoids building most of the triangles and reduces cleverly, and handles 1600 points (islands and loops) in 1.394 s; Kruskal alone, for islands only, takes 0.23 s there. For millions of galaxies in three dimensions, research codes replace the Rips complex by the much smaller alpha complex built on the Delaunay tessellation, which gives the same shapes as growing balls without the explosion of cells [5, 269].

12.13 The cosmic web: clusters, tunnels and voids

The matter of the late universe is arranged in a web: dense clusters at the nodes, filaments joining them, thin walls between filaments, and large nearly empty voids (see Chapter T1 for the objects and the cosmology notes for their physics). The same three-dimensional density contrast $\delta(\boldsymbol{x})$ has already been met twice. In Section T3.6 we measured the genus of its excursion sets in

simulated Gaussian boxes and found it equal to the Gaussian formula Eq. (T3.16): a Gaussian field is a sponge at the median density. In Section 12.5.4 the expected Euler characteristic of a three-dimensional Gaussian field was derived from its critical points. Both results are about one number per threshold, the alternating sum $\chi = \beta_0 - \beta_1 + \beta_2$. What is new here is to split that sum into its three counts and to follow each feature through the thresholds, as Pranav and collaborators proposed for the cosmic web [5] and Wilding and collaborators carried out for ΛCDM simulations [4].

In three dimensions the dry land $X_\nu = \{\delta \geq \nu\}$ (with δ in units of its standard deviation) has three kinds of features. β_0 counts its separate pieces: *clusters*. β_1 counts independent *tunnels*: loops in the land around which a string could be threaded, the holes of a doughnut, formed for example by a ring of filaments around an empty channel. β_2 counts *voids*: pockets of low density completely sealed inside the land, like the hollow of a ball. A sum with signs can hide its terms; two clusters and one hollow ball have the same $\chi = 2$ (Figure T3.14). The separate counts cannot hide them.

Example: three solids. A solid ball has one piece, no tunnel and no sealed pocket: $(\beta_0, \beta_1, \beta_2) = (1, 0, 0)$, $\chi = 1$. A solid doughnut has one piece and one tunnel, through its hole: $(1, 1, 0)$, $\chi = 0$. A hollow ball, a thick shell around an empty pocket, has one piece and one void: $(1, 0, 1)$, $\chi = 2$. A pretzel with two holes has $(1, 2, 0)$ and $\chi = -1$. In each case $\chi = \beta_0 - \beta_1 + \beta_2$, and only the separate counts say whether a negative χ comes from tunnels or a positive one from clusters rather than voids.

The question. As the density threshold is lowered through a Gaussian random field in three dimensions, at which levels are clusters, tunnels and voids most numerous, and do the three Betti curves, combined with their signs, reproduce the Euler characteristic predicted for a Gaussian field?

12.13.1 Which events create and destroy which features

Section 12.5.1 showed that the topology of the dry land changes only when the level passes the height of a critical point, a point where the gradient of the field vanishes, and that the type of the point fixes the type of change. In three dimensions there are four types, told apart by how many of the three principal curvatures of the field are negative (how many directions the field falls away in). Take them in the order in which the falling level meets them.

- A *maximum* (the field falls in all three directions) is the top of a cluster. Passing it creates a new cluster: β_0 increases by one.
- A *filament saddle* (falls in two directions, rises in one) is the lowest point along the spine of a filament between two higher regions. Passing it adds a bridge of land. If the bridge joins two clusters, one of them dies (β_0 decreases); if it joins a cluster to itself, it closes a ring, and a tunnel is born (β_1 increases).
- A *wall saddle* (falls in one direction, rises in two) is the lowest point of a thin sheet stretched across an opening. Passing it adds a membrane of land. If the membrane spans a ring, the tunnel is closed off and dies (β_1 decreases); if it closes the last opening of a pocket, a void is sealed and born (β_2 increases).
- A *minimum* (the field rises in all directions) is the bottom of a void. Passing it fills the void (β_2 decreases).

Figure 12.30 summarises the list. It explains the order in which the three Betti curves rise and fall: clusters are born at maxima, which are mostly high; tunnels at filament saddles, typical of

intermediate levels; voids at wall saddles and they die at minima, mostly low. For a Gaussian field, which is as likely as its negative, the list is symmetric: the voids of δ are the clusters of $-\delta$, so β_2 at level ν should behave on average like β_0 at $-\nu$, and β_1 should be symmetric about $\nu = 0$.

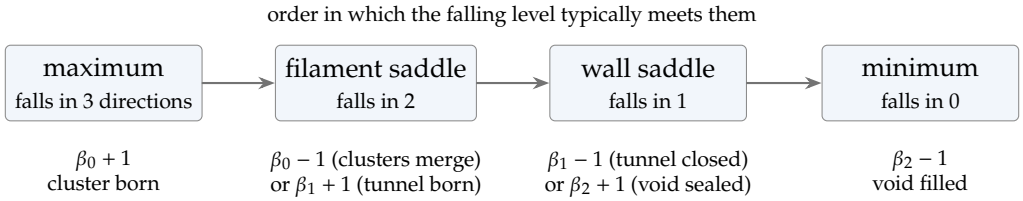


Figure 12.30: The four kinds of critical point of a three-dimensional field and what passing each does to the dry land $\{\delta \geq \nu\}$ as the level falls. Every birth or death in the three persistence diagrams happens at one of them.

12.13.2 Betti curves of Gaussian boxes

`code/ch12/b05_cosmic_web.py` draws 24 periodic Gaussian boxes of 80^3 cells with a power spectrum $P(k) \propto k^{-1.5} e^{-k^2R^2}$, $R = 2.5$ cells (the simulator of Section 7.4 in three dimensions), and computes the persistence of the superlevel sets on the 3-torus with gudhi's `PeriodicCubicalComplex`. The Betti curves are then counted from the diagrams with (12.41). A from-scratch check: on the first box, the number of clusters at $\nu = 1$ and $\nu = 2$ counted by `scipy.ndimage.label` (26 neighbours, the three-dimensional version of 8), with labels glued across the periodic faces, differs from gudhi's by 0.

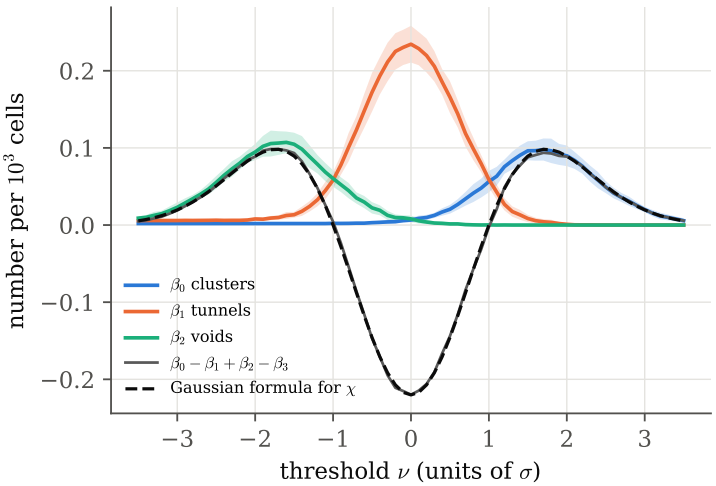


Figure 12.31: Betti curves of three-dimensional Gaussian boxes: clusters (β_0), tunnels (β_1) and voids (β_2) per thousand cells, mean of 24 boxes with the box-to-box scatter shaded. The grey curve is $\beta_0 - \beta_1 + \beta_2 - \beta_3$ (the last term is the single three-dimensional class of the whole torus at the lowest levels) and the dashed curve the Gaussian formula for the Euler characteristic.

Figure 12.31 shows the result. The clusters peak at $\nu = 1.7$ with 0.0979 per thousand cells; the tunnels peak at $\nu = 0$, at 0.2345 per thousand cells, more than twice as many as the other two at their peaks; the voids peak at $\nu = -1.6$ with 0.1072. The order is the one Figure 12.30 predicts, and the symmetry is visible: the void curve is the mirror image of the cluster curve, and the tunnel curve is centred on zero. The alternating sum agrees with the Gaussian formula for χ (Section 12.5.4, Eq. (T3.16)) to within 2.205% of its peak at every threshold, where one Monte Carlo error is up to 2.772%; over all 71 thresholds the largest deviation is 1.8 of its own

Monte Carlo errors. At this smoothing the finite cells, which the formula for a continuous field does not know about, leave no visible trace.

Example: what χ hides at the median density. At $\nu = 0$ the Euler characteristic is -0.22 per thousand cells. Read alone, it says "more tunnels than clusters and voids". The Betti curves say how: at $\nu = 0$ there are 0.00651 clusters and 0.007975 voids per thousand cells, almost none, against 0.2345 tunnels. At $\nu = +1.6$, where χ is 0.09082, and at -1.6, where it has the same size, the curves say which side we are on: at $+1.6$ the positive χ is clusters, at -1.6 it is voids. The Euler characteristic cannot tell these two apart without the sign of ν; for a non-Gaussian field, where the two sides need not mirror each other, only the separate counts can.

Verification and research. The agreement of the alternating sum with the formula verifies the code: it is a known answer reproduced. The Betti curves themselves have no known answer. For Gaussian fields there are no exact expressions for the expected Betti numbers, only approximations and numerical studies [5], and for the non-Gaussian density of the late universe there is nothing analytic at all. Here the Monte Carlo is not a check of a formula but the only instrument we have: the expected curves and their scatter are whatever the simulations say, and part 12c turns exactly these simulated means and covariances into a likelihood.

12.13.3 What gravity does to the diagram

The Gaussian box is the starting point of structure formation, not its result. Section 12.1.3 drew the distinction: a monotone relabelling of the density, such as $\delta \to e^{\delta}$, changes no excursion set and no persistence pair beyond relabelling the levels (we checked this on a map in Section 12.11.3), but gravitational evolution moves matter around, merges structures and empties voids, so the topology itself changes. Wilding and collaborators followed this in ΛCDM N-body simulations, computing the Betti curves and persistence diagrams of the density estimated on the Delaunay tessellation of the particles at redshifts from 3.8 to 0 [4]. Their diagrams grow sharp apexes, and the apex of the cluster diagram sits at the density at which matter detaches from the expansion and begins to collapse; the features near the apex form a self-similar population that reflects the hierarchical growth of structure. The topology there is not a quantity protected from the physics; it is the record of the physics. Reproducing those diagrams needs N-body simulations and the Delaunay density estimate, far beyond laptop scale; Section 12.15.3 reproduces the simpler reference case that the same group used to learn to read such diagrams, a random scatter of points.

12.14 Polarization singularities: a charge that a remapping cannot change

The CMB polarization at a point is a headless stick: a direction without an arrowhead, described by the Stokes parameters Q and U combined into $P = Q + iU = |P|e^{2i\alpha}$, with α the angle of the stick (Eq. (T3.12); see the cosmology notes for where the polarization comes from). Where $Q = U = 0$ the stick has no direction. Walk once around such a point and the stick turns by half a turn, with the walk or against it: the *index* of the singularity is $+\frac{1}{2}$ or $-\frac{1}{2}$ (Section T3.3.1). Section T3.3.2 counted them on simulated maps and Section T3.4 showed, on lensed patches, that gravitational lensing moves them without changing their indices. Treated as data, the singularities are a catalogue of points with charges $\pm\frac{1}{2}$, and the questions are the statistical ones: what exactly protects the charges, how do we measure them on a pixel map, and when does the protection fail? In the language of Section 12.1, lensing is a nuisance operator $\mathcal{N}$ that acts by

remapping the sky, and we want to know precisely which statistics T satisfy $T[\mathcal{N}[P]] = T[P]$ and under which conditions.

The question. If a polarization map is deformed continuously, or remapped by $\boldsymbol{x} \mapsto \boldsymbol{x} + \boldsymbol{d}(\boldsymbol{x})$, which properties of its singularities are guaranteed to survive, which can change, and how do they change?

12.14.1 Why the charge cannot change continuously

Take a closed loop Γ on which the polarization never vanishes, and define the charge inside it as the number of turns of the stick along the loop,

$$q(\Gamma) = \frac{1}{2\pi} \oint_\Gamma \mathrm{d}\alpha = \frac{1}{4\pi} \oint_\Gamma \mathrm{d}\arg P, \tag{12.46}$$

which is the sum of the indices of the singularities inside Γ ([1], their eq. 5). Now let the map change continuously with a parameter t (time, or the strength of a lensing deflection), P_t, and suppose that no singularity crosses Γ while t changes, so $P_t \neq 0$ on Γ throughout.

$$\begin{aligned} q_t(\Gamma) &\overset{(1)}{=} \frac{1}{4\pi} \times (\text{a whole number of turns of } \arg P_t \text{ along } \Gamma) \times 2\pi \in \tfrac{1}{2}\mathbb{Z} \\ &\overset{(2)}{=} \text{a continuous function of } t \\ &\overset{(3)}{=} q_0(\Gamma) \quad \text{for all } t. \end{aligned} \tag{12.47}$$

(1) Going once around Γ brings P_t back to its starting value, so its phase changes by 2π times a whole number, and (12.46) divides by 4π: q is a multiple of $\frac{1}{2}$.
(2) $\arg P_t$ at each point of Γ changes continuously with t because P_t does and never passes through 0 there, so the integral changes continuously.
(3) A continuous function whose values are restricted to the separated numbers $\ldots, -\frac{1}{2}, 0, \frac{1}{2}, \ldots$ cannot jump from one to another, so it is constant.

So the total charge inside a fixed loop can change only when a singularity crosses the loop. Inside the loop, singularities can move, and a $+\frac{1}{2}$ and a $-\frac{1}{2}$ can be created together or annihilate together, because that leaves the total unchanged; a single charge can never appear or vanish on its own [1].

12.14.2 The index of a zero, and what a remapping does to it

The simplest zero first. Near a zero at the origin, Q and U both vanish linearly, so $Q + iU \approx f(z) = c_1 z + c_2 \bar{z}$, with $z = x + iy$ and complex constants c_1, c_2. The real map $(x, y) \mapsto (Q, U)$ has the Jacobian matrix J with rows ∇Q and ∇U. Its determinant is

$$\begin{aligned} \det J &\overset{(1)}{=} Q_x U_y - Q_y U_x \overset{(2)}{=} \operatorname{Im}\left(\overline{f_x}\, f_y\right) \\ &\overset{(3)}{=} \operatorname{Im}\left((\bar{c}_1 + \bar{c}_2)(ic_1 - ic_2)\right) \\ &\overset{(4)}{=} \operatorname{Re}\left(|c_1|^2 - |c_2|^2 + c_1\bar{c}_2 - \bar{c}_1 c_2\right) \overset{(5)}{=} |c_1|^2 - |c_2|^2. \end{aligned} \tag{12.48}$$

(1) The determinant of the 2×2 matrix with rows (Q_x, Q_y) and (U_x, U_y); subscripts are partial derivatives.
(2) With $f_x = Q_x + iU_x$ and $f_y = Q_y + iU_y$, the product $\overline{f_x} f_y = (Q_x Q_y + U_x U_y) + i(Q_x U_y - U_x Q_y)$.
(3) From $f = c_1(x + iy) + c_2(x - iy)$: $f_x = c_1 + c_2$ and $f_y = ic_1 - ic_2$.

(4) Multiply out and use $\mathrm{Im}(iw) = \mathrm{Re}\, w$.
(5) $c_1\bar{c}_2 - \bar{c}_1 c_2$ is a number minus its complex conjugate, which is purely imaginary.

Which way does P turn? If $|c_1| > |c_2|$, write $f = c_1 z\,(1 + w)$ with $w = (c_2/c_1)(\bar{z}/z)$. On a circle around the origin $|w| = |c_2/c_1| < 1$, so $1 + w$ stays in a disc around 1 that does not contain 0 and its phase returns to where it started without turning; the phase of f then turns exactly as z does, once, with the loop. The index is $+\frac{1}{2}$. If $|c_2| > |c_1|$, the same argument with $f = c_2\bar{z}(1 + \dots)$ gives one turn against the loop: index $-\frac{1}{2}$. With (12.48) both cases read

$$\text{index of a zero} = \tfrac{1}{2}\ \mathrm{sgn}\det J. \tag{12.49}$$

This is the rule the plaquette finder of Section T3.3.1 applies implicitly: summing the phase steps around a 2×2 block measures the sign of the local Jacobian.

A remapping. The remapped map is $\tilde{P}(\boldsymbol{x}) = P(\boldsymbol{F}(\boldsymbol{x}))$ with $\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{x} + \boldsymbol{d}(\boldsymbol{x})$, and its zeros are the points $\boldsymbol{x}_0$ with $\boldsymbol{F}(\boldsymbol{x}_0) = \boldsymbol{y}_0$ a zero of P. By the chain rule the Jacobian of $\tilde{P}$ at $\boldsymbol{x}_0$ is $J(\boldsymbol{y}_0)\,\mathsf{A}(\boldsymbol{x}_0)$, where $\mathsf{A} = \mathbb{1} + \partial\boldsymbol{d}$ is the Jacobian of the remapping, the magnification matrix of lensing (Section T3.4.2). Then

$$\begin{aligned}\text{index of } \tilde{P} \text{ at } \boldsymbol{x}_0 &\overset{(1)}{=} \tfrac{1}{2}\ \mathrm{sgn}\det\big(J(\boldsymbol{y}_0)\mathsf{A}(\boldsymbol{x}_0)\big) \overset{(2)}{=} \tfrac{1}{2}\ \mathrm{sgn}\det J(\boldsymbol{y}_0)\ \mathrm{sgn}\det\mathsf{A}(\boldsymbol{x}_0)\\ &\overset{(3)}{=} \mathrm{sgn}\det\mathsf{A}(\boldsymbol{x}_0) \times \big(\text{index of } P \text{ at } \boldsymbol{y}_0\big).\end{aligned} \tag{12.50}$$

(1) The rule (12.49) applied to $\tilde{P}$, whose Jacobian is $J\mathsf{A}$.
(2) The determinant of a product is the product of the determinants.
(3) The rule (12.49) read backwards for P at $\boldsymbol{y}_0$.

Where the remapping keeps orientation, $\det\mathsf{A} > 0$, every zero keeps its index. Where the sky is folded over itself, $\det\mathsf{A} < 0$, the image is a mirror image and the index flips.

Folds create pairs. If $\det\mathsf{A} > 0$ everywhere, the remapping is one-to-one and each zero of P has exactly one image, with the same index. If the displacement is strong enough to fold the sky, a point $\boldsymbol{y}_0$ can have three preimages: one before the fold, one inside it (with $\det\mathsf{A} < 0$), one after it. Their indices are, by (12.50), $+q$, $-q$, $+q$, which add up to q. The one-dimensional cartoon of Figure 12.32 shows why the signs alternate: the curve $F(x)$ must cross the level y_0 going up, then down, then up again. Seen from the remapped sky, a fold has created a pair of opposite charges and kept the total. This is the exact form of the statement that a continuous change can only create or annihilate pairs.

12.14.3 Testing the protection on simulated maps

`code/ch12/b06_pol_remap.py` draws 6 E-mode polarization maps on periodic $10^\circ \times 10^\circ$ patches with 512^2 pixels of 1.172′, smoothed by a 10′ beam, from the fiducial C_ℓ^{EE}; each has about 919 singularities. For each map it draws a smooth random potential, coherent over about a degree, takes $\boldsymbol{d}$ as its gradient, and scales it so that the convergence $\kappa = -\frac{1}{2}\nabla\cdot\boldsymbol{d}$ has a chosen root-mean-square value κ_{rms}, from 0.03 (the size of real CMB lensing on these scales) up to 0.8. The remapped maps $\tilde{Q}(\boldsymbol{x}) = Q(\boldsymbol{x} + \boldsymbol{d})$ and $\tilde{U}(\boldsymbol{x}) = U(\boldsymbol{x} + \boldsymbol{d})$ are computed by cubic interpolation. (On the curved sky the polarization must also be carried from the displaced point by parallel transport [3], Section T3.4.1; on a flat patch that step is the identity.) For every singularity of the remapped map the code follows $\boldsymbol{d}$ to the point $\boldsymbol{x} + \boldsymbol{d}(\boldsymbol{x})$ and looks there for a singularity of the original map within a pixel and a half; it checks the index rule (12.50) with the sign of $\det\mathsf{A}$ at that point; and for every original singularity it adds up the indices of all its images and compares with its own index.

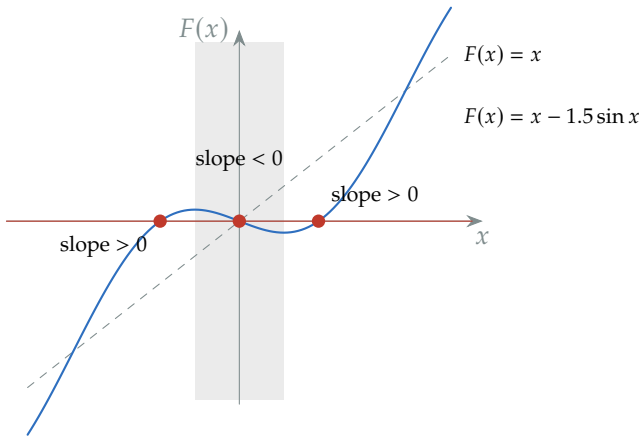


Figure 12.32: A one-dimensional remapping $F(x) = x - 1.5\sin x$ that folds: in the shaded strip its slope is negative. The level $F = 0$ (red line), the position of one singularity of the original map, is reached at three points (dots) with slopes of signs $+, -, +$. The three images carry indices $+q, -q, +q$, which add up to the original q. The dashed line is the identity map, which has a single crossing.

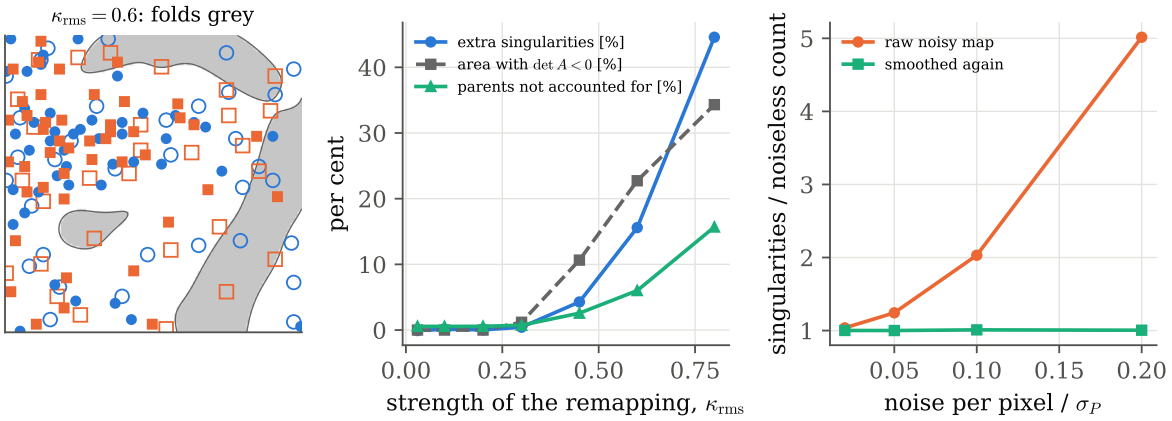


Figure 12.33: Left: a $2.5° \times 2.5°$ corner of one map remapped with $\kappa_{\rm rms} = 0.6$; grey regions are folded ($\det A < 0$). Open markers: singularities of the original map; filled: of the remapped map (circles $+\frac{1}{2}$, squares $-\frac{1}{2}$). Middle: against the strength of the remapping, the fraction of extra singularities, the folded area, and the fraction of original singularities whose images do not add up to their index. Right: noise is not a remapping; white noise added to the map multiplies the number of singularities (orange), and smoothing the noisy map again removes the extra ones (green).

Example: realistic lensing strength. At $\kappa_{\rm rms} = 0.03$ no point of any map is folded. The remapped maps have the same number of singularities as the originals (the change is -0.07254%); 99.82% of them sit on an original singularity at the far end of their displacement, and 99.66% also carry the index the rule (12.50) predicts. The few exceptions are zeros that fall at the corner shared by two 2×2 blocks, where the pixel finder can put them in either block. This is the result of Section T3.4.4 seen from the other side: the catalogue of charges is unchanged, point by point, up to the resolution of the grid.

Example: folding the sky. The first folds appear at $\kappa_{\rm rms} \approx 0.3$ (Figure 12.33, middle). At $\kappa_{\rm rms} = 0.6$, 22.73% of the area is folded and the remapped maps have 15.6% more singularities than the originals, all of them created in pairs, as the left panel shows: inside and beside the grey regions, one original singularity has three images, the middle one with the opposite index. For 93.99% of the original singularities the indices of the images still add up to the

original index, and 95.48% of the images obey the sign rule. The failures grow with the folding because the remapped field is squeezed near the fold lines, and some of the created pairs are closer together than a pixel: the grid can no longer resolve them. Real CMB lensing, with a convergence of a few hundredths, is about ten times too weak to fold the sky; the experiment shows what the protection rests on, and where it would end.

Example: noise is not a remapping. White noise added to Q and U in each pixel, at a level of 5% of the polarization amplitude, raises the number of singularities by a factor 1.243; at 20%, by a factor 5.015, and only 37.05% of the singularities of the noisy map lie on one of the original map. The noise creates pairs on the scale of a pixel, wherever the polarization is weak. Smoothing the noisy map over two pixels removes almost all of them: the count returns to 1.005 times that of the equally smoothed clean map, and 94.29% of the singularities match. The order matters, as Section T3.4.3 found for the beam: a smoothing that acts on the remapped, noisy sky is part of the nuisance, and its effect must be measured by simulating it.

12.14.4 What is protected, and what is not

Putting the pieces together, for a map changed by a smooth remapping or any continuous deformation:

- *Protected:* the total charge inside any loop that no singularity crosses (12.47); and, while the remapping is one-to-one, the whole catalogue, singularity by singularity, with its indices (12.50).
- *Not protected:* the positions (they move with the remapping); the number of singularities once folds appear (pairs are created); the charge inside a loop that singularities cross, which is the situation at the edge of any finite patch or mask; pairs closer than the resolution, which a beam or a pixel grid can merge or hide; and anything that noise touches, because noise is not a deformation of the map but an addition to it.
- *A non-local step breaks it:* the E-mode and B-mode maps are built from Q and U by a non-local filter, so the singularities of an E-only map are not those of the full map and are not protected (Section T3.4.3).

This is the caution of Section 12.1.2 made concrete: a topological statistic is protected only against a named class of transformations, and every analysis must say which class its nuisance belongs to.

12.15 Reproducing the literature

Three papers carry the quantitative claims about persistence and singularities: Huterer and Vachaspati counted the polarization singularities of simulated CMB maps and measured how their charges are arranged [2], testing a prediction of Vachaspati and Lue [1]; and Pranav and collaborators built the reference Betti curves and diagrams against which the cosmic web is read [5]. We redo each of their measurements with our own code, starting from what they asked and what they assumed, deriving the equation behind the number they plot, and then putting their numbers next to ours and tracing every difference to a cause.

12.15.1 How many singularities: Huterer and Vachaspati's counts

Their question. In 2004 the first CMB polarization maps were arriving, and Huterer and Vachaspati asked what the singularities of a polarization map from the standard cosmological

model should look like, so that the predictions would be ready when the data were. The first number they report is the simplest: how many singularities a full-sky map has, and how that grows with the angular resolution of the map.

Their setup. They computed the spectra of a flat ΛCDM model with CMBFAST, drew Gaussian E and B coefficients from them up to a maximum multipole $\ell_{\max}$, made HEALPix maps at $N_{\rm side} = 256$ and found the singularities with the winding rule of Section T3.3.1. With no tensor modes and no lensing, $C_\ell^{BB} = 0$, so the maps are pure E. Their parameters, as stated in the paper:

quantity	value	meaning
Ω_m	0.3	matter fraction of the critical density
$\Omega_m h^2$, $\Omega_b h^2$	0.127, 0.021	physical matter and baryon densities ($h = 0.6506$ follows)
n_s	1	primordial spectral index (scale-invariant)
tensors, lensing	none	so $B = 0$
optical depth τ	not stated	sets the large-scale reionization signal in C_ℓ^{EE}

They found about 6000 singularities for $\ell_{\max} = 100$ and about 150 000 for $\ell_{\max} = 500$, growing "roughly as the square of" $\ell_{\max}$.

Their key equation, here. The paper gives no formula; the counts come from the maps. But the expected count follows from the density formula derived in Eq. (T3.13), $n = \sigma_1^2/(4\pi\sigma_0^2)$ per steradian, with $\sigma_0^2 = \langle Q^2\rangle$ and $\sigma_1^2 = \langle|\nabla Q|^2\rangle$. On a flat patch with pure E modes, a Fourier mode $\boldsymbol{\ell}$ of angle $\varphi_{\boldsymbol{\ell}}$ contributes $Q(\boldsymbol{\ell}) = E(\boldsymbol{\ell})\cos 2\varphi_{\boldsymbol{\ell}}$ (see the cosmology notes for this relation), and a gradient brings down a factor $i\boldsymbol{\ell}$. So

$$\frac{\sigma_1^2}{\sigma_0^2} \overset{(1)}{=} \frac{\int \frac{\mathrm{d}^2\ell}{(2\pi)^2}\,\ell^2\,C_\ell^{EE}\cos^2 2\varphi_{\boldsymbol{\ell}}}{\int \frac{\mathrm{d}^2\ell}{(2\pi)^2}\,C_\ell^{EE}\cos^2 2\varphi_{\boldsymbol{\ell}}} \overset{(2)}{=} \frac{\int_0^{\ell_{\max}} \ell^3 C_\ell^{EE}\,\mathrm{d}\ell}{\int_0^{\ell_{\max}} \ell\, C_\ell^{EE}\,\mathrm{d}\ell} \equiv \langle\ell^2\rangle_E,$$

$$N_{\rm sky} \overset{(3)}{=} 4\pi\, n \overset{(4)}{=} \langle\ell^2\rangle_E \;\longrightarrow\; \frac{\sum_\ell (2\ell+1)\,\ell(\ell+1)\,C_\ell^{EE}}{\sum_\ell (2\ell+1)\,C_\ell^{EE}}. \tag{12.51}$$

(1) The variance of a Gaussian field is the integral of its spectrum over all modes; for the gradient each mode is weighted by $|i\boldsymbol{\ell}|^2 = \ell^2$.

(2) Write $\mathrm{d}^2\ell = \ell\,\mathrm{d}\ell\,\mathrm{d}\varphi$. The angular integral of $\cos^2 2\varphi$ is π in both numerator and denominator and cancels, as do the $(2\pi)^2$; the spectrum is cut at $\ell_{\max}$.

(3) A full sky is 4π steradians.

(4) Insert $n = \sigma_1^2/(4\pi\sigma_0^2)$. The arrow is the same average written on the sphere, Eq. (T3.14), where a ring of $\mathrm{d}\ell$ in the flat plane, of circumference $2\pi\ell$, becomes the $2\ell+1$ modes of each multipole, and ℓ^2 becomes the eigenvalue $\ell(\ell+1)$ of the Laplacian.

The number of singularities on the sky is the power-weighted average of $\ell(\ell+1)$ below $\ell_{\max}$. Because C_ℓ^{EE} rises with ℓ up to the first acoustic peaks, most of the weight sits near the cut, the average is a fixed fraction of $\ell_{\max}^2$, and that is the "roughly as the square".

What we compute. `code/ch12/b07_hv_counts.py` runs CAMB for their model twice, once with an optical depth $\tau = 0.06$ and once with reionization switched off, since their τ is not given; evaluates (12.51) for $\ell_{\max}$ from 50 to 500; and counts the singularities directly on 12 simulated $30^\circ \times 30^\circ$ periodic patches at each $\ell_{\max}$, with at least four pixels per half wavelength, scaling the counts to the full sky. The grid version of (12.51), the same average over the discrete Fourier modes of the patch, matches the patch counts to 4.663% at every $\ell_{\max}$, at most 2.3 times the Monte Carlo error of 12 patches. The same formula with the fiducial *Planck* spectra (Section 7.11.1) repeats the numbers of Section T3.3.3.

Side by side.

$\ell_{\max}$	H&V 2005	(12.51), $\tau = 0.06$	(12.51), no reionization	our patches	fiducial
100	≈ 6000	5245	6082	6142 ± 127	5311
500	$\approx 150\,000$	138531	139047	137854 ± 513	133373

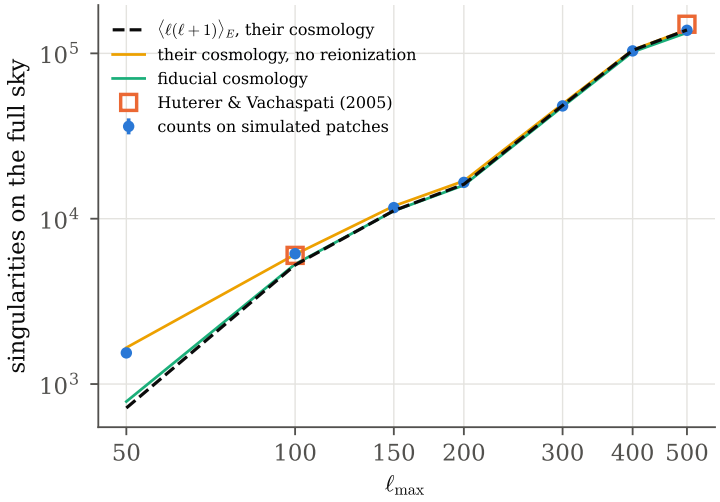


Figure 12.34: Number of polarization singularities on the full sky against the maximum multipole of the map. Dashed: the formula (12.51) for their cosmology with $\tau = 0.06$; orange: the same without reionization; green: the fiducial *Planck* cosmology; dots: direct counts on simulated patches scaled to the full sky; squares: the two numbers of Huterer and Vachaspati. The curves differ only at low $\ell_{\max}$, where the reionization signal at $\ell < 12$ carries a large share of the variance.

Why they differ. At $\ell_{\max} = 100$ the formula gives 5245 with $\tau = 0.06$ and 6082 without reionization: the reionization bump of C_ℓ^{EE} at $\ell < 12$ adds variance σ_0^2 but almost no gradient variance σ_1^2, so it lowers the count by 14%. The published 6000 is the no-reionization number. Our patches agree with that number for a different reason: a 30° patch cannot hold modes with ℓ below about $360°/30° = 12$, so it does not contain the bump at all (Figure 12.34; at $\ell_{\max} = 50$ the patches follow the orange curve, 1659, far from the 716 of the dashed one). At $\ell_{\max} = 500$ the bump no longer matters and the three numbers for their cosmology (formula with and without reionization, and patches) agree to about one per cent; the published count is 8% above them. We cannot trace that difference from the paper: their value is quoted as a round number, their spectrum came from a different Boltzmann code, and their counting on HEALPix pixels at $N_{\text{side}} = 256$ samples an $\ell_{\max} = 500$ map with only about three pixels per wavelength. The growth from 100 to 500 is a power $\ell_{\max}^{2.123}$ in our formula, close to their "square".

The lesson. Counting singularities measures one moment of the polarization spectrum, $\langle \ell(\ell+1) \rangle$, and that moment is sensitive to whatever adds large-scale variance, here the reionization of the universe, which has nothing to do with the small-scale structure where the singularities live. The count itself carries no information beyond C_ℓ^{EE}; any new information must be in how the charges are arranged, which the charge statistics below probe.

12.15.2 How the charges are arranged: Vachaspati and Lue's exponent

Their question. Vachaspati and Lue noticed that a polarization map is a field of headless rods, like the molecules of a nematic liquid crystal, and that the theory of defects in condensed matter predicts how the defects of such a field are arranged [1]. The prediction concerns the net charge inside a large loop: it fluctuates around zero, and its root-mean-square value should grow as the square root of the loop's length. Huterer and Vachaspati measured the exponent on simulated maps [2].

Their key equation, derived. Take a loop Γ of length L, much longer than the correlation length ξ of the stick directions. Cut it into $m = L/\xi$ stretches of length ξ, and let $\delta\alpha_k$ be the change of the stick angle along stretch k. By (12.46) the charge inside is $q = \frac{1}{2\pi}\sum_k \delta\alpha_k$. The increments have zero mean (no direction of turning is preferred), and increments along stretches farther apart than ξ are independent; call their common variance s^2. Then

$$\sigma^2(q) \overset{(1)}{=} \frac{1}{4\pi^2}\,\mathrm{Var}\Big(\sum_{k=1}^{m}\delta\alpha_k\Big) \overset{(2)}{=} \frac{1}{4\pi^2}\sum_{k=1}^{m}\mathrm{Var}(\delta\alpha_k) \overset{(3)}{=} \frac{m\,s^2}{4\pi^2} = \frac{s^2}{4\pi^2}\,\frac{L}{\xi},$$

$$\sigma(q) = \frac{s}{2\pi}\Big(\frac{L}{\xi}\Big)^{1/2} \qquad (\nu = \tfrac{1}{2}). \tag{12.52}$$

(1) The variance of a constant times a random variable is the constant squared times its variance.
(2) The variance of a sum of independent terms is the sum of their variances (the covariances vanish).
(3) All m terms have the same variance s^2, and $m = L/\xi$.

This is [1], their eq. 6, with $\beta = \frac{1}{2}$. *What the exponent means.* Compare a sky on which the singularities were scattered at random with random signs, with density n. The charge inside a circle of circumference L, area $A = L^2/4\pi$, would then be a sum of $\langle N\rangle = nA$ independent charges $\pm\frac{1}{2}$:

$$\sigma^2(q) \overset{(1)}{=} \langle N\rangle\ \mathrm{Var}(\pm\tfrac{1}{2}) \overset{(2)}{=} \frac{nA}{4} \overset{(3)}{=} \frac{n\,L^2}{16\pi}, \qquad \sigma(q) \propto L \quad (\nu = 1). \tag{12.53}$$

(1) For a Poisson number of independent terms with zero mean, the variance of the sum is the mean number times the variance of one term.
(2) A charge $\pm\frac{1}{2}$ with equal probabilities has variance $\frac{1}{4}$.
(3) $A = \pi R^2$ and $L = 2\pi R$ give $A = L^2/4\pi$.

So $\nu = \frac{1}{2}$ is the signature of charges that are not independent: a $+\frac{1}{2}$ almost always has a $-\frac{1}{2}$ nearby, and only the pairs that the loop cuts in two contribute to the net charge. Their number grows with the length of the loop, not with its area. A measured ν near $\frac{1}{2}$ says "the charges are screened"; near 1, "they are not".

Their setup. On a HEALPix map, the pixels of a ring at colatitude θ form a loop of length $L = 2\pi\sin\theta$ around the north pole; the winding of the polarization angle along the ring gives the charge of the polar cap (with a correction of one unit for the pole, where the local axes themselves turn once; it is the same for every map and drops out of the spread). They rotated one map in a hundred random directions, measured $\sigma(q)$ ring by ring, and fit $\sigma \propto L^{\nu}$ over rings plotted at $L = 0.01, 0.03, \ldots, 0.99$; values from 0.01 to 1 in equal steps match $\sin\theta$, the ring length divided by 2π, rather than the length itself, and we read their axis that way. They found $\nu = 0.54$ for $\ell_{\max} = 100$ and $\nu = 0.45$ for $\ell_{\max} = 500$, each with an error of about 0.01. They also measured the distance from each singularity to its nearest neighbour of the same and of the opposite index, with means of 3.24° and 2.69° at $\ell_{\max} = 100$.

What we compute. `code/ch12/b08_charge_stats.py` repeats their ring measurement on 100 independent full-sky maps at $N_{\mathrm{side}} = 256$, their resolution, with their spectrum (without reionization, which matched their counts in Section 12.15.1), at the 50 rings with $\sin\theta = 0.01, 0.03, \ldots, 0.99$. The error on the fitted exponent comes from refitting it many times on maps drawn at random, with repeats, from the 100 (a bootstrap, Section 3.15). On periodic 40° flat patches it also catalogues the singularities themselves, measures the charge inside circles of circumference L with the true indices and with the indices shuffled among the singularities (which turns the catalogue into the random-sign sky of (12.53) at the same positions), and measures nearest-neighbour distances.

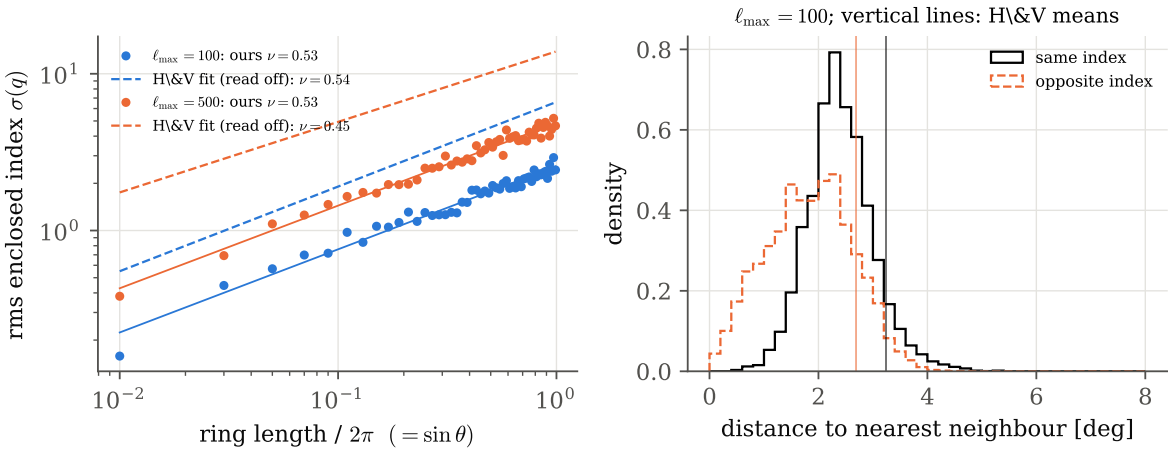


Figure 12.35: Left: root-mean-square charge of the polar caps against the ring length divided by 2π, for maps cut at $\ell_{max} = 100$ and 500 (dots, with power-law fits), and the fitted laws of Huterer and Vachaspati read off their figure (dashed). The slopes are close; the normalisations differ by a factor of two to three. Right: distance from each singularity to its nearest neighbour of the same index and of the opposite index at $\ell_{max} = 100$; vertical lines are their two means.

Side by side.

	H&V 2005	ours
ν, $\ell_{max} = 100$ (rings)	0.54 ± 0.01	0.5305 ± 0.01527
ν, $\ell_{max} = 500$ (rings)	0.45 ± 0.01	0.5273 ± 0.009313
ν with indices shuffled, $\ell_{max} = 100$ / 500 (patches)	–	0.8951 / 0.9041
$\sigma(q)$ at $\sin\theta = 1$, $\ell_{max} = 100$ / 500	≈ 7 / ≈ 14 (read off)	2.436 / 4.656
nearest neighbour, same index	3.24°	2.404°
nearest neighbour, opposite index	2.69°	1.839°
ratio same / opposite	1.20	1.307

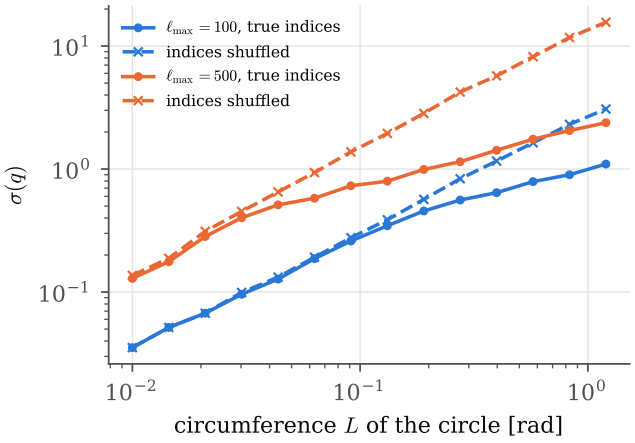


Figure 12.36: Charge inside circles on flat patches, with the true indices (dots) and with the indices shuffled among the same positions (crosses). For small circles, which hold at most one singularity, the two agree. For circles much longer than the distance between a $+\frac{1}{2}$ and its partner, the true charge grows as about $L^{1/2}$ and the shuffled one as about L.

Why they differ. At $\ell_{max} = 100$ the two exponents differ by 0.5 combined standard errors. At $\ell_{max} = 500$ ours is 0.08 higher, 5.7 combined standard errors. Two things make the published

error bar too small to judge either difference. Their hundred ring sets come from rotations of one map, so they share that map's particular arrangement of charges, and the error of 0.01 does not include the scatter from one sky to another, which ours does. And $\sigma(q)$ is not an exact power of L: on small rings the cap holds at most one singularity, and (12.53) applies, with ν near 1; only on rings much longer than the pair separation does (12.52) take over (Figure 12.36). The fitted slope depends on how much of each regime the fit range contains, and it is noisy: on the patches, where we fit only circles longer than 0.2 radians, the slope from one set of simulations to another scatters by about ± 0.05. The normalisation is a separate matter. On the smallest ring, $\sin\theta = 0.01$, the polar cap has an area of about $\pi \times 10^{-4}$ steradians, and with 6082 singularities on 4π steradians it holds on average $6082 \times 10^{-4}/4 \approx 0.15$ of them, so the charge there, made of independent single charges, has a spread of about $\frac{1}{2}\sqrt{0.15} \approx 0.19$ in units of the fundamental charge $\frac{1}{2}$, by (12.53); we measure 0.1581. Their figure shows about 0.4 there, and the ratio stays between two and three at all lengths, so their plotted charge must be in a different unit, which the paper does not state; a constant unit does not affect the exponent. Finally, their nearest-neighbour distances are both about 1.4 times ours, while the ratio of the two, the measure of attraction against repulsion, is close (1.20 against 1.307). Our distances are consistent with our count: for random points at our density the mean nearest-neighbour distance would be 1.325°, and we find 1.684° to the nearest singularity of either index. Their own text quotes a mean nearest-neighbour distance "slightly above 2°" for about 6000 singularities, where our maps with the same number give 1.684°, so their absolute distances belong to a sparser map than the one described, and cannot be compared with ours; the qualitative result, opposite charges closer than equal ones, is reproduced.

The lesson. The exponent ν is a slope fitted over a range of loop sizes on which the charge spread is not an exact power law; its second decimal place depends on the range and on how many independent skies are used. What is robust is the contrast between $\nu \approx \frac{1}{2}$, screened charges born in pairs, and $\nu \approx 1$, independent charges, and the shuffling test shows that contrast directly.

A third look: the angular correlation function. The charge spread looks at the net charge in a big loop, and the nearest-neighbour distance at one singularity and its closest partner. A third question puts the two together: at a separation θ, are there more singularities, or fewer, than points thrown down at random would give, and does the answer depend on the signs of the two charges? Huterer and Vachaspati answer with the angular correlation function $w(\theta)$, borrowed from the clustering of galaxies. Its definition is a statement about pairs. If the singularities were independent of each other, the chance that a given one has a second within a ring of separations $[\theta_1, \theta_2]$ around it would just be the share of the sky inside the ring, $f = \frac{1}{2}(\cos\theta_1 - \cos\theta_2)$ on a sphere, or $f = \pi(\theta_2^2 - \theta_1^2)/A$ on a flat periodic patch of area A. We then write the true chance as $f\,(1 + w)$: $w = 0$ is random, $w < 0$ is repulsion, $w > 0$ is attraction (their eqs. 4–5). Counting over all pairs,

$$\text{pairs found in the ring} \overset{(1)}{=} \frac{N(N-1)}{2}\, f\,\big(1 + w(\theta)\big),$$
$$w(\theta) \overset{(2)}{=} \frac{DD(\theta)}{RR(\theta)} - 1, \qquad RR \equiv \frac{N(N-1)}{2}\, f. \tag{12.54}$$

(1) There are $N(N-1)/2$ pairs among N singularities, and each lands in the ring with chance $f(1 + w)$.

(2) DD is the number of pairs in the data that fall in the ring and RR the number a random scatter would give; solving for w gives the Peebles–Hauser estimator [2] (their eq. 6, where the two numbers of points differ and so a ratio $N_{\text{rand}}/N_{\text{map}}$ appears). For pairs of equal charge the $N(N-1)/2$ is replaced by the number of such pairs, $N_+(N_+-1)/2 + N_-(N_--1)/2$,

and for opposite charges by N_+N_-.

With no mask there is no selection function to model, and RR is exact; the estimator needs no random catalogue. `code/ch12/b09_wtheta.py` computes (12.54) on 60 periodic 40° patches at $\ell_{\max} = 100$ in bins of 2.5° (Figure 12.37).

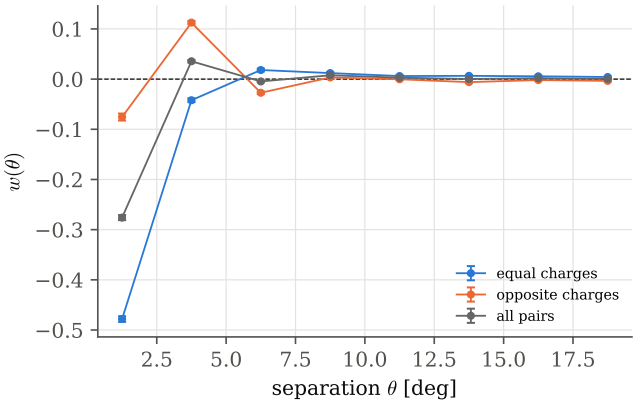


Figure 12.37: Angular correlation function of the singularities of E-mode maps cut at $\ell_{\max} = 100$, mean of 60 flat patches with the error of the mean, for all pairs and for pairs of equal and of opposite charge. Closer than about 5° equal charges are strongly deficient and opposite charges are in excess; beyond about 10° all three curves are flat at zero.

In the first bin, closer than 2.5°, equal charges are almost absent ($w = -0.4781 \pm 0.005635$) and opposite charges are also deficient but much less so (-0.07559 ± 0.007006); in the next bin, 2.5° to 5°, which is where the typical distance to the nearest partner sits, opposite charges are in excess and equal charges are still deficient. Beyond 10° every bin has $|w| \leq 0.006497$: the singularities are placed as if at random. (The tiny positive values there are the counterpart of the deficit at short range, since the pair counts of all bins must add up to $N(N-1)/2$.) This is their result: attraction of opposite and repulsion of equal charges below about 10°, randomness above. Their bins have a half-width of 5°, four times the width of ours, so they cannot resolve the dip below 2.5° that the finer bins show; the two agree on the sign of the effect and on the scale, about 10°, at which it ends. On the physics: the correlation length of the polarization stick at $\ell_{\max} = 100$ is a few degrees, so there is nothing to correlate beyond it. They also repeated all these measurements for maps with a large tensor fraction, a very tilted primordial spectrum ($n_s = 0$ and 2) and strongly skewed, non-Gaussian $a_{\ell m}$, and found ν and $w(\theta)$ unchanged within errors, while the number of singularities changed with the total power; only a grossly anisotropic spectrum altered the statistics [2]. We did not repeat these variations. The pairing of charges is a property of any smooth random headless-vector field, not of a particular cosmology, and that is why these statistics hardly respond to the cosmological model.

12.15.3 Betti curves of shot noise: Pranav and collaborators' baseline

Their question. To read the Betti numbers and persistence diagrams of the real cosmic web one needs reference cases where the answer is understood. Pranav and collaborators built such templates from simple point processes [5]. They point out that even for a Gaussian random field no exact expressions for the Betti numbers or the persistence diagrams exist, so the templates must be computed. The first template is the most featureless one: points scattered uniformly at random, a Poisson process, whose topology is pure shot noise.

Their setup. They put 500 000 particles at random in a periodic box of side $200\,h^{-1}$Mpc, estimated the density at every particle with the Delaunay tessellation field estimator (DTFE,

which uses the volume of the tetrahedra around each particle), and computed the persistence of the superlevel sets. They plotted the Betti curves against the density threshold divided by the standard deviation of the density, and reported that β_0 peaks at a normalised level of about 1.8, β_1 at 0.6 and β_2 at 0.3; that the three peaks are well separated, so the topology is dominated in turn by clusters, tunnels and voids; and that for different mean separations of the particles the normalised peak positions coincide.

Their key relation. The Betti number at level ν is the number of diagram points with $b \geq \nu > d$ (their §5.3), which is (12.41); and the order of the peaks follows from Figure 12.30.

What we compute. The Poisson part of `code/ch12/b05_cosmic_web.py` puts particles at random in a periodic box of 64^3 cells with mean separations of 2, 3 and 4 cells (three realisations each), estimates the density on the grid from the distance to the 5th nearest particle, $\rho = 5/(\frac{4}{3}\pi r_5^3)$, which adapts to the local density as the DTFE does, and computes the Betti curves of ρ/σ_ρ with gudhi's periodic cubical complex.

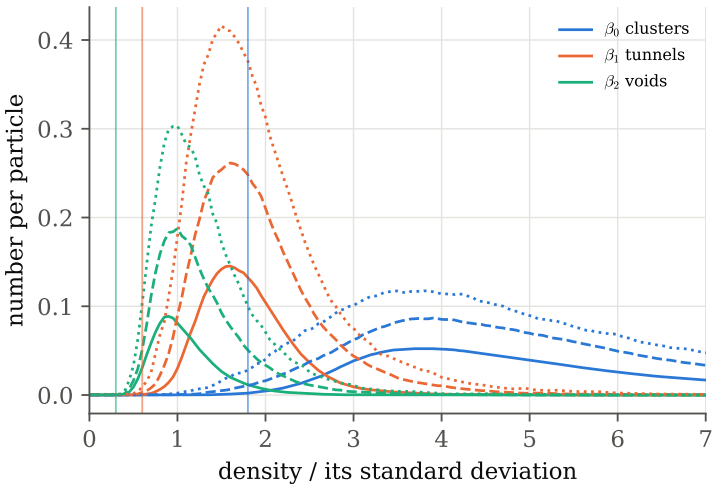


Figure 12.38: Betti curves of the density of randomly placed points, per particle, against the density divided by its standard deviation, for mean separations of 2 (solid), 3 (dashed) and 4 (dotted) cells. Vertical lines: the peak positions reported by Pranav and collaborators for their Delaunay density estimate. The order of the peaks and their coincidence across separations are reproduced; the positions themselves depend on the estimator.

Side by side.

	Pranav et al. 2017 (DTFE)	ours (nearest neighbours)
peak of β_0 (clusters)	≈ 1.8	3.833
peak of β_1 (tunnels)	≈ 0.6	1.567
peak of β_2 (voids)	≈ 0.3	0.9667
order of the peaks	$\beta_0, \beta_1, \beta_2$	$\beta_0, \beta_1, \beta_2$
peak positions for different separations	coincide	coincide within 0.1
Gaussian field: exact Betti curves	none known	Monte Carlo, Figure 12.31

Why they differ. The levels are absolute densities in units of the standard deviation, not deviations from the mean, so the peak positions depend on the ratio of the mean density to its spread, and that ratio belongs to the estimator, not to the points. The DTFE gives very high densities to particles closely surrounded by others and very low ones in empty regions, a broad distribution with a long tail; Pranav and collaborators note its tendency to form spikes at clusters. Our nearest-neighbour estimate on a grid averages over five particles and is much narrower, so in units of its own spread its mean sits higher and all three peaks move up.

The order of the peaks and their independence of the particle density do not depend on the estimator, and both are reproduced. The coincidence across separations is not an accident: with the estimator tied to the mean separation, the density field of a Poisson process looks the same in units of that separation, whatever its value.

The lesson. A Betti curve of a point set describes the points and the estimator together. Two analyses can only be compared, or a measurement compared with a model, if the simulated points go through the same density estimate as the data: the same rule as for the beam and the noise in Section 12.10.

12.16 How it is done in research

The Gaussian patches, boxes and random points of the experiments are scaled-down versions of analyses that are run in research. The steps are the same; the data, the sizes and the care taken at each step are not. Figure 12.39 lays out the pipeline as a real analysis runs it. Under each step of the pipeline we say what a real analysis involves, what we did instead, and what the reduction costs.

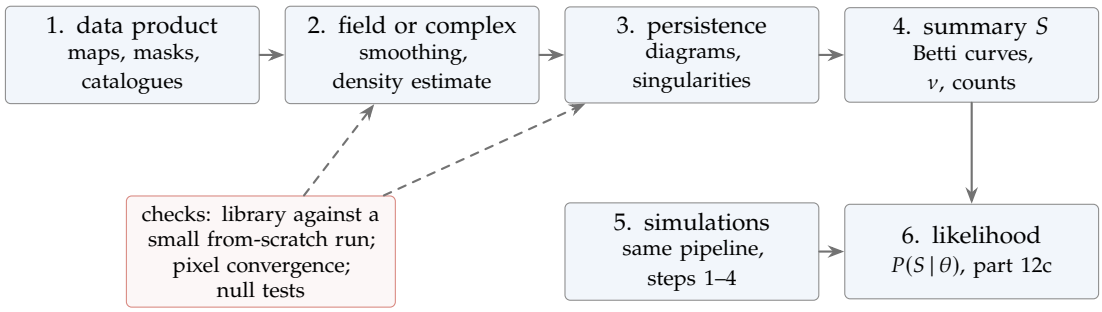


Figure 12.39: The pipeline of a persistence or singularity analysis. The data and every simulation go through exactly the same steps 1–4; the simulations then give the sampling distribution of the summary, which part 12c turns into a likelihood. The checks (red) guard the two steps where conventions are easiest to get wrong.

1. Data products. For CMB polarization the inputs are the component-separated Q and U maps of *Planck* (HEALPix, $N_{\mathrm{side}} = 2048$), or arcminute-resolution maps from ground-based telescopes over part of the sky, together with the mask of the Galaxy and point sources and a large set of end-to-end simulations of the instrument's noise and processing (about a thousand for *Planck*). For the cosmic web the inputs are galaxy catalogues with positions and redshifts, or particle snapshots of N-body simulations; Wilding and collaborators used five ΛCDM runs, each with 256^3 particles in a box of 300 h^{-1}Mpc, at several redshifts [4]. *Ours:* Gaussian simulations only, flat patches of 10° to 40° and full-sky maps at $N_{\mathrm{side}} = 256$, periodic Gaussian boxes of 80^3 cells and Poisson points on a 64^3 grid. *Cost:* no real noise, no mask, no non-Gaussian structure; these are the baselines a real analysis compares against, not the analysis.

2. From data to a field or a complex. Maps are smoothed to a common resolution chosen so that the pixel noise is well below the field on that scale; Section 12.10 showed why: the stability theorem bounds the damage by the largest noise value, and smoothing is what brings that value down. Particles are turned into a density by the Delaunay tessellation field estimator, or the Delaunay tessellation itself becomes the complex (an alpha complex) [5]. *Ours:* Gaussian smoothing on a grid and a nearest-neighbour density. *Cost:* the peak positions of the Betti

curves move with the estimator (Section 12.15.3), so our curves are templates of the same kind as the published ones, not the same templates.

3. Persistence and singularities. On the sphere the cubical grid is replaced by HEALPix pixels with their neighbours from `healpy.get_all_neighbours`, and persistence is computed with gudhi or a dedicated cubical code; for N-body data, by the reduction of the Delaunay complex with PHAT [4, 5]. Polarization singularities are found on HEALPix by following the polarization angle around small circuits of neighbouring pixels, with the polarization at neighbouring pixels carried to a common tangent plane by parallel transport before the angles are compared [1], and with the correction at the poles of the coordinate system [2]. The pixel size must resolve the map: Huterer and Vachaspati require $N_{\rm side} > \ell_{\rm max}/4$. *Ours:* flat periodic grids, where none of these geometric corrections arise, and HEALPix rings only for the ring charges. *Cost:* none in the statistics, which are local; but every geometric step we skipped is a place where a real analysis can go wrong.

4. The summary. A diagram is a set of points of varying number, which a likelihood cannot take directly. The standard summaries are the Betti curves on a fixed grid of levels, histograms of the diagram points in birth–death bins (the "intensity maps" of [5]), or smooth versions of them; for singularities, the counts by index, the charge spread $\sigma(q)$ on a grid of loop sizes, nearest-neighbour statistics and the angular correlation function of the singularities [2]. Each is a fixed-length vector S.

5. Simulations through the same pipeline. Every effect we cannot model analytically, the mask cutting features at its edges, the noise creating short bars and pairs of singularities, the beam merging them, the estimator, is put into the simulations and goes through steps 1–4 unchanged. The simulated summaries give the mean and the covariance of S under each model. *Ours:* a few dozen to a hundred realisations per measurement, each script under ten minutes on a laptop. *Cost:* Monte Carlo errors of a few per cent on means, as quoted with each result, and too few realisations for an accurate covariance of a long summary vector, which part 12c addresses.

Checks. Three checks recur in every serious analysis, and each appeared here. A library run is compared with a from-scratch run on a small map (Section 12.11.3), which catches mixed conventions. The pixel or particle resolution is changed and the summary must not move (the count of singularities against pixel size, Section 12.15.1). And null tests destroy the structure being measured while keeping everything else: shuffling the indices among fixed positions (Figure 12.36) is one; rotating the map, or flipping the sign of a Gaussian field, are others.

Recap

- Lowering a threshold through a field turns its excursion sets into a filtration. Each island, lake, tunnel or void has a birth level and a death level (Section 12.9); with the elder rule the record is unique. The persistence diagram is the set of pairs, the barcode draws them as intervals, and the Betti number at any level is the number of pairs alive there (12.41). Persistence $b - d$ is the range of thresholds over which a feature exists.
- Stability: if the field changes by at most ϵ everywhere, the diagram moves by at most ϵ in the bottleneck distance; features created by the change have persistence at most 2ϵ and sit by the diagonal (Section 12.10). Verified on noisy maps: the ratio never exceeded 0.9132.

- Islands are computed by union–find with the elder rule, lakes by the same code on $-\phi$ with a sea around the map (Alexander duality in the plane, not on the torus); our code and gudhi, ripser, scipy and scikit-image agree exactly (Section 12.11).
- For points, the Vietoris–Rips filtration; its island diagram is the minimum spanning tree (Kruskal, single linkage, friends-of-friends), and loops need the reduction of the boundary matrix (Section 12.12).
- In three dimensions clusters, tunnels and voids are born and die at maxima, filament saddles, wall saddles and minima; for a Gaussian field their curves peak at $\nu \approx 1.7$, 0 and -1.6, and their alternating sum reproduces the Gaussian χ (Section 12.13). Gravity changes this topology; a monotone relabelling does not.
- The index of a polarization singularity is $\frac{1}{2}\operatorname{sgn}\det J$; a remapping multiplies it by $\operatorname{sgn}\det \mathsf{A}$, so a one-to-one remapping keeps every charge, a fold creates pairs, and noise creates pairs on the pixel scale (Section 12.14).
- Reproduced: the singularity counts of Huterer and Vachaspati (the published 6000 at $\ell_{\max} = 100$ is the no-reionization value), the exponent $\nu \approx \frac{1}{2}$ of Vachaspati and Lue and its meaning (screened pairs, against $\nu \approx 1$ for shuffled charges), and the ordering of the Betti peaks of Pranav and collaborators (Section 12.15).

12.17 Weak-lensing maps to practise on: Gaussian and lognormal convergence

Every topological summary of Sections 12.2 and 12.9 was tried so far on fields we could simulate in a fraction of a second: Gaussian maps of the CMB, shot noise, Gaussian boxes. The field this part is about is the weak-lensing convergence κ of Chapter T1: the projected matter between us and a population of distant galaxies, read off from the tiny stretches of their images. The surveys measure it on thousands of square degrees, and the questions are the ones Section T1.7 ended with. How well do the maps measure the amount of matter, Ω_m, and how clumped it is, σ_8? And can the peaks, islands and lakes of the map shorten the long valley along $S_8 = \sigma_8(\Omega_m/0.3)^{0.5}$ that the power spectrum leaves open?

To answer with Monte Carlo, in the way of Chapters 6 and 8, we need many maps whose statistics we trust, and we need them for several cosmologies. The real tools are ray-traced N-body simulations, each of which costs thousands of processor hours (Section 12.21). On a laptop we need a field that can be made from a power spectrum by fast Fourier transforms, as in Section 7.4, and yet looks like lensing where it matters for topology. The Gaussian field of Section 7.2 is the obvious first try. It fails in a way that a single physical fact makes plain.

The question. What is the smallest value the convergence can take, and what kind of random field respects that floor while keeping the power spectrum the survey measures?

12.17.1 The floor: an empty line of sight

Recall from Section T1.6 how the convergence is built. Cut the line of sight to the galaxies into thick slices; then κ is a weighted sum of the density contrast δ_s averaged through each slice, $\kappa = \sum_s w_s \delta_s \, \Delta\chi$, with weight $w_s = \frac{3}{2}(H_0/c)^2 \Omega_m \, g(\chi_s)\chi_s/a(\chi_s)$. Here χ is comoving distance (the universe is flat, so $f_K = \chi$), a the scale factor, and $g(\chi)$ the lensing efficiency of Section T1.6: the fraction of the source light that lies behind χ, weighted by how strongly matter at χ bends it. Every weight is positive.

A density contrast is $\delta = \rho/\bar{\rho} - 1$, and a density cannot be negative, so $\delta \geq -1$, with equality for a slice that is completely empty. A sum of positive weights times numbers that are each at least -1 is at least minus the sum of the weights. So the lowest convergence any line of sight can have is the one with no matter at all along it, an **empty beam**:

$$\kappa \overset{(1)}{=} \sum_s w_s\, \delta_s\, \Delta\chi \overset{(2)}{\geq} -\sum_s w_s\, \Delta\chi \overset{(3)}{=} -\frac{3}{2}\Big(\frac{H_0}{c}\Big)^2 \Omega_m \int_0^{\chi_{\max}} \mathrm{d}\chi\, \frac{g(\chi)\,\chi}{a(\chi)} \equiv -\lambda. \tag{12.55}$$

(1) The convergence as a sum over independent slices, Section T1.6.
(2) Each $w_s > 0$ and each $\delta_s \geq -1$, so each term is at least $-w_s\Delta\chi$.
(3) Thin slices turn the sum into an integral; we give the result a name, λ, the **shift**.

On the other side there is no ceiling: a line of sight that passes through the core of a cluster has δ of hundreds in a few slices, and κ can reach several tenths. So the distribution of κ is lopsided by construction: a hard wall a little below zero, a long tail above it. A Gaussian has neither.

Example: the size of the floor. `lib_wl.limber_and_shift` evaluates the integral in (12.55) on the same redshift grid as the Limber integral of Section T1.7.3, for the source sample of Section T1.7.2 ($p_z \propto z^2 e^{-(z/0.5)^{1.5}}$, median redshift 0.71) and the Planck cosmology ($\Omega_m = 0.3153$, $\sigma_8 = 0.8111$). It gives $\lambda = 0.0452$. The rms of the convergence in a pixel of 1.5′ is $\sigma_\kappa = 0.0129$ (the square root of the integral of C_ℓ over the modes the pixels carry, (7.13)). The wall therefore sits $\lambda/\sigma_\kappa = 0.0452/0.0129 \approx 3.5$ standard deviations below the mean. A Gaussian puts a fraction 2.3×10^{-4} of its pixels beyond 3.5σ on each side; it would give about 15 pixels of a 256×256 map a convergence that no arrangement of matter can produce, and, worse, would make the high tail exactly as thin as the low one.

Example: how the floor moves with the cosmology. The shift does not depend on σ_8 at all: (12.55) contains no clumping, only the geometry and the mean density. It does depend on Ω_m. Changing Ω_m by ±5% at fixed h gives $\lambda = 0.0465$ and 0.0437, a logarithmic slope

$$\frac{\mathrm{d}\ln\lambda}{\mathrm{d}\ln\Omega_m} \approx \frac{\ln 0.0465 - \ln 0.0437}{\ln 1.05 - \ln 0.95} = 0.62.$$

The explicit factor Ω_m alone would give 1. The rest comes from the distances: more matter decelerates the expansion more, so the same source redshift lies at a smaller comoving distance, and the integral of $g\chi/a$ over a shorter path is smaller. This small number will matter in Section 12.20: it is a property of the map that responds to Ω_m but not to σ_8, which the power spectrum, responding to both in nearly the same way, cannot provide.

12.17.2 The shifted lognormal field

What is the simplest random field with a floor at $-\lambda$, a mean of zero, and a prescribed power spectrum? State the inputs first, and separate physics from convenience.

- *Physical inputs.* The convergence never goes below $-\lambda$ (12.55); its mean over the sky is zero (lensing conserves the mean, since the matter fluctuations average to zero); its power spectrum is the Limber $C_\ell^{\kappa\kappa}$ of Section T1.6.
- *Mathematical choice.* Build κ from a Gaussian field G by the simplest increasing transformation with a floor: an exponential. This is a choice, not a law. Its only justification is that it is the simplest such map and that it describes simulated matter fields reasonably well [55].

The **shifted lognormal** convergence is then

$$\kappa(\boldsymbol{\theta}) = \lambda\Big[e^{\,G(\boldsymbol{\theta})-s^2/2} - 1\Big], \qquad G \text{ Gaussian, mean } 0, \text{ variance } s^2, \text{ correlation function } \xi_G(\vartheta). \tag{12.56}$$

The exponential is positive, so $\kappa > -\lambda$ automatically. Two questions remain, and each is answered by one property of the Gaussian. Why the $-s^2/2$ in the exponent? And which ξ_G gives the convergence the correlation function we want? (Xavier and collaborators write the same field as $X = e^Z - \lambda$ with a mean μ for Z; their $\alpha = \langle X\rangle + \lambda$ is our λ when $\langle X\rangle = 0$ [55, eqs. 1, 4, 6].)

The one fact we use. For a Gaussian variable Y with mean m and variance v,

$$\langle e^Y\rangle = e^{m+v/2}, \tag{12.57}$$

the moment generating function of the Gaussian at $t = 1$ (complete the square in the exponent of the density, as in Section 7.2). A sum of jointly Gaussian variables is again Gaussian, so the fact applies to any sum $G_1 + G_2$ as well.

The mean. Take $Y = G - s^2/2$, with mean $-s^2/2$ and variance s^2:

$$\langle\kappa\rangle \overset{(1)}{=} \lambda\Big[\langle e^{G-s^2/2}\rangle - 1\Big] \overset{(2)}{=} \lambda\Big[e^{-s^2/2+s^2/2} - 1\Big] = 0. \tag{12.58}$$

(1) The average of a sum is the sum of the averages; λ is a constant.
(2) Equation (12.57) with $m = -s^2/2$ and $v = s^2$.

The $-s^2/2$ is exactly what cancels the upward pull of the exponential: e^G alone has mean $e^{s^2/2} > 1$, because the exponential lifts the high values of G more than it lowers the low ones.

The correlation function. Take two directions a separation ϑ apart, write G_1, G_2 for the Gaussian field there and κ_1, κ_2 for the convergence. Their sum $G_1 + G_2$ is Gaussian with mean zero and variance $s^2 + s^2 + 2\xi_G(\vartheta)$. Then

$$\begin{aligned}
\xi_\kappa(\vartheta) \equiv \langle\kappa_1\kappa_2\rangle &\overset{(1)}{=} \lambda^2\Big\langle\big(e^{G_1-s^2/2} - 1\big)\big(e^{G_2-s^2/2} - 1\big)\Big\rangle \\
&\overset{(2)}{=} \lambda^2\Big[e^{-s^2}\big\langle e^{G_1+G_2}\big\rangle - 1 - 1 + 1\Big] \\
&\overset{(3)}{=} \lambda^2\Big[e^{-s^2}\,e^{s^2+\xi_G(\vartheta)} - 1\Big] = \lambda^2\Big[e^{\xi_G(\vartheta)} - 1\Big],
\end{aligned} \tag{12.59}$$

$$\xi_G(\vartheta) \overset{(4)}{=} \ln\Big[1 + \frac{\xi_\kappa(\vartheta)}{\lambda^2}\Big]. \tag{12.60}$$

(1) The definition (12.56) at the two points; the mean of κ is zero (12.58), so $\langle\kappa_1\kappa_2\rangle$ is the covariance.
(2) Multiply out. The two cross terms are each $\langle e^{G-s^2/2}\rangle = 1$ by (12.58).
(3) Equation (12.57) for $Y = G_1 + G_2$, with $m = 0$ and $v = 2s^2 + 2\xi_G$, so $\langle e^Y\rangle = e^{s^2+\xi_G}$.
(4) Solve for ξ_G.

Equations (12.59) and (12.60) are eqs. (6) and (7) of [55], in our notation. They are the recipe: from the wanted ξ_κ compute ξ_G point by point, turn it into a power spectrum P_G, and make G by the method of Section 7.4.2. The transformation acts point by point in real space, so the correlation function, not the power spectrum, is the natural place to apply it. At zero separation it fixes the variance of G: $s^2 = \xi_G(0) = \ln(1 + \sigma_\kappa^2/\lambda^2)$.

The skewness. How lopsided is this field? Write $Y = e^{G-s^2/2}$, so $\kappa = \lambda(Y - 1)$ and the skewness of κ is that of Y. By (12.57) with $m = -ns^2/2$ and $v = n^2s^2$ for the variable $nG - ns^2/2$,

$\langle Y^n\rangle = e^{n(n-1)s^2/2}$, that is $1, e^{s^2}, e^{3s^2}$ for $n = 1, 2, 3$. Call $q \equiv e^{s^2} - 1 = \sigma_\kappa^2/\lambda^2$ (the zero-separation case of (12.59)). Then

$$\begin{aligned}
\big\langle (Y-1)^2 \big\rangle &\overset{(1)}{=} e^{s^2} - 2 + 1 = q, \\
\big\langle (Y-1)^3 \big\rangle &\overset{(2)}{=} e^{3s^2} - 3e^{s^2} + 3 - 1 \overset{(3)}{=} (1+q)^3 - 3(1+q) + 2 \overset{(4)}{=} q^2(3+q), \\
\gamma_\kappa \equiv \frac{\langle (Y-1)^3\rangle}{\langle (Y-1)^2\rangle^{3/2}} &\overset{(5)}{=} \sqrt{q}\,(3+q) = r\,(3+r^2), \qquad r \equiv \frac{\sigma_\kappa}{\lambda}.
\end{aligned} \tag{12.61}$$

(1) Expand the square and use $\langle Y\rangle = 1$, $\langle Y^2\rangle = e^{s^2}$.
(2) Expand the cube, $(Y-1)^3 = Y^3 - 3Y^2 + 3Y - 1$, and use the three moments.
(3) $e^{s^2} = 1 + q$ and $e^{3s^2} = (1+q)^3$.
(4) Multiply out: $1 + 3q + 3q^2 + q^3 - 3 - 3q + 2 = 3q^2 + q^3$.
(5) Divide by $q^{3/2}$; $\sqrt{q} = \sigma_\kappa/\lambda$.

This is eq. (10) of [55] with their $\alpha = \lambda$ and $v = \sigma_\kappa^2$. The whole one-point shape is controlled by one number, the ratio r of the rms to the depth of the floor. A shallow floor (small λ) or a large rms makes a strongly skewed field; when $r \to 0$ the field becomes Gaussian.

Example: the skewness by hand. With $\sigma_\kappa = 0.0129$ and $\lambda = 0.0452$, $r = 0.29$ and $\gamma_\kappa = 0.29 \times (3 + 0.084) \approx 0.89$; the unrounded numbers give 0.88. The skewness of a Gaussian is zero. A map smoothed with a Gaussian of 3′, as the peak and topology analyses will use, has a smaller rms, 0.0064, so $r \approx 0.0064/0.0452 = 0.14$ and (12.61) suggests $\gamma \approx 0.14 \times 3.02 \approx 0.43$. This last number is only a guide: smoothing averages neighbouring pixels, and an average of lognormal variables is not exactly lognormal.

Example: why the skewness carries what the spectrum does not. The literature often quotes the **reduced skewness**, the third moment divided by the square of the second, $S_3 \equiv \langle\kappa^3\rangle/\langle\kappa^2\rangle^2 = \gamma_\kappa/\sigma_\kappa$. For a weakly skewed lognormal field (12.61) gives $S_3 \approx 3r/\sigma_\kappa = 3/\lambda$. The clumping amplitude has dropped out: σ_8 raises σ_κ but not λ. What remains is $1/\lambda \propto \Omega_m^{-0.62}$. Perturbation theory says the same about the real field: the reduced skewness of the lensing signal is independent of the normalisation of the power spectrum and scales roughly as Ω_m^{-1} [BartelmannSchneider2001 p. 146]. A low-Ω_m, high-σ_8 universe and a high-Ω_m, low-σ_8 one with the same S_8 have nearly the same C_ℓ (Section T1.7.4), but the first has fewer lenses, each more clumped, and so a more lopsided map. The one-point shape is a lever on the valley of Section T1.7.

12.17.3 Making the maps

The patch is 256×256 pixels of 1.5′, a square of 6.4° on a side (24.4 such patches make 1000 deg^2), periodic as every FFT simulation is. `w01_lognormal.py` computes the Limber spectrum with our own integral over CAMB's halofit matter spectrum (the hand-made Limber code of Section T1.7.3, now inside `lib_wl`) and λ for the fiducial cosmology and for Ω_m and σ_8 moved by ±5%, and caches all five for the later scripts. The lognormal recipe is four lines:

Listing 12.11: From the target spectrum to a shifted lognormal map (`code/ch12/lib_wl.py`)

```
127 def lognormal_spectrum(P, lam):
128     """Spectrum P_G of the Gaussian field G such that kappa = lam (e^{G - s^2/2} - 1) has
        spectrum P.
129
130     xi_kappa(x) = ifft2(P)/A_pix on the grid;  xi_G = ln(1 + xi_kappa / lam^2);  P_G =
        A_pix fft2(xi_G).
```

```
    Negative P_G (a few modes at the highest l) are set to zero.
    """
    xi = np.fft.ifft2(P).real / A_PIX
    xig = np.log1p(xi / lam**2)
    PG = np.fft.fft2(xig).real * A_PIX
    PG[0, 0] = 0.0
    return np.clip(PG, 0.0, None), xig[0, 0]


def lognormal_map(PG, lam, white):
    """Shifted lognormal convergence with zero mean, built from the Gaussian spectrum PG."""
    G = gaussian_map(PG, white)
    s2 = G.var()
    k = lam * (np.exp(G - s2 / 2) - 1.0)
    return k - k.mean()
```

Line 133 turns the target spectrum on the FFT grid into the correlation function at every pixel separation (the inverse transform of the Wiener–Khinchin relation, Section 7.3.2, with the grid normalisation of (7.19)). Line 134 is (12.60), point by point. Line 135 transforms ξ_G back into the spectrum P_G of the Gaussian field. A spectrum must be non-negative, and nothing guarantees that the transform of $\ln(1 + \xi_\kappa/\lambda^2)$ is; this is the limitation Xavier and collaborators analyse, which bites hardest when several correlated lognormal fields must be made at once [55, §2.2]. For our single map the negative modes are a fraction 0.00% of the grid, and line 137 sets them to zero. Lines 142–145 draw G from white noise, exponentiate as in (12.56), and remove the tiny mean that one finite map has. A Gaussian map with the same spectrum is made from the same white noise (`gaussian_map`), so the two maps of Figure 12.40 have the same phases and differ only through the transformation.

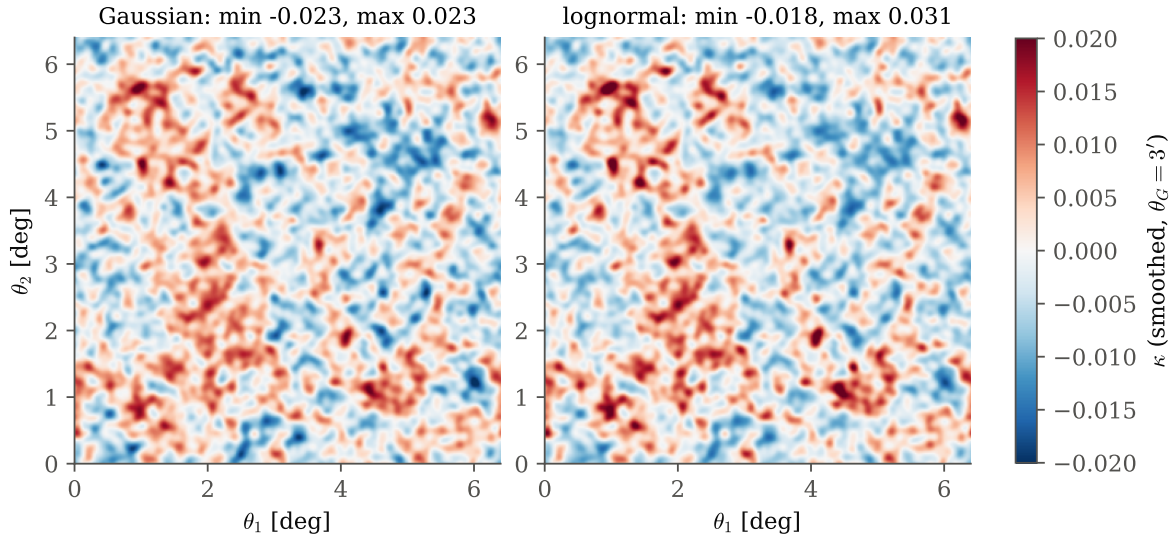


Figure 12.40: A Gaussian convergence map (left) and a shifted lognormal one (right), built from the same white noise and with the same power spectrum, both smoothed with a 3′ Gaussian. The large-scale pattern is identical. The lognormal map has compact, intense red spots, where many lenses line up, and shallow, wide blue regions, because it cannot fall far below zero: compare the minimum and maximum in the titles.

What the maps show. Over 100 maps, the measured skewness of the lognormal pixels is 0.88, against 0.88 from (12.61); the Gaussian maps give 0.001. After 3′ smoothing the lognormal skewness is 0.44 (the guide above said about 0.43) and the Gaussian 0.001. The lowest lognormal

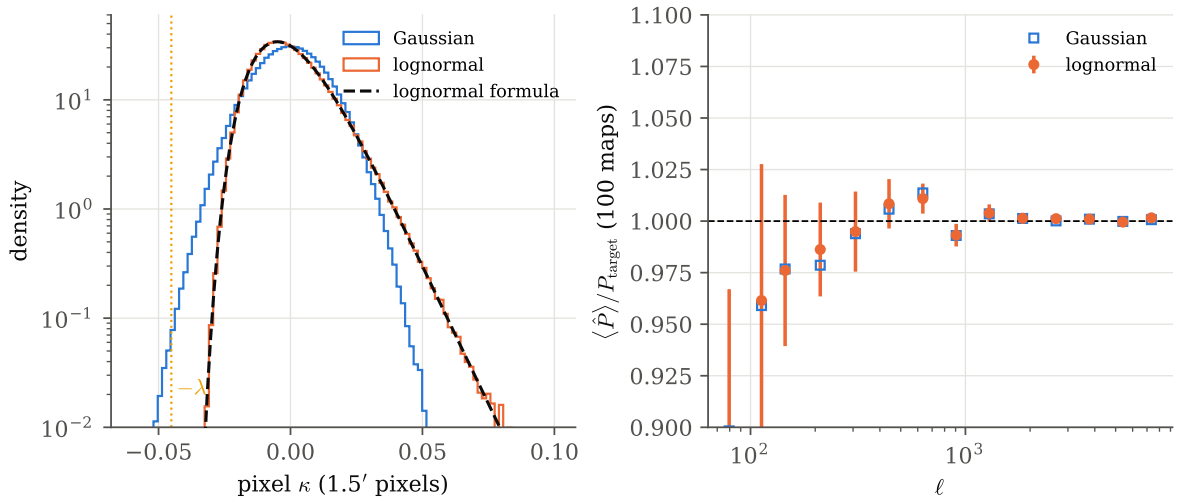


Figure 12.41: Left: the distribution of pixel values in 20 Gaussian and 20 lognormal maps (1.5′ pixels, logarithmic vertical axis), with the lognormal density obtained from (12.56) by a change of variables (dashed). The Gaussian is symmetric; the lognormal stops at the floor $-\lambda$ (dotted line) and has a long tail to high κ. Right: the band powers of 100 maps of each kind divided by the target spectrum. Both reproduce it; the scatter at low ℓ, the same for both since the maps share their white noise, is the sampling variance of the few modes there.

pixel in 20 maps is −0.0356, inside the wall at −0.0452. The power spectrum survives the transformation: for $\ell > 300$ the mean band powers of the lognormal maps differ from the target by at most 1.1%; including the lowest bands the largest deviation is 10.9%, the same as the Gaussian maps' 10.2%, which is the sampling scatter of a handful of modes and not an error of the method (Figure 12.41, right). So the two families of maps share their two-point statistics and differ in everything else. That is exactly the situation in which a summary beyond C_ℓ can win.

What the lognormal model gets wrong. It is a model of the one-point shape with one free number, and the number we chose, the empty-beam shift, is not the best one. Xavier and collaborators made convergence maps the honest way, as sums of correlated lognormal density slices along the line of sight, and asked which shift a lognormal fitted to them needs. For sources at $z = 2.1$ the empty-beam value is 0.2115, while matching the first three moments gives 0.1288 [55, Table 2]. The empty beam overestimates the shift by about 60%, so by $S_3 \approx 3/\lambda$ it underestimates the skewness by a similar factor: our maps are less non-Gaussian than the real sky, and any gain we find from non-Gaussian summaries is, if anything, too small. The reason is physical: a real line of sight is never empty, and the deepest voids have $\delta \approx -0.8$, not -1. A sum of lognormal slices is also not itself lognormal [55, §3.3], and a lognormal field has no haloes, no filaments and no lensing of lensing. For each of these the research analyses use ray-traced N-body maps [51]; we return to the cost in Section 12.21.

12.18 From galaxy shapes to a mass map

A weak-lensing survey measures galaxy shapes. In the weak limit, the measured ellipticity $\varepsilon \approx \varepsilon^{(0)} + \gamma$ combines the intrinsic ellipticity $\varepsilon^{(0)}$ with the shear γ of Section T1.6. The shear is a spin-2 quantity, represented by a stick with no arrowhead. Peaks, islands and lakes are properties of a map of κ, so before any of the summaries of Sections 12.2 and 12.9 can be computed, the convergence must be reconstructed from the sticks. Three things stand between the sticks and the map. The reconstruction itself is non-local: the shear at one point is

caused by matter everywhere. The shapes carry noise some thirty times larger than the signal (Section T1.8). And part of every survey is missing, behind bright stars and beyond its edges.

The question. Given the shear on a patch of sky, which convergence produced it? And how faithful is the answer when each galaxy's shape is mostly its own, and when some of the patch is blank?

12.18.1 The inversion, mode by mode

Recall the two second-derivative relations of Section T1.6, in the flat angles θ_1, θ_2 of Figure T1.7: the convergence is half the Laplacian of the lensing potential, $\kappa = \frac{1}{2}(\psi_{,11} + \psi_{,22})$, and the shear is $\gamma \equiv \gamma_1 + i\gamma_2$ with $\gamma_1 = \frac{1}{2}(\psi_{,11} - \psi_{,22})$ and $\gamma_2 = \psi_{,12}$, a comma meaning a derivative. Both are linear in one unknown function, so we should eliminate it. Derivatives become multiplications in Fourier space. Write each field on the patch as a sum of plane waves, $f(\boldsymbol{\theta}) = \int \frac{\mathrm{d}^2\ell}{(2\pi)^2} \hat{f}(\boldsymbol{\ell}) e^{i\boldsymbol{\ell}\cdot\boldsymbol{\theta}}$, with $\boldsymbol{\ell} = (\ell_1, \ell_2) = \ell(\cos\varphi_\ell, \sin\varphi_\ell)$, so that $\partial/\partial\theta_a$ acting on a wave multiplies it by $i\ell_a$. Then, for every $\boldsymbol{\ell} \neq \mathbf{0}$,

$$\hat{\kappa} \overset{(1)}{=} \tfrac{1}{2}\big[(i\ell_1)^2 + (i\ell_2)^2\big]\hat{\psi} = -\tfrac{1}{2}\ell^2\hat{\psi},$$

$$\hat{\gamma} \equiv \hat{\gamma}_1 + i\hat{\gamma}_2 \overset{(2)}{=} \tfrac{1}{2}\big[(i\ell_1)^2 - (i\ell_2)^2\big]\hat{\psi} + i\,(i\ell_1)(i\ell_2)\hat{\psi} = -\tfrac{1}{2}\big(\ell_1^2 - \ell_2^2 + 2i\ell_1\ell_2\big)\hat{\psi}$$

$$\overset{(3)}{=} -\tfrac{1}{2}(\ell_1 + i\ell_2)^2\hat{\psi} \overset{(4)}{=} -\tfrac{1}{2}\ell^2 e^{2i\varphi_\ell}\hat{\psi} \overset{(5)}{=} e^{2i\varphi_\ell}\,\hat{\kappa}, \tag{12.62}$$

$$\hat{\kappa} \overset{(6)}{=} e^{-2i\varphi_\ell}\,\hat{\gamma}. \tag{12.63}$$

(1) The Laplacian of a plane wave: each derivative brings down $i\ell_a$, and $i^2 = -1$.
(2) The two shear components with the same rule, combined into one complex number.
(3) $(\ell_1 + i\ell_2)^2 = \ell_1^2 - \ell_2^2 + 2i\ell_1\ell_2$.
(4) $\ell_1 + i\ell_2 = \ell e^{i\varphi_\ell}$, the polar form of the wavevector.
(5) The first line: $-\frac{1}{2}\ell^2\hat{\psi} = \hat{\kappa}$.
(6) Multiply both sides by $e^{-2i\varphi_\ell}$, which has modulus one.

Equation (12.63) is the **Kaiser–Squires inversion** [46]. In the notation of [BartelmannSchneider2001 p. 88, eqs. 5.2–5.3] it reads $\hat{\kappa} = \pi^{-1}\hat{\mathcal{D}}^*\hat{\gamma}$ with $\hat{\mathcal{D}} = \pi e^{2i\varphi_\ell}$, the row of the table in Section T1.6. In words: a convergence wave of any orientation produces a shear wave of the same wavelength whose sticks are turned by twice the angle of the wavevector, and undoing the turn returns the convergence. Nothing is lost in either direction, because a phase has modulus one.

Two remarks complete the inversion. First, the chain needs $\boldsymbol{\ell} \neq \mathbf{0}$: a uniform convergence κ_0 comes from $\psi = \frac{1}{2}\kappa_0\theta^2$, whose $\psi_{,11} = \psi_{,22}$ and $\psi_{,12} = 0$, so it makes no shear at all. The map is therefore known only up to an added constant, the **mass-sheet degeneracy** [BartelmannSchneider2001 p. 88]. On a periodic simulated patch we fix the constant by requiring zero mean; a peak height or a threshold measured from the mean is unaffected by it. Second, in real space the multiplication by $e^{-2i\varphi_\ell}$ is a convolution with a kernel that falls only as $1/\theta^2$ [BartelmannSchneider2001 eq. 5.4]: the convergence at a point is a weighted sum of the shear over the whole patch. This is why a hole in the data will spoil the map well beyond its edge.

***E* and *B* from any stick field.** Apply (12.63) to a stick field that did *not* come from a convergence, such as one with noise or an error in it. The result is a complex map, and we name its two parts:

$$\kappa_E(\boldsymbol{\theta}) + i\kappa_B(\boldsymbol{\theta}) \equiv \int \frac{\mathrm{d}^2\ell}{(2\pi)^2}\, e^{-2i\varphi_\ell}\,\hat{\gamma}(\boldsymbol{\ell})\, e^{i\boldsymbol{\ell}\cdot\boldsymbol{\theta}}. \tag{12.64}$$

If the sticks came from a real convergence, the integrand is $\hat{\kappa}(\boldsymbol{\ell})$, whose inverse transform is the real map κ, so $\kappa_E = \kappa$ and $\kappa_B = 0$: lensing makes no B, the statement proved in real space in (T1.7). Now turn every stick by 45°. A stick at angle ϕ is $|\gamma|\, e^{2i\phi}$, so the turn multiplies γ by $e^{2i\pi/4} = i$, and by linearity the output becomes $i\kappa$: the same map, moved entirely into κ_B. This is the swirl pattern of Figure T1.6(b). Pure E and pure B are the two ways a stick field can be arranged around each point, and (12.64) sorts any field into them.

Listing 12.12: Shear from convergence and back (`code/ch12/lib_wl.py`)

```
148 def shear_from_kappa(kappa):
149     """gamma_1 + i gamma_2 from kappa: multiply each Fourier mode by e^{2 i phi_l}."""
150     g = np.fft.ifft2(E2IPHI * np.fft.fft2(kappa))
151     return g.real, g.imag
152
153
154 def kaiser_squires(g1, g2):
155     """kappa_E + i kappa_B: multiply each mode of gamma_1 + i gamma_2 by e^{-2 i phi_l}."""
156     k = np.fft.ifft2(np.conj(E2IPHI) * np.fft.fft2(g1 + 1j * g2))
157     return k.real, k.imag
```

The two functions are (12.62) and (12.64) on the FFT grid: `E2IPHI` is the array $e^{2i\varphi_\ell} = (\ell_1 + i\ell_2)^2/\ell^2$ (step 3 of the chain, which avoids computing any angle), set to zero at $\boldsymbol{\ell} = \mathbf{0}$, where the mass sheet lives. Lines 150 and 156 multiply every mode by the phase or its conjugate.

Example: the round trip. Take one lognormal map of Section 12.17.2, make its shear, and invert. `w02_ks.py` finds that the recovered κ_E differs from the input by at most 2.7×10^{-15} of the map's rms, and κ_B is at most 1.8×10^{-15} of it: zero to the precision of double arithmetic. The input map has rms 0.0126 and each shear component 0.0089. Why smaller by $\sqrt{2}$? Because a phase of modulus one keeps the power of every mode, the complex shear has the variance of κ, $\langle|\gamma|^2\rangle = \langle\gamma_1^2\rangle + \langle\gamma_2^2\rangle = \sigma_\kappa^2$, and on an isotropic sky the two components share it equally: $\sigma_{\gamma_1} = \sigma_\kappa/\sqrt{2} = 0.0126/1.414 = 0.0091$.

12.18.2 Shape noise in the mass map

Recall the noise of (T1.11). With $\sigma_e = 0.27$ the rms of one ellipticity component and $\bar{n}$ galaxies per unit solid angle, a pixel of solid angle A holds about $\bar{n}A$ galaxies and its mean ellipticity scatters by $\sigma_e/\sqrt{\bar{n}A}$ in each component, independently from pixel to pixel. For the KiDS-like density $\bar{n} = 6.2\,\mathrm{arcmin}^{-2}$ and our 1.5′ pixels this is $0.27/\sqrt{6.2 \times 2.25} = 0.0723$ per component, against a signal of 0.0089: the per-pixel signal-to-noise ratio is about 1/8. The inversion keeps the power of every mode, so the noise power $\sigma_e^2/\bar{n}$ of each component goes, unchanged in total, into κ_E and κ_B, each receiving $\sigma_e^2/\bar{n}$; fifty pure-noise maps give 1.002 and 0.990 times $\sigma_e^2/\bar{n} = 9.95 \times 10^{-10}$ in E and B. The noise is white, the same at every ℓ, while the signal falls with ℓ; at the pixel scale the map is all noise. Shrinking the pixels makes it worse without limit, the "infinite noise" of a direct inversion [BartelmannSchneider2001 p. 88]. The cure is to smooth.

How much noise survives smoothing? Smooth with a normalised Gaussian of standard deviation θ_G, $W(\boldsymbol{\theta}) = e^{-\theta^2/2\theta_G^2}/(2\pi\theta_G^2)$. White noise $n(\boldsymbol{\theta})$ with power N has the correlation $\langle n(\boldsymbol{\theta})n(\boldsymbol{\theta}')\rangle = N\,\delta^{(2)}(\boldsymbol{\theta} - \boldsymbol{\theta}')$ (a field of independent pixels, Section T1.6). The smoothed noise is $n_s(\boldsymbol{\theta}) = \int \mathrm{d}^2\theta'\, W(\boldsymbol{\theta} - \boldsymbol{\theta}')n(\boldsymbol{\theta}')$, so

$$\sigma^2_{\mathrm{noise}} \equiv \langle n_s^2\rangle \overset{(1)}{=} \int \mathrm{d}^2\theta_1 \int \mathrm{d}^2\theta_2\, W(\boldsymbol{\theta} - \boldsymbol{\theta}_1)W(\boldsymbol{\theta} - \boldsymbol{\theta}_2)\, N\delta^{(2)}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) \overset{(2)}{=} N \int \mathrm{d}^2\theta'\, W^2(\boldsymbol{\theta}')$$
$$\overset{(3)}{=} \frac{N}{(2\pi\theta_G^2)^2} \int_0^\infty 2\pi\vartheta\, e^{-\vartheta^2/\theta_G^2}\, \mathrm{d}\vartheta \overset{(4)}{=} \frac{N}{4\pi^2\theta_G^4}\,\pi\theta_G^2 \overset{(5)}{=} \frac{\sigma_e^2}{4\pi\theta_G^2\,\bar{n}}. \tag{12.65}$$

(1) Square the smoothed noise and average; only the noise is random.
(2) The delta function removes one integral; shift the variable.
(3) Insert W^2 and use polar coordinates.
(4) $\int_0^\infty 2\pi\vartheta e^{-\vartheta^2/\theta_G^2}\mathrm{d}\vartheta = \pi\theta_G^2$ (substitute $u = \vartheta^2$).
(5) $N = \sigma_e^2/\bar{n}$ for κ_E.

The smoothing area $4\pi\theta_G^2$ counts the galaxies that the filter effectively averages. For $\theta_G = 3'$ and $\bar{n} = 6.2\,\mathrm{arcmin}^{-2}$, $\sigma^2_{\mathrm{noise}} = 0.0729/(4\pi \times 9 \times 6.2) = 1.04 \times 10^{-4}$, so $\sigma_{\mathrm{noise}} = 0.0102$. On the pixel grid the same quantity is a lattice sum over the modes, and it gives 0.0102; fifty noise maps give 0.0102. The smoothed signal has rms 0.0064 (Section 12.17.3), so even after smoothing each pixel is dominated by noise. Every peak, island and lake of the mass map will be measured in units of this σ_{noise}, the signal-to-noise map $\nu(\boldsymbol{\theta}) \equiv \kappa_{E,s}(\boldsymbol{\theta})/\sigma_{\mathrm{noise}}$.

Example: a published noise level, recovered. Kratochvil and collaborators quote the smoothed noise as $\sigma^2_{\mathrm{noise}} = \langle\sigma_\gamma^2\rangle/(2\pi\theta_G^2 n_{\mathrm{gal}})$ [52, eq. 9], with σ_γ the rms of one shear component. Compared with (12.65) the 4π has become 2π. The two agree if their θ_G is the width in $e^{-\theta^2/\theta_G^2}$, a standard deviation of $\theta_G/\sqrt{2}$: then $4\pi(\theta_G/\sqrt{2})^2 = 2\pi\theta_G^2$. Their own numbers confirm this reading. With $\sigma_\gamma = 0.15 + 0.035z = 0.22$ for sources at $z = 2$, $n_{\mathrm{gal}} = 15\,\mathrm{arcmin}^{-2}$ and $\theta_G = 1'$, the formula gives $\sigma^2_{\mathrm{noise}} = 0.0484/(2\pi \times 15) = 5.1 \times 10^{-4}$, $\sigma_{\mathrm{noise}} = 0.0227$, and their detection threshold of $4.5\sigma_{\mathrm{noise}}$ is 0.102, the value they quote. A smoothing scale is a number with a convention attached, like the σ_e of Section T1.6.2.

12.18.3 Aperture mass: a filtered map without the inversion

Kaiser–Squires builds the convergence at a point from the shear of the whole patch, and then we smooth it. The peak analyses of Dietrich and Hartlap and of Heydenreich and collaborators do not do it in two steps. They compute the **aperture mass**, a filtered convergence obtained straight from the sticks [53, 56]. *Can the filter be put inside the inversion?* Choose a round filter $U(|\boldsymbol{\theta}|)$ and ask for

$$M(\boldsymbol{\theta}_0) \equiv \int \mathrm{d}^2\theta\, U(|\boldsymbol{\theta} - \boldsymbol{\theta}_0|)\,\kappa(\boldsymbol{\theta}), \qquad \int \mathrm{d}^2\theta\, U = 0, \tag{12.66}$$

the convergence averaged around $\boldsymbol{\theta}_0$ with weight U. The second condition makes the filter **compensated**: positive in the middle and negative in a ring, so that its total weight is zero. We express M through the shear. The convolution theorem turns (12.66) into a sum over modes, and (12.63) replaces $\hat{\kappa}$:

$$\begin{aligned} M(\boldsymbol{\theta}_0) &\overset{(1)}{=} \int \frac{\mathrm{d}^2\ell}{(2\pi)^2}\,\hat{U}(\ell)\,\hat{\kappa}(\boldsymbol{\ell})\,e^{i\boldsymbol{\ell}\cdot\boldsymbol{\theta}_0} \overset{(2)}{=} \int \mathrm{d}^2\theta\,\gamma(\boldsymbol{\theta})\,K(\boldsymbol{\theta}_0 - \boldsymbol{\theta}), \\ K(\boldsymbol{x}) &\equiv \int \frac{\mathrm{d}^2\ell}{(2\pi)^2}\,\hat{U}(\ell)\,e^{-2i\varphi_\ell}\,e^{i\boldsymbol{\ell}\cdot\boldsymbol{x}} \\ &\overset{(3)}{=} \int_0^\infty \frac{\ell\,\mathrm{d}\ell}{(2\pi)^2}\,\hat{U}(\ell) \int_0^{2\pi} \mathrm{d}\varphi_\ell\, e^{-2i\varphi_\ell}\,e^{i\ell|\boldsymbol{x}|\cos(\varphi_\ell - \varphi_x)} \\ &\overset{(4)}{=} -e^{-2i\varphi_x}\,Q(|\boldsymbol{x}|), \qquad Q(\vartheta) \equiv \int_0^\infty \frac{\ell\,\mathrm{d}\ell}{2\pi}\,\hat{U}(\ell)\,J_2(\ell\vartheta). \end{aligned} \tag{12.67}$$

(1) The transform of a filtered map is the product of the transforms (U is even, so $\hat{U}$ is real and depends on ℓ only).

(2) Insert $\hat{\kappa} = e^{-2i\varphi_\ell}\hat{\gamma}$ and write $\hat{\gamma}(\boldsymbol{\ell}) = \int \mathrm{d}^2\theta\, \gamma(\boldsymbol{\theta}) e^{-i\boldsymbol{\ell}\cdot\boldsymbol{\theta}}$; the $\boldsymbol{\ell}$ integral moves inside and collects into the kernel K.
(3) Polar coordinates in $\boldsymbol{\ell}$, with φ_x the polar angle of $\boldsymbol{x}$, so that $\boldsymbol{\ell}\cdot\boldsymbol{x} = \ell\,|\boldsymbol{x}|\cos(\varphi_\ell - \varphi_x)$.
(4) Shift the angle, $\varphi_\ell = \varphi_x + \alpha$, which pulls out $e^{-2i\varphi_x}$, and use the Bessel integral $\int_0^{2\pi} e^{-2i\alpha} e^{iz\cos\alpha}\mathrm{d}\alpha = 2\pi\, i^{-2} J_{-2}(z) = -2\pi J_2(z)$.

For a map that is pure E the result is real, so we may keep the real part. With $\boldsymbol{x} = \boldsymbol{\theta}_0 - \boldsymbol{\theta}$ the angle φ_x differs from the polar angle φ of $\boldsymbol{\theta} - \boldsymbol{\theta}_0$ by π, which leaves $e^{-2i\varphi_x} = e^{-2i\varphi}$, and

$$M(\boldsymbol{\theta}_0) = \int \mathrm{d}^2\theta\, Q(|\boldsymbol{\theta} - \boldsymbol{\theta}_0|)\, \gamma_t(\boldsymbol{\theta}; \boldsymbol{\theta}_0), \qquad \gamma_t(\boldsymbol{\theta}; \boldsymbol{\theta}_0) \equiv -\mathrm{Re}\big[\gamma(\boldsymbol{\theta})\, e^{-2i\varphi}\big]. \tag{12.68}$$

The number γ_t is the **tangential shear**: the component of the stick at $\boldsymbol{\theta}$ along the direction that circles $\boldsymbol{\theta}_0$. For a stick tangential to the circle, $\gamma = -|\gamma|\, e^{2i\varphi}$ (the pattern around a lump found in Section T1.6), so $\gamma_t = +|\gamma|$; a radial stick gives $-|\gamma|$. The real-space form of the weight, $Q(\vartheta) = (2/\vartheta^2)\int_0^\vartheta \vartheta' U(\vartheta')\mathrm{d}\vartheta' - U(\vartheta)$, is the one quoted by [56, eq. 2]; the Bessel form above is what falls out of our inversion. Three things follow at once. *No mass sheet.* The compensation condition says $\hat{U}(0) = \int U\, \mathrm{d}^2\theta = 0$. A constant κ_0 lives only in the mode $\boldsymbol{\ell} = \boldsymbol{0}$, which $\hat{U}$ removes, so M is blind to it: the degeneracy of Section 12.18.1 is built out of the observable. *Holes cost only what they hold.* Equation (12.68) is an integral over the sticks that exist. In the estimator below a masked galaxy is a term missing from a sum, and the aperture around a hole is merely less well measured, whereas a Kaiser–Squires map is spoiled over the whole range of its $1/\theta^2$ kernel (Section 12.18.4). Heydenreich and collaborators discard the pixels for which at least half the aperture is masked [56, §2.2]. *A noise level.* Replace the integral by a sum over the $\bar{n}$ galaxies per unit solid angle, each carrying its measured ellipticity $\varepsilon_{t,i} = \varepsilon^{(0)}_{t,i} + \gamma_t$: $\hat{M} = \bar{n}^{-1}\sum_i Q_i\, \varepsilon_{t,i}$. The intrinsic tangential components are independent with variance σ_e^2, so

$$\sigma_M^2 \overset{(1)}{=} \frac{1}{\bar{n}^2}\,\bar{n}\int \mathrm{d}^2\theta\, Q^2\, \sigma_e^2 \overset{(2)}{=} \frac{\sigma_e^2}{\bar{n}}\int \mathrm{d}^2\theta\, Q^2(|\boldsymbol{\theta}|). \tag{12.69}$$

(1) The variance of a sum of independent terms is the sum of the variances, and the number of galaxies in $\mathrm{d}^2\theta$ is $\bar{n}\,\mathrm{d}^2\theta$.
(2) Collect the constants.

This is the shape of (12.65), in which the Gaussian W played the part of Q: smoothed noise is always $\sigma_e^2/\bar{n}$ times the integral of the square of the weight. The aperture-mass **signal-to-noise map** is $\nu = M/\sigma_M$, and it is the map of the three published analyses; Heydenreich and collaborators choose Q as a smooth function shaped like the lensing signal of a typical dark-matter halo, so that the peaks of ν are tuned to haloes [56, §2.2.1]. We use the Gaussian-smoothed Kaiser–Squires map, which is easier to simulate on a periodic grid and has the same noise algebra. The two maps are two different linear filters of the same shear, so the qualitative behaviour of their peaks and islands, which is what the comparisons of Section 12.20.4 rest on, is the same, and the numbers are not.

12.18.4 What a mask does to the map

Real data have holes. Our mask removes 40 discs of radius 3′ to 9′ around imagined bright stars and a strip of 0.4° along one edge, keeping a fraction $f_{\mathrm{obs}} = 0.91$ of the patch. Where there is no galaxy, the measured shear is zero. What does (12.64) do with a shear map that is suddenly zero inside a hole?

A masked shear map is not the shear of any convergence. Inside a hole the true κ is not zero, yet the sticks around it are gone, and the inversion, which builds κ at a point from the sticks

everywhere with a $1/\theta^2$ kernel, attributes the missing pattern partly to a different convergence and partly to B. In Fourier language the mask multiplies the shear by a window $W(\boldsymbol{\theta})$, which convolves its transform: the mode $\boldsymbol{\ell}$ of the masked shear is a mixture of modes $\boldsymbol{\ell}'$ of the true one, each carrying the phase $e^{2i\varphi_{\ell'}}$ of its own wavevector. Undoing it with the phase $e^{-2i\varphi_{\ell}}$ of the new wavevector leaves the factor $e^{2i(\varphi_{\ell'}-\varphi_{\ell})}$, whose imaginary part feeds B. The equation is derived in Section 12.19.1. The effect should be strongest next to a hole, where the missing sticks are close, and fade with distance.

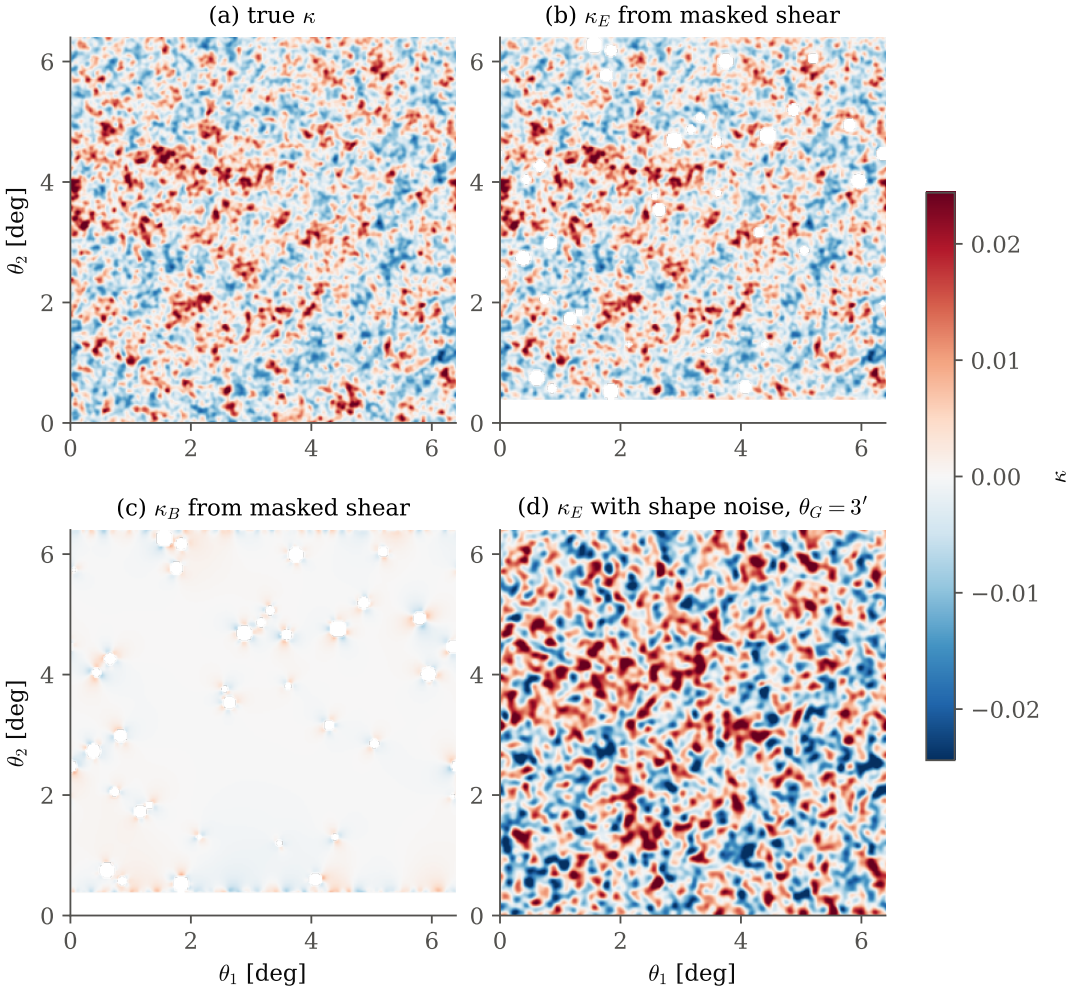


Figure 12.42: (a) A lognormal convergence map (1.5′ smoothing for display). (b) The E map reconstructed from its shear with the mask applied: away from the holes and the edge strip (grey) it is the map of panel (a). (c) The B map of the same masked shear: zero far from the mask, rings around every hole. (d) The E map from the full, unmasked shear with KiDS-like shape noise, smoothed with $\theta_G = 3'$: the strongest peaks of panel (a) survive, the rest of the structure is mixed with noise.

Figure 12.42 shows it, and the numbers confirm it. Within 3′ of a masked pixel the error of κ_E has an rms of 0.37 times the rms of the map, and κ_B has 0.36 of it; farther than 30′ from any masked pixel the two fall to 0.023 and 0.018. In power, the masked B map carries 2% of the E power at the lowest band ($\ell \approx 200$) and 0.9% at the highest ($\ell \approx 5000$), and the E power is lowered to 0.90 and 0.90 of its true value, against an observed fraction $f_{\rm obs}$ of the area: most of the loss is simply the missing area. For the topological summaries the lesson is local: features near holes are unreliable, and a research analysis either excludes a margin around the mask or, better, applies the same mask to every simulation, so that the distortion is part of the model;

Heydenreich and collaborators handle masked pixels with relative homology for the same reason [56, §2.3.3]. We sidestep the problem for the summaries of Section 12.20 by analysing full periodic patches, and keep the mask for the power spectrum, where Section 12.19.1 shows how to undo it.

12.18.5 Reproducing a published test: how much to smooth a mass map

Their question. The Dark Energy Survey published mass maps of its Year 3 footprint, $4143\,\mathrm{deg}^2$ of sky. Jeffrey, Gatti and collaborators asked which reconstruction method, and which amount of smoothing, gives the map closest to the truth [54]. The answer matters for anyone who wants to count peaks or voids in the map: too little smoothing and the map is noise; too much and the structures are washed out.

Their setup. A full-sky ray-traced convergence map of [51] played the true sky. They placed the DES Y3 galaxies on it, with the real positions, footprint and weights, $n_{\mathrm{eff}} = 5.59$ galaxies per square arcminute, and shape noise drawn by rotating the real ellipticities at random. They reconstructed κ_E by Kaiser–Squires on the sphere at HEALPix $N_{\mathrm{side}} = 1024$ (3.44′ pixels), and by three methods with priors, and smoothed the Kaiser–Squires map with Gaussians of standard deviation θ from 0 to 40′.

Their key equation. Their inversion is (12.63); their eq. (22) is its flat-sky form. Their score is the **Pearson correlation coefficient** between the true map κ and the reconstruction $\hat{\kappa}$ [54, eq. 35], the correlation coefficient of Chapter 1 computed over pixels, $r = \langle \kappa\hat{\kappa}\rangle / \sqrt{\langle\kappa^2\rangle\langle\hat{\kappa}^2\rangle}$ (both maps with zero mean). Why should r have a maximum as θ grows? Write the smoothed reconstruction mode by mode as $\hat{\kappa}_\ell = W_\ell(\kappa_\ell + n_\ell)$, with W_ℓ the transform of the smoothing kernel ($e^{-\ell^2\theta^2/2}$ for a Gaussian), C_ℓ the signal power and N the noise power. Different modes are uncorrelated, and the noise is uncorrelated with the signal, so each average is a sum over modes:

$$r \overset{(1)}{=} \frac{\sum_\ell W_\ell C_\ell}{\sqrt{\sum_\ell C_\ell}\,\sqrt{\sum_\ell W_\ell^2 (C_\ell + N)}} \overset{(2)}{=} \frac{\sum_\ell a_\ell b_\ell}{\sqrt{\sum_\ell C_\ell}\,\sqrt{\sum_\ell a_\ell^2}} \overset{(3)}{\le} \left[\frac{\sum_\ell C_\ell^2/(C_\ell + N)}{\sum_\ell C_\ell}\right]^{1/2}, \tag{12.70}$$

(1) Parseval's theorem turns each pixel average into a sum over modes: $\langle\kappa\hat{\kappa}\rangle$ picks up $W_\ell C_\ell$ from each mode, $\langle\hat{\kappa}^2\rangle$ picks up $W_\ell^2(C_\ell + N)$.

(2) Name $a_\ell \equiv W_\ell\sqrt{C_\ell + N}$ and $b_\ell \equiv C_\ell/\sqrt{C_\ell + N}$, so that $a_\ell b_\ell = W_\ell C_\ell$.

(3) The Cauchy–Schwarz inequality, $\sum ab \le \sqrt{\sum a^2}\sqrt{\sum b^2}$, with equality when $a_\ell \propto b_\ell$.

The bound is reached when $W_\ell \propto C_\ell/(C_\ell + N)$: pass a mode in proportion to the fraction of its power that is signal. That filter is the Wiener filter, which is why the Wiener reconstruction of Jeffrey and collaborators beats every smoothed Kaiser–Squires map in their Fig. 6. A Gaussian W_ℓ can only imitate it: it passes everything below $\ell \sim 1/\theta$ and nothing above. If θ is too small, it passes noise-dominated modes, which add to the denominator only; if too large, it removes signal-dominated modes from the numerator. The best θ puts the cut near the multipole where $C_\ell = N$, which moves to higher ℓ, so to smaller θ, as the galaxy density grows. That is the prediction to test.

What we compute. `w02_ks.py` makes 20 lognormal maps, adds the shape noise of three surveys (DES Y3, $n = 5.59$, $\sigma_e = 0.268$; our KiDS-like 6.2 and 0.27; a Stage-IV-like 30 and 0.27), inverts, smooths with θ from 0.75′ to 20′, and computes r against the true map at our 1.5′

resolution and against the true map smoothed to the size of an $N_{\mathrm{side}} = 1024$ pixel, which is what they compared with.

Listing 12.13: The Pearson test (`code/ch12/w02_ks.py`)

```
    for i in range(nmap):
        kt = wl.lognormal_map(PG, lam, rng.standard_normal((wl.N, wl.N)))
        t1, t2 = wl.shear_from_kappa(kt)
        # the truth at a HEALPix-like resolution (N_side = 1024 pixels are 3.4 arcmin wide)
        kt_hp = wl.smooth(kt, 3.4 / np.sqrt(12))
        white = rng.standard_normal((2, wl.N, wl.N))
        for name, (n, se) in surveys.items():
            s = wl.sigma_pix(n, se)
            e, _ = wl.kaiser_squires(t1 + s * white[0], t2 + s * white[1])
            for j, th in enumerate(thetas):
                es = wl.smooth(e, th)
                res[name][i, j] = pearson(es, kt)
                res_hp[name][i, j] = pearson(es, kt_hp)
```

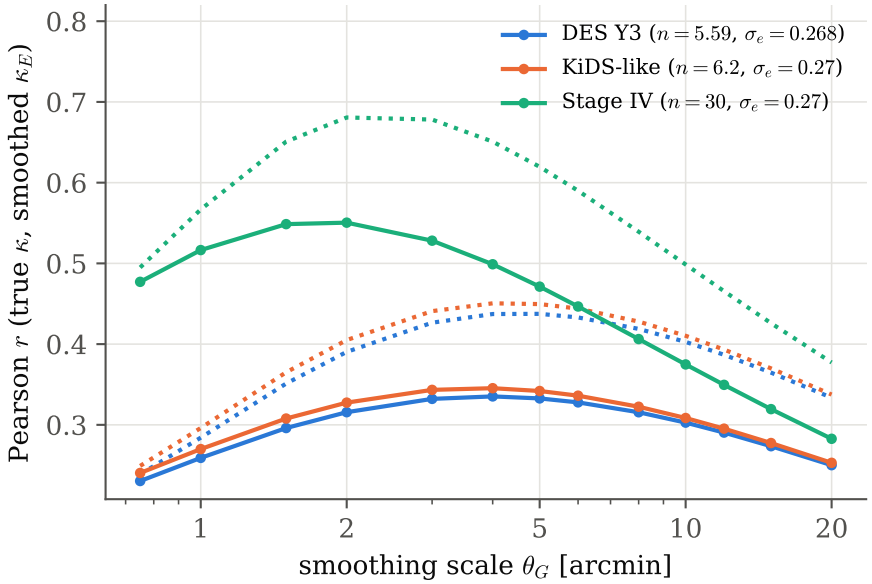


Figure 12.43: Pearson correlation between the true convergence and the smoothed Kaiser–Squires map, as a function of the smoothing scale, for three galaxy densities (solid: truth at our 1.5′ pixels; dotted: truth averaged to the size of a $N_{\mathrm{side}} = 1024$ pixel, as in the DES Y3 test). Each curve rises while smoothing removes noise faster than signal and falls once it removes signal; the optimum moves to smaller scales as the galaxy density grows.

Side by side. Figure 12.43 has the predicted shape. For the DES-like density our curve peaks at $\theta = 4'$ with $r = 0.34$; against the pixel-averaged truth at 5′ with $r = 0.44$. The KiDS-like case peaks at 4′ ($r = 0.35$), and the Stage-IV case at 2′ ($r = 0.55$). Jeffrey and collaborators find the Kaiser–Squires r rising from 0.20 unsmoothed to about 0.31 at 10′, nearly flat between 5′ and 15′, and slowly falling beyond [54, Fig. 6].

Why they differ. The shapes agree and the optimum lies in the same range, a few to ten arcminutes, with ours at the small end and our r a little higher. Three causes push in this direction. Their truth is a ray-traced N-body map, more strongly non-Gaussian than our lognormal one, with more of its signal in compact peaks that smoothing removes. Their footprint has edges and holes everywhere, and Section 12.18.4 showed how a mask lowers

the agreement near them; our patch has none. And their effective density varies over the footprint with the depth of the images, while ours is uniform. Their grid of smoothing scales $(0, 5, 10, 15, \dots)$ also cannot resolve a maximum this flat to better than 5′. The galaxy density moves the optimum as (12.70) predicts: from 4′ at $n = 5.59$ to 2′ at $n = 30$.

The lesson. The best smoothing scale is not a property of the method but of the data: it is set by where the signal stops beating the noise, and a mass map from a deeper survey deserves less smoothing. The summaries of Section 12.20 use $\theta_G = 3'$, close to the optimum for our density.

12.19 The two-point analysis: band powers, their errors, and the banana

Every lensing survey so far has drawn its headline number from a two-point statistic: KiDS-1000 from band powers and related statistics, DES Year 3 from the correlation functions $\xi_\pm$, HSC from power spectra (Section T1.7.6). Before asking what the peaks and the topology of the mass map can add, we need the number they must beat, computed in the same setting as everything else: our KiDS-like galaxy density, our source sample, our patches. That calls for three pieces, each met before on the CMB. The band powers must be measured on a patch with holes, which is the pseudo-C_ℓ problem of Section 8.9.3, now for a spin-2 field. Their errors must be known, which is the covariance problem of Section 8.13. And the errors must be turned into errors on (Ω_m, σ_8), which is the Fisher matrix of Section 10.3.

The question. On a KiDS-like survey of 1000 deg^2, can the convergence power spectrum be measured without bias through a mask, and how well does it then measure Ω_m, σ_8 and S_8?

12.19.1 Band powers through a mask: E leaks into B

Recall the pseudo-C_ℓ idea of Section 8.9.3: the power spectrum of a masked map is not the true spectrum, but its *average* is a known linear combination of the true band powers, through a coupling matrix that depends only on the mask; inverting that matrix (MASTER, Section 8.11) removes the bias. For a scalar CMB map the coupling comes from the window alone. Section 12.18.4 found a second effect for shear: the mask also moves power from E to B. We now derive both at once, on the flat periodic patch.

The masked shear, mode by mode. Let the true shear be pure E, $\hat\gamma(\boldsymbol{\ell}) = e^{2i\varphi_\ell}\hat E(\boldsymbol{\ell})$, where $\hat E$ is the transform of the real map κ (12.62). The mask $W(\boldsymbol{\theta})$, one where we see and zero where we do not, multiplies the shear, so its transform $\hat W$ convolves it. On the grid, with $N_{\rm pix}$ pixels, the convolution is a sum over the grid modes $\boldsymbol{\ell}'$. Apply the inversion (12.64) to the masked shear, and call the result $X(\boldsymbol{\ell}) = \hat E_W(\boldsymbol{\ell}) + i\hat B_W(\boldsymbol{\ell})$, where $\hat E_W$ and $\hat B_W$ are the transforms of the two real maps κ_E and κ_B made from the masked shear:

$$
\begin{aligned}
X(\boldsymbol{\ell}) &\overset{(1)}{=} e^{-2i\varphi_\ell}\frac{1}{N_{\rm pix}}\sum_{\boldsymbol{\ell}'}\hat W(\boldsymbol{\ell}-\boldsymbol{\ell}')\,e^{2i\varphi_{\ell'}}\hat E(\boldsymbol{\ell}') = \frac{1}{N_{\rm pix}}\sum_{\boldsymbol{\ell}'}\hat W(\boldsymbol{\ell}-\boldsymbol{\ell}')\,e^{2i(\varphi_{\ell'}-\varphi_\ell)}\hat E(\boldsymbol{\ell}'),\\
\hat E_W(\boldsymbol{\ell}) &\overset{(2)}{=} \tfrac12\big[X(\boldsymbol{\ell}) + X(-\boldsymbol{\ell})^*\big] \overset{(3)}{=} \frac{1}{N_{\rm pix}}\sum_{\boldsymbol{\ell}'}\hat W(\boldsymbol{\ell}-\boldsymbol{\ell}')\cos\big[2(\varphi_{\ell'}-\varphi_\ell)\big]\hat E(\boldsymbol{\ell}'),\\
\hat B_W(\boldsymbol{\ell}) &\overset{(4)}{=} \tfrac{1}{2i}\big[X(\boldsymbol{\ell}) - X(-\boldsymbol{\ell})^*\big] = \frac{1}{N_{\rm pix}}\sum_{\boldsymbol{\ell}'}\hat W(\boldsymbol{\ell}-\boldsymbol{\ell}')\sin\big[2(\varphi_{\ell'}-\varphi_\ell)\big]\hat E(\boldsymbol{\ell}').
\end{aligned}
\tag{12.71}
$$

(1) The transform of a product is the convolution of the transforms (with the $1/N_{\rm pix}$ of the discrete transform); then multiply by the phase of the inversion.

(2) A real map has a transform with $\hat{f}(-\boldsymbol{\ell}) = \hat{f}(\boldsymbol{\ell})^*$. X is the transform of the complex map $\kappa_E + i\kappa_B$, and the transform of its real part is the average of $X(\boldsymbol{\ell})$ and $X(-\boldsymbol{\ell})^*$.
(3) In $X(-\boldsymbol{\ell})^*$ rename $\boldsymbol{\ell}' \to -\boldsymbol{\ell}'$ and use $\hat{W}(-\boldsymbol{q})^* = \hat{W}(\boldsymbol{q})$ and $\hat{E}(-\boldsymbol{\ell}')^* = \hat{E}(\boldsymbol{\ell}')$ (both maps are real), and $\varphi_{-\ell} = \varphi_\ell + \pi$, which leaves $e^{2i\varphi}$ unchanged. It becomes the same sum with $e^{-2i(\varphi_{\ell'} - \varphi_\ell)}$, and $\frac{1}{2}(e^{ix} + e^{-ix}) = \cos x$.
(4) The imaginary part in the same way, with $\frac{1}{2i}(e^{ix} - e^{-ix}) = \sin x$.

Each true mode $\boldsymbol{\ell}'$ feeds the masked mode $\boldsymbol{\ell}$ with the window's weight $\hat{W}(\boldsymbol{\ell} - \boldsymbol{\ell}')$, split between E and B by the angle between the two wavevectors. Without a mask $\hat{W}$ is a spike at zero, only $\boldsymbol{\ell}' = \boldsymbol{\ell}$ contributes, the angle is zero, and everything stays in E. With a mask, nearby wavevectors at an angle leak a fraction $\sin^2$ of their power into B.

The average power. Different true modes are uncorrelated, $\langle \hat{E}(\boldsymbol{\ell}')\hat{E}(\boldsymbol{\ell}'')^* \rangle = N_{\rm pix} A_{\rm pix}^{-1} C_{\ell'}\, \delta_{\boldsymbol{\ell}'\boldsymbol{\ell}''}$ on the grid (7.19). Squaring (12.71) and averaging keeps only the diagonal terms:

$$\langle \tilde{C}_\ell^E \rangle = \sum_{\boldsymbol{\ell}'} w_2(\boldsymbol{\ell} - \boldsymbol{\ell}') \cos^2\big[2(\varphi_{\ell'} - \varphi_\ell)\big] C_{\ell'}, \qquad \langle \tilde{C}_\ell^B \rangle = \sum_{\boldsymbol{\ell}'} w_2(\boldsymbol{\ell} - \boldsymbol{\ell}') \sin^2\big[2(\varphi_{\ell'} - \varphi_\ell)\big] C_{\ell'}, \tag{12.72}$$

with $w_2(\boldsymbol{q}) \equiv \big|\hat{W}(\boldsymbol{q})\big|^2 / N_{\rm pix}^2$ the power of the window. Three checks. Without a mask w_2 is one at $\boldsymbol{q} = \boldsymbol{0}$ and zero elsewhere, and (12.72) returns C_ℓ in E and nothing in B. By Parseval, $\sum_{\boldsymbol{q}} w_2(\boldsymbol{q}) = N_{\rm pix}^{-1} \sum_{\boldsymbol{\theta}} W^2 = f_{\rm obs}$, the observed fraction, so a flat spectrum $C_\ell = N$ put into both E and B (shape noise, Section 12.18.2) comes out as $f_{\rm obs} N$ in each, whatever the angles, since $\cos^2 + \sin^2 = 1$: the noise bias of a masked map is simply $f_{\rm obs} N$. And the formula is the flat-sky version of the spin-2 coupling kernels of the MASTER family [168], the scalar case of Eq. (8.60) with the cosine and sine as the only new ingredient. Binning (12.72) into bands gives four matrices $M_{bb'}^{EE}$, $M_{bb'}^{BE}$, and, by the 45° rotation of Section 12.18.1, $M^{BB} = M^{EE}$ and $M^{EB} = M^{BE}$; the decoupled band powers are the measured pseudo band powers, noise subtracted, multiplied by the inverse of the 2×2 block matrix, as in (8.68).

Listing 12.14: The coupling matrices from the exact average (`code/ch12/w03_twopoint.py`)

```
def coupling_exact(mask, P):
    """The same matrices from the exact expectation (flat-sky MASTER):
    <p_E(l)> = sum_l' w2(l - l') cos^2[2(phi_l - phi_l')] P(l'),  w2 = |FFT(W)|^2 / N_pix^2,
    and sin^2 for B; cos^2 = (1 + cos 4 dphi)/2 turns each sum into two FFT convolutions."""
    npix = mask.size
    w2 = np.abs(np.fft.fft2(mask.astype(float))) ** 2 / npix**2
    W2x = np.fft.ifft2(w2)
    e4 = wl.E2IPHI**2

    def conv(g):                                  # sum_l' w2(l - l') g(l') on the periodic grid
        return npix * np.fft.fft2(W2x * np.fft.ifft2(g))

    nb = len(EDGES) - 1
    idx = np.digitize(wl.LMAG, EDGES) - 1
    MEE = np.zeros((nb, nb))
    MBE = np.zeros((nb, nb))
    for j in range(nb):
        Pj = P * (idx == j)
        c0 = conv(Pj).real
        c4 = (e4 * conv(Pj * np.conj(e4))).real
        pE, pB = 0.5 * (c0 + c4), 0.5 * (c0 - c4)
        for b in range(nb):
            MEE[b, j] = pE[idx == b].mean()
            MBE[b, j] = pB[idx == b].mean()
    return np.block([[MEE, MBE], [MBE, MEE]]), MEE, MBE
```

`coupling_exact` evaluates (12.72) without a double sum over modes. The identity $\cos^2 x = \frac{1}{2}(1+\cos 2x)$ turns the $\cos^2[2\Delta\varphi]$ into $\frac{1}{2}[1+\mathrm{Re}\, e^{4i(\varphi_{\ell'}-\varphi_\ell)}]$, and each term is a convolution of w_2 with C times a phase, done by FFT in line 70 (the convolution theorem again, now used in reverse). Line 79 is the $e^{4i\varphi}$ term, lines 80–83 add or subtract the two and average over each band. The bands are the eight analysis bands from $\ell = 112$ to 5000 plus one below and one above, since power outside the analysed range also leaks into it. As a check the same matrices are measured by brute force (`coupling`): 100 Gaussian maps with power in one band at a time, masked, inverted, measured. The diagonals agree to 2.4%, the Monte Carlo scatter of 100 maps.

Example: what the coupling matrix says. For our mask ($f_{\rm obs} = 0.907$) the largest off-diagonal entry of M^{EE}, relative to the diagonal, is 2%, and the largest $E \to B$ entry is 1.3% of the diagonal: a mask of small holes couples neighbouring bands only weakly and leaks about one per cent of E into B, the same order as the 2% to 0.9% found from the maps in Section 12.18.4.

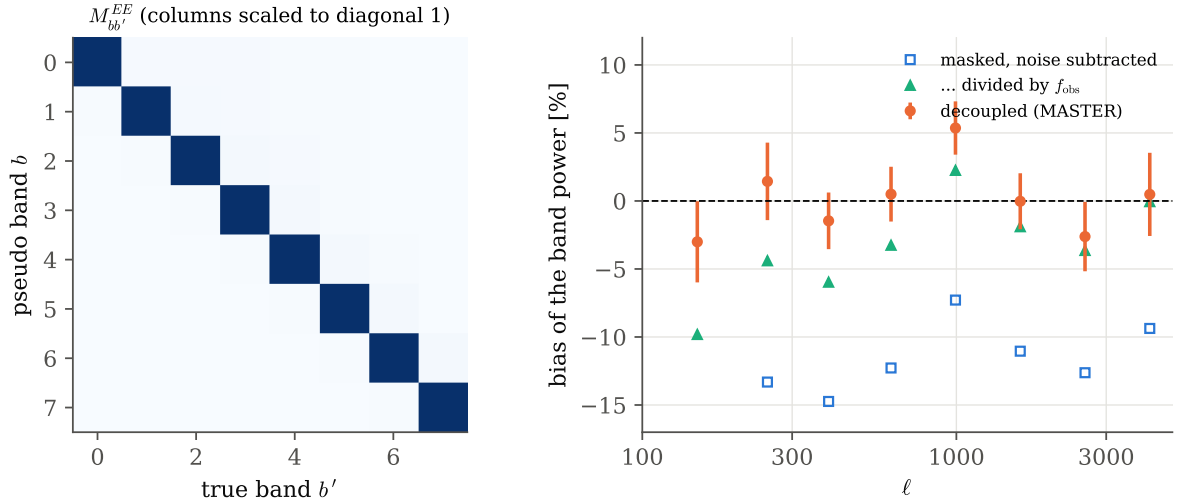


Figure 12.44: Left: the coupling matrix M^{EE} of the masked patch, each column divided by its diagonal entry; it is nearly diagonal. Right: the bias of the band powers over 200 lognormal maps with KiDS-like shape noise. Masked and noise-subtracted (squares): between −18% and −7%, mostly the missing area. Divided by $f_{\rm obs}$ (triangles): up to 10%, varying with ℓ. Decoupled by the inverse coupling matrix (circles, with the Monte Carlo error of the mean): consistent with zero.

Example: the recovered spectrum. `w03_twopoint.py` makes 200 lognormal maps with KiDS-like noise, masks them and measures the band powers three ways (Figure 12.44). The raw masked band powers, noise subtracted, are −18% to −7% off. Dividing by $f_{\rm obs}$, the simplest correction, leaves errors up to 10%. The decoupled band powers deviate by at most 5.4% with a Monte Carlo error of the mean up to 3.1%; over the 8 bands the χ^2 of the mean against the truth is 10.4. The decoupled B band powers give $\chi^2 = 6.9$ for 8 degrees of freedom: consistent with zero, which is the null test of (T1.7) passed on a masked map. The price of the mask is a scatter 1.07 times that of the unmasked patch, close to $1/\sqrt{f_{\rm obs}} = 1.05$: fewer independent modes, as in Section 8.4.2.

12.19.2 How large are the errors? Gaussian formula and simulations

The Gaussian band-power error of (T1.16) assumes that each Fourier mode is a Gaussian variable. On a flat patch of area A it reads as follows. The grid modes form a lattice of spacing $2\pi/L$, with $L^2 = A$, so a ring of radius ℓ and width $\mathrm{d}\ell$ holds $2\pi\ell\,\mathrm{d}\ell/(2\pi/L)^2 = \ell A\,\mathrm{d}\ell/2\pi$ modes.

The modes come in pairs $\pm\boldsymbol{\ell}$ with the same power (the map is real), so a band of n_b grid modes has $n_b/2$ independent ones. For a Gaussian mode the measured power $|\hat\kappa|^2$ follows an exponential distribution, whose standard deviation equals its mean $C_\ell + N$ (the χ^2_2 of (7.75), scaled). Averaging $n_b/2$ independent such powers,

$$\operatorname{Var}\hat C_b = \frac{(C_b+N)^2}{n_b/2} = \frac{2(C_b+N)^2}{n_b}, \qquad n_b \simeq \frac{A}{2\pi}\int_{\text{band}} \ell\,\mathrm{d}\ell. \tag{12.73}$$

On the full sky $A = 4\pi f_{\text{sky}}$ and $n_b \simeq \int (2\ell+1) f_{\text{sky}} \mathrm{d}\ell$, which is (T1.16) again: the same mode counting in a different coordinate system. Our bands hold from 28 grid modes (the lowest) to 15208 (the highest).

A lognormal field is not Gaussian, and its band powers are not exponential sums: a single bright spot puts power in many modes at once, so the modes are correlated and the variance is larger. `w05_costs.py` compares (12.73) with the scatter of the 1000 lognormal and 600 Gaussian fiducial maps of Section 12.20 (unmasked, with shape noise). For the Gaussian maps the ratio of the simulated to the predicted variance lies between 0.93 and 1.03, within the Monte Carlo error of a variance from 600 maps, $\sqrt{2/(n-1)} = 0.058$; the largest correlation between two bands is 0.11, also noise. For the lognormal maps the ratio runs from 0.98 to 1.04, within their own error of 0.045, and no two bands are correlated by more than 0.08, less than the noise correlation of the Gaussian maps (Figure 12.45): in these maps the excess is too small to see. Shape noise is what rescues the formula: where it dominates $C_b + N$, the noise, which is Gaussian and white, dilutes the non-Gaussian excess of the signal.

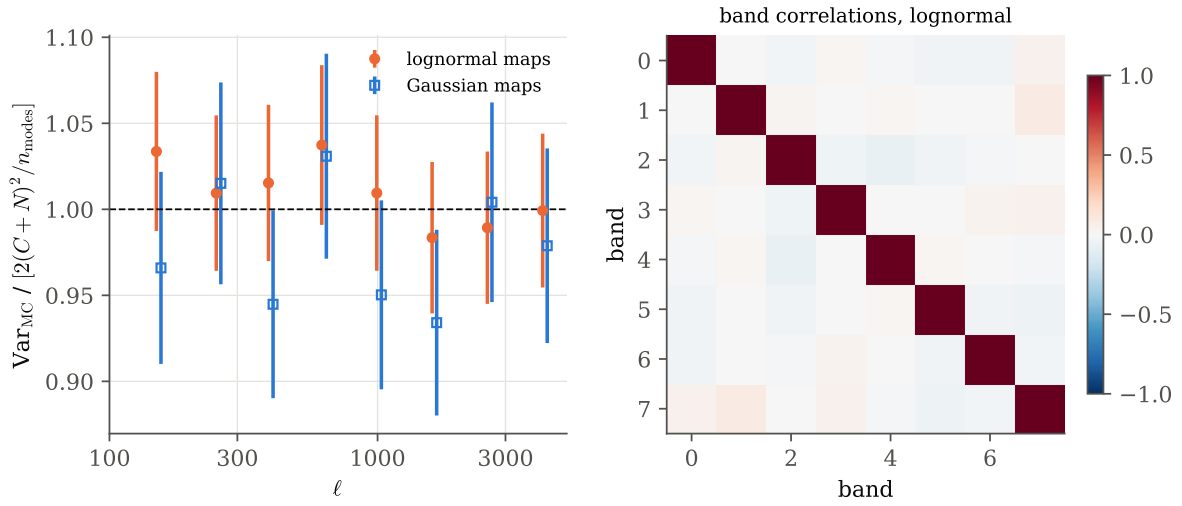


Figure 12.45: Left: the variance of each band power over the simulated maps divided by the Gaussian prediction (12.73), for lognormal and Gaussian maps with the same spectrum and the same shape noise. Right: the correlation matrix of the eight band powers of the lognormal maps. Both sets of maps agree with the formula within the Monte Carlo error, and the lognormal bands are no more correlated than noise allows: with this shape noise the non-Gaussian excess stays below what 1000 maps can detect.

This is the situation Section 8.13 described for the CMB, now with a physical cause: the analytic covariance is a good start, and simulations are needed for the part the analytic form leaves out. Research analyses add to the Gaussian term a non-Gaussian term from the trispectrum and a **super-sample** term, the response of all modes to density fluctuations larger than the survey, and check the sum against hundreds of simulations [45, 51]. Our forecast below uses the Gaussian formula, and Section 12.20 uses the simulated covariance throughout.

12.19.3 From band powers to (Ω_m, σ_8): the Fisher forecast

Recall the Gaussian Fisher matrix (10.20): for a data vector with mean $\boldsymbol{\mu}(\boldsymbol{\theta})$ and a covariance that does not depend on the parameters, $F_{ij} = \partial_i \boldsymbol{\mu}^{\mathsf{T}} \mathbf{C}^{-1} \partial_j \boldsymbol{\mu}$. With independent bands, $\mathbf{C}$ is diagonal with the variances (12.73), so

$$F_{ij} = \sum_b \frac{n_b}{2} \frac{\partial_i C_b \, \partial_j C_b}{(C_b + N)^2}, \qquad \boldsymbol{\theta} = (\Omega_m, \sigma_8), \tag{12.74}$$

the same expression as the signal-to-noise sum (T1.17) with the derivative of the amplitude replaced by the derivatives with respect to the parameters. The derivatives come from the Limber spectra at Ω_m and σ_8 moved by $\pm 5\%$, as central differences (Section 10.5), and the survey is $1000\,\mathrm{deg}^2$ with $n = 6.2\,\mathrm{arcmin}^{-2}$ and $\sigma_e = 0.27$ in one source bin.

Listing 12.15: The Gaussian band-power Fisher matrix (`code/ch12/w03_twopoint.py`)

```
87 def fisher(dC, C, N, edges, area_deg2):
88     """Gaussian band-power Fisher matrix: F_ij = sum_b n_modes,b/2 dC_i dC_j / (C+N)^2,
89     with n_modes,b = 2 pi l dl * A / (2 pi)^2 summed over the bin (mode counting on the
       sky)."""
90     A = area_deg2 * (np.pi / 180) ** 2
91     F = np.zeros((2, 2))
92     for lo, hi, in zip(edges[:-1], edges[1:]):
93         ell = np.arange(int(np.ceil(lo)), int(np.floor(hi)) + 1)
94         nm = ell * A / (2 * np.pi)          # modes per unit l: 2 pi l A /(2 pi)^2, both
       halves
95         w = nm / 2 / (C(ell) + N) ** 2
96         d = [f(ell) for f in dC]
97         for i in range(2):
98             for j in range(2):
99                 F[i, j] += np.sum(w * d[i] * d[j])
100    return F
```

Line 94 counts the modes per unit ℓ on the survey area, line 95 is the weight $n/2\,(C+N)^{-2}$, and the double loop adds the products of derivatives.

The answer. For $100 \le \ell \le 3000$ the marginal errors are $\sigma(\Omega_m) = 0.097$ and $\sigma(\sigma_8) = 0.145$, with a correlation coefficient $r = -0.999$ between them, and the error on S_8 is $\sigma(S_8) = 0.0218$, a precision of 2.6%. The individual errors are large and the combination is small: the ellipse is a long needle along the valley of Section T1.7 (Figure 12.46). Had Ω_m been known, σ_8 would be measured to 0.0069; freeing it inflates the error by the factor $1/\sqrt{1-r^2} = 21.2$ of (T1.19).

Example: why S_8 is not quite the needle. The long axis of the ellipse, in the logarithms of the two parameters, is the direction $\sigma_8 \Omega_m^{\alpha} = \text{const}$ with $\alpha = 0.58$, the exponent found multipole by multipole in Section T1.7.5 (there 0.57 at $\ell = 500$). The survey convention fixes $\alpha = 0.5$. Write

$$\ln S_8 \overset{(1)}{=} \ln \sigma_8 + 0.5 \ln \Omega_m + \text{const} \overset{(2)}{=} \ln\!\left(\sigma_8 \Omega_m^{\alpha}\right) - (\alpha - 0.5) \ln \Omega_m + \text{const},$$
$$\sigma(\ln S_8) \overset{(3)}{\approx} |\alpha - 0.5|\; \sigma(\ln \Omega_m) \approx 0.08 \times \frac{0.097}{0.315} \approx 0.025. \tag{12.75}$$

(1) The definition $S_8 = \sigma_8 (\Omega_m/0.3)^{0.5}$.
(2) Add and subtract $\alpha \ln \Omega_m$.
(3) The first term is the best-measured combination, nearly fixed; the second inherits the large error of Ω_m along the needle.

So $\sigma(S_8) \approx 0.025 \times 0.83 \approx 0.021$, which is most of the 0.0218: the error on S_8 is set less by the width of the needle than by the small tilt between the needle and the line of constant S_8. A survey that quotes $\sigma_8\Omega_m^{\alpha}$ with its own α, as HSC does with $\alpha = 0.45$ [49], quotes the narrowest number its data allow.

The banana. The Fisher ellipse is the quadratic approximation of the log-likelihood at the fiducial point (Section 10.1). Away from it the valley curves, because the constant-spectrum line is $\sigma_8 \propto \Omega_m^{-\alpha}$, a curve in the (Ω_m, σ_8) plane. `w03_twopoint.py` also evaluates the Gaussian likelihood of the band powers on a 300×300 grid, with a power-law emulator $C_b(\Omega_m, \sigma_8) = C_b^{\rm fid}(\sigma_8/\sigma_8^{\rm fid})^{a_b}(\Omega_m/\Omega_m^{\rm fid})^{b_b}$ built from the same logarithmic derivatives. Its contours (Figure 12.46) follow the ellipse near the centre and bend into the banana familiar from the published contours. From the grid, with a flat prior over the plotted box, $\sigma(S_8) = 0.0210$ and $\sigma(\Omega_m) = 0.099$, close to the Fisher values: along S_8 the likelihood is Gaussian, and the curvature only bends the long direction.

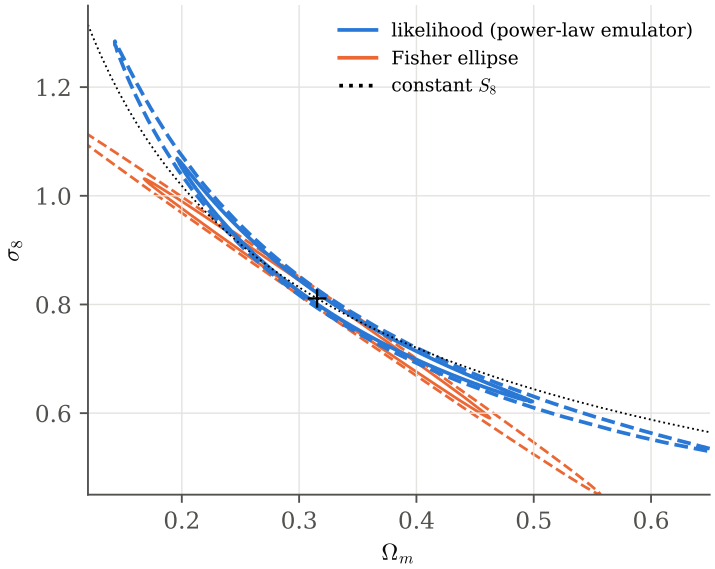


Figure 12.46: The 1σ and 2σ regions ($\Delta\chi^2 = 2.30, 6.18$) for (Ω_m, σ_8) from the convergence power spectrum of a KiDS-like 1000 deg^2 survey, $100 \le \ell \le 3000$, one source bin. The Fisher ellipse (orange) is the tangent approximation; the likelihood on a grid (blue) curves into a banana along the dotted line of constant S_8.

How the error depends on the scales used. With $\ell_{\rm max} = 1500$ instead of 3000 the forecast becomes $\sigma(S_8) = 0.0427$; with $\ell_{\rm max} = 5000$, 0.0170. The gain flattens because beyond the crossing multipole, where $C_\ell = N$ ($\ell \approx 286$ for our density, $\ell \approx 1064$ for a Stage-IV density of 30 arcmin^{-2}), each mode carries the factor $[C/(C+N)]^2$ of Section T1.8.4 and adds little.

12.19.4 Against the published surveys

Section T1.7.6 listed the published precisions: $\sigma(S_8) = 0.0225$ for KiDS-1000 (symmetrised), 0.024 for DES Year 3 and 0.0315 for HSC Year 1. Our 0.0218 for a KiDS-like survey is the same size. That agreement is a coincidence of errors that point in opposite directions, and it is worth taking apart, because each piece is a choice every analysis makes.

- *We use more scales.* KiDS-1000 uses band powers for $100 \le \ell \le 1500$ [45], HSC $300 \le \ell \le 1900$

[49]; small scales are cut because baryonic feedback and the non-linear model are uncertain there (Section T1.6). Cutting ours at 1500 doubles our error, to 0.0427.

- *We have no nuisance parameters.* Photometric-redshift shifts, the shear calibration m, intrinsic alignments and baryons are all marginalised in the real analyses. Each adds a direction in parameter space along which the spectrum can change, and (T1.19) says what that costs; Section 12.21 computes the cost of m.
- *We have no tomography.* KiDS-1000 splits its sources into five redshift bins. Each bin has its own exponent α (Figure T1.10), so the five needles are not parallel, and their intersection is much shorter than any one of them. This is what lets KiDS, with nuisances and a lower $\ell_{\max}$, reach 0.0225 while our single bin to $\ell = 1500$ reaches only 0.0427.
- *Our covariance is Gaussian.* The non-Gaussian variance, hidden by shape noise in Figure 12.45 but present in the signal, and the super-sample term both widen the real errors.

The lesson of the comparison: a forecast that matches a published error bar has not been validated by it; only a forecast that includes the same scales, nuisances and covariance terms can be compared number to number. What we can do honestly is compare summaries with each other under identical simplifications: whatever the nuisances do to the spectrum, they do to the peaks and the islands too.

12.20 Beyond two points: peaks, shapes and persistence of the mass map

The power spectrum of Section 12.19 pins down one combination of Ω_m and σ_8 and leaves the other almost free. Section 12.17.2 found a property of the map that responds to the two parameters differently: the lopsidedness of the convergence, whose reduced skewness is $3/\lambda$, set by the empty-beam shift and so by Ω_m alone. Two universes on the same S_8 line have the same spectrum, but the one with less matter, more clumped, has rarer and taller peaks and wider, shallower voids. The power spectrum cannot see this; it is blind to the phases of the Fourier modes, where that arrangement lives (Section 7.13). The peaks, islands and lakes of the map can.

The question. For the noisy signal-to-noise map of a KiDS-like survey, how well does each of four summaries (peak counts, Minkowski functionals, Betti curves, persistent Betti numbers) measure (Ω_m, σ_8), how much does each add to the power spectrum, and does the gain disappear, as it must, when the map is Gaussian?

12.20.1 Four summaries of one map

All four are computed from the same object: the κ_E map reconstructed from noisy shear (Section 12.18.1), smoothed with a Gaussian of $\theta_G = 3'$ and divided by the rms of the smoothed shape noise, $\sigma_{\rm noise} = 0.0102$ (12.65). This is the **signal-to-noise map** $\nu(\boldsymbol{\theta}) = \kappa_{E,s}(\boldsymbol{\theta})/\sigma_{\rm noise}$. The divisor is a fixed number, the same for every cosmology, so that a threshold $\nu = 2$ means the same convergence in every map, and the parameters can move the summaries. (Dividing each map by its own rms, as one does for a Gaussian field in Section 12.2, would remove the amplitude information we want to keep.) The thresholds are in the range $-1.5 \le \nu \le 3$, where the excursion sets are neither empty nor everything.

- *Peak counts.* A **peak** is a pixel higher than all eight of its neighbours. We histogram the heights ν of all peaks in eleven bins from -1 to 12. A peak is a local event; its count at

height ν is the number of maxima of the field there, the quantity whose Gaussian average the Kac–Rice method of Section 12.5.2 gives.

- *Minkowski functionals.* At each of ten thresholds, the area fraction V_0, the boundary length per unit area V_1 and the Euler characteristic per unit area V_2 of the excursion set $\{\nu \geq \nu_t\}$ (Section 12.4.1): thirty numbers.
- *Betti curves.* At the same thresholds, the number of islands β_0 and of lakes β_1 of the excursion set (Section 12.9.3): twenty numbers. In the plane $V_2 \times$ area $= \beta_0 - \beta_1$ at every threshold where the set does not wrap around the periodic patch.
- *Persistent Betti numbers.* From the persistence diagrams of the superlevel filtration (Section 12.9), for each pair of thresholds $\nu_{\rm lo} \leq \nu_{\rm hi}$ from the grid $\{-1, 0, 1, 2, 3\}$, the number of islands, and separately of lakes, that exist at every level between them.

Working definition 12.6 (Persistent Betti number)

Lower a threshold ν through a map. Each island (or lake) of the excursion set $\{\nu(\boldsymbol{\theta}) \geq \nu\}$ is born at a level b and dies at a level $d < b$ (Section 12.9). The **persistent Betti number**

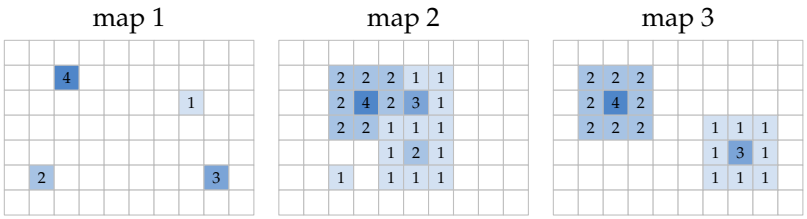
$$\beta_n(\nu_{\rm hi}, \nu_{\rm lo}) \equiv \#\big\{(b,d) \text{ in the diagram of dimension } n:\ b \geq \nu_{\rm hi} \text{ and } d < \nu_{\rm lo}\big\}, \qquad \nu_{\rm lo} \leq \nu_{\rm hi},$$

counts the features alive over the whole range $[\nu_{\rm lo}, \nu_{\rm hi}]$ [born at or above the upper level, not yet dead at the lower one]. For $\nu_{\rm lo} = \nu_{\rm hi} = \nu$ it is the ordinary Betti number $\beta_n(\nu)$ of (12.41).

This is eq. (10) of Heydenreich, Brück and Harnois-Déraps [56], translated. They lower nothing: they raise a threshold t through the signal-to-noise map and study the sets $\{\nu \leq t\}$, the **sublevel** filtration, in which a minimum of the map starts a component and a maximum starts a hole. Our superlevel filtration of ν is their sublevel filtration of $-\nu$, so their interval $[t_i, t_j)$ is our pair $(b, d) = (-t_i, -t_j)$ and their condition $t_i \leq t \leq t' < t_j$ is ours with $\nu_{\rm hi} = -t$, $\nu_{\rm lo} = -t'$. In our convention the peaks of the map are islands, which matches the way we think of a mass map.

What persistence sees that counts do not: three small maps. Heydenreich and collaborators illustrate the difference with three tiny maps [56, Fig. 4], redrawn in Figure 12.47: a 7×10 grid of signal-to-noise values, zero except in a few pixels. Map 1 has four isolated spikes of heights 4, 3, 2 and 1. Map 2 has the same four peak heights, but the 4, the 3 and the 2 sit on one plateau of 1s and 2s, with only the 1 isolated. Map 3 has two separate hills, a 4 on a ring of 2s and a 3 on a ring of 1s. Pixels touch through edges and corners, as in Section 12.2.2.

Work through them in the order of the question *what does each summary record? Peaks.* A pixel counts if it is higher than all eight neighbours. In map 2 the 3 (third row) has neighbours $2, 1, 1, 2, 1, 1, 1, 1$, all lower, so it is a peak; so is the 2 in the fifth row, surrounded by 1s, and the lone 1, surrounded by 0s. Maps 1 and 2 both have peaks of heights $4, 3, 2, 1$: identical peak counts. *Betti numbers.* Lower the threshold. In map 2 the set $\{\nu \geq 3\}$ is the 4 and the 3, two islands (the 2 between them is not yet dry); at $\nu = 2$ the 2s join the 4 and the 3 into one island while the lower 2 appears on its own, still two; at $\nu = 1$ the plateau is one island and the lone 1 another, still two; only the 4 is above 4. So $\beta_0 = 2, 2, 2, 1$ at $\nu = 1, 2, 3, 4$, the same as map 3, where the two hills stay apart at every level above zero. *Persistence.* The elder rule of Section 12.9 records when each island is absorbed. In map 2 the 3 is absorbed into the 4 at level 2, the pass between them, so its pair is $(3, 2)$; the 2 in the fifth row joins the plateau at level 1, $(2, 1)$; the lone 1 joins at 0, when the background goes dry, $(1, 0)$. In maps 1 and 3 every island survives until the background joins it at 0. The diagrams differ for all three maps. A single



	map 1	map 2	map 3
peak heights	$4, 3, 2, 1$	$4, 3, 2, 1$	$4, 3$
islands $\beta_0(\nu)$ at $\nu = 1, 2, 3, 4$	$4, 3, 2, 1$	$2, 2, 2, 1$	$2, 2, 2, 1$
island diagram (b, d)	$(4, \cdot), (3, 0), (2, 0), (1, 0)$	$(4, \cdot), (3, 2), (2, 1), (1, 0)$	$(4, \cdot), (3, 0)$
$\beta_0(\nu_{\text{hi}}{=}2, \nu_{\text{lo}}{=}1)$	3	1	2

Figure 12.47: Three signal-to-noise maps after [56, Fig. 4] (darker means higher). Peak counts cannot tell map 1 from map 2; Betti curves cannot tell map 2 from map 3; the persistence diagrams, and the persistent Betti number in the last row, differ for all three. The dot in $(4, \cdot)$ marks the island that never dies.

persistent Betti number already separates them: $\beta_0(2, 1)$ counts the islands born at or above 2 that are still separate at level 1, which is 3 in map 1 (the 4, 3 and 2), 1 in map 2 (only the 4; the 3 and the 2 merged before level 1) and 2 in map 3. Peaks record heights, Betti numbers record how many pieces there are at each level, and persistence records how the pieces are connected across levels: whether a peak stands alone or sits on the shoulder of a bigger one.

`w04_beyond.py` repeats the three maps by machine, with the union–find code of Section 12.11.1 and with its periodic version used for the simulated maps; the diagrams of the two *agree* exactly.

The periodic version needs one more idea, for the lakes. On a patch that wraps around, there is no open sea to tell a lake from the rest of the water, so the trick of Section 12.11.1 (surround the map by sea) is not available. What is available is a duality between land and water. *When is a lake born?* At the level s where the land closes a ring around a body of water, which is the level where, seen from the water's side, two water bodies stop being connected: the pass between them. *When does it die?* At the level m where the last water inside the ring dries up, which is the water body's lowest point. So a lake of the land with pair $(b, d) = (s, m)$ is the same object as a water body that, as the level is *raised*, appears at its bottom m and merges with another body at the pass s. Raising the level through ν is lowering it through $-\nu$, so the water bodies are the islands of $-\nu$, and the lakes of the map come from the same union–find run on $-\nu$, with four neighbours instead of eight (two water pixels that touch only at a corner are separated by the land pixels that touch there). The two rings that wrap around the periodic patch never close up and have no pair; they are left out. The persistence diagrams of all 1600 fiducial maps were also computed separately, from the same seeds, with `gudhi`'s `PeriodicCubicalComplex` (`w04b_gudhi_check.py`); the Betti and persistent Betti numbers of the two codes differ by at most 0.

On a simulated 256×256 map, β_0 from the persistence diagram matches the component count of `lib_topo` at all ten thresholds (largest difference 0), and $\beta_0 - \beta_1$ matches the cubical Euler characteristic of Section 12.2.3 at every threshold $\nu \geq 1$, where nothing wraps around the patch (largest difference 0).

Redundancy between the summaries. The diagonal of the persistent Betti numbers repeats the Betti curves, and V_2 repeats $\beta_0 - \beta_1$. In a combined data vector such exact repetitions make the covariance matrix singular, and its inverse meaningless. When persistent Betti numbers are combined with Minkowski functionals we therefore keep only their off-diagonal entries,

$\nu_{\rm lo} < \nu_{\rm hi}$, which hold the information the counts at single levels do not.

12.20.2 The machinery: simulated covariance, common random numbers, and a biased Fisher matrix

None of these summaries has a formula for its mean as a function of (Ω_m, σ_8) for a non-Gaussian map, nor for its covariance. Both come from simulations. The Fisher matrix of a data vector $\boldsymbol{s}$ with mean $\boldsymbol{\mu}(\boldsymbol{\theta})$ and covariance $\mathbf{C}$ is still $F_{ij} = \partial_i \boldsymbol{\mu}^{\mathsf{T}} \mathbf{C}^{-1} \partial_j \boldsymbol{\mu}$ (eq. 10.20, treating $\boldsymbol{s}$ as Gaussian, which a sum over many peaks or pixels nearly is), but now every ingredient is estimated, and each estimate has its own error.

The covariance. 1000 lognormal maps of the fiducial cosmology, each with independent shape noise, give the sample covariance $\hat{\mathbf{C}}$. Its inverse is biased high, and (8.78) gives the correction: multiply $\hat{\mathbf{C}}^{-1}$ by $(n-p-2)/(n-1)$, with n maps and p entries in the data vector [6]. For the largest data vector, $p = 65$, the factor is 0.93.

The derivatives. At each of 800 seeds we make four maps, with Ω_m and σ_8 in turn moved by $\pm 10\%$, and difference them: $\boldsymbol{D}_k = [\boldsymbol{s}(\theta_i + h_i) - \boldsymbol{s}(\theta_i - h_i)]/2h_i$ for seed k, and the derivative is the average over seeds. Why the same seed for all four? The scatter of a difference is

$$\operatorname{Var}\big[s(\theta+h) - s(\theta-h)\big] = \operatorname{Var} s(\theta+h) + \operatorname{Var} s(\theta-h) - 2\operatorname{Cov}\big[s(\theta+h), s(\theta-h)\big], \quad (12.76)$$

and if the two maps share their white noise and their shape noise, they differ only through the small change of cosmology, the covariance is almost as large as the variances, and the difference is almost noise-free. These are **common random numbers**, the antithetic pairs of Section 6.15.1 used the other way round: there a negative covariance shrinks the variance of a sum, here a positive one shrinks the variance of a difference. For the power spectrum the cancellation is nearly perfect. For counts it is not: a 10% change of σ_8 lifts some peaks across a bin edge and joins some islands, and these whole-unit jumps differ from seed to seed. That is why the steps are $\pm 10\%$ here, twice those of the power spectrum, at the cost of four more CAMB runs (`w04a_spectra.py`): the jumps grow in proportion to the step, the difference divided by $2h$ shrinks, and the noise of the derivative falls.

The bias this noise causes. Write the estimated derivative as the true one plus an error, $\hat{\boldsymbol{d}}_i = \boldsymbol{d}_i + \boldsymbol{\epsilon}_i$, with $\langle \boldsymbol{\epsilon}_i \rangle = 0$ and, for $N_{\rm der}$ independent seeds, $\langle \boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_j^{\mathsf{T}} \rangle = \boldsymbol{\Sigma}_{ij}/N_{\rm der}$, where $\boldsymbol{\Sigma}_{ij}$ is the covariance of the per-seed derivatives $\boldsymbol{D}_k$. The Fisher matrix is quadratic in the derivatives, and a quadratic form of a noisy vector is biased, exactly as the power of a noisy map is (8.19):

$$\begin{aligned} \big\langle \hat{\boldsymbol{d}}_i^{\mathsf{T}} \mathbf{C}^{-1} \hat{\boldsymbol{d}}_j \big\rangle &\overset{(1)}{=} \boldsymbol{d}_i^{\mathsf{T}} \mathbf{C}^{-1} \boldsymbol{d}_j + \boldsymbol{d}_i^{\mathsf{T}} \mathbf{C}^{-1} \langle \boldsymbol{\epsilon}_j \rangle + \langle \boldsymbol{\epsilon}_i \rangle^{\mathsf{T}} \mathbf{C}^{-1} \boldsymbol{d}_j + \big\langle \boldsymbol{\epsilon}_i^{\mathsf{T}} \mathbf{C}^{-1} \boldsymbol{\epsilon}_j \big\rangle \\ &\overset{(2)}{=} F_{ij} + \operatorname{Tr}\big[\mathbf{C}^{-1} \langle \boldsymbol{\epsilon}_j \boldsymbol{\epsilon}_i^{\mathsf{T}} \rangle\big] \overset{(3)}{=} F_{ij} + \frac{\operatorname{Tr}\big[\mathbf{C}^{-1} \boldsymbol{\Sigma}_{ji}\big]}{N_{\rm der}}. \end{aligned} \quad (12.77)$$

(1) Insert $\hat{\boldsymbol{d}} = \boldsymbol{d} + \boldsymbol{\epsilon}$ and multiply out.
(2) The cross terms vanish because $\langle \boldsymbol{\epsilon} \rangle = 0$; a number equals its own trace, and $\operatorname{Tr}(\boldsymbol{a}^{\mathsf{T}} \mathsf{M} \boldsymbol{b}) = \operatorname{Tr}(\mathsf{M} \boldsymbol{b} \boldsymbol{a}^{\mathsf{T}})$.
(3) The covariance of an average of $N_{\rm der}$ independent terms is that of one term divided by $N_{\rm der}$.

The bias is positive on the diagonal: noise in the derivatives looks like extra sensitivity. On its own it is small: for the largest data vector, 0.7% of the largest diagonal entry. But the quantity we care about is not on the diagonal. The ellipse is a needle because its two derivative vectors are nearly parallel. With correlation coefficient r between the two parameter errors, the smallest eigenvalue of $\mathbf{F}$, the one that measures the width across the needle, is a fraction of order $1 - r^2$ of

the diagonal (for two equal diagonal entries a and off-diagonal ra the eigenvalues are $a(1 \pm r)$, and $1 - |r| \approx (1 - r^2)/2$). For the power spectrum $r = -0.998$, so $1 - r^2 \approx 0.4\%$ and the smallest eigenvalue is about 0.2% of the diagonal. A bias of 0.7% of the diagonal is then larger than the true width: noise in the derivatives turns two parallel vectors into two vectors at an angle, and an angle is exactly what breaks a degeneracy. The symptom is unmistakable in the Gaussian control of Section 12.20.3: without a correction, the Minkowski functionals of Gaussian maps give a smaller error on S_8 than the power spectrum (0.0139 against 0.0164), and adding them to the spectrum multiplies the figure of merit by 1.48. Both are impossible, since for a Gaussian field the spectrum holds all the information (Section 7.2).

Watch out: A Fisher matrix from simulated derivatives is optimistic

Estimated derivatives add the positive term (12.77) to the Fisher matrix, and its effect on the error along a degeneracy can be larger than the true information. Use common random numbers, steps large enough that counts change by many units, and many seeds; subtract the bias; and bracket the answer with an estimate that is biased the other way.

We report three estimates. The *plain* one, $\hat{\boldsymbol{d}}^{\mathsf{T}}\hat{\mathbf{C}}^{-1}\hat{\boldsymbol{d}}$ with the Hartlap factor, is biased high. The *corrected* one subtracts (12.77), with $\boldsymbol{\Sigma}_{ij}$ estimated from the scatter of the per-seed derivatives. The *compressed* one splits the seeds in two halves: the first half defines two combinations of the data vector, $t_i = \hat{\boldsymbol{d}}_i^{(1)\mathsf{T}}\hat{\mathbf{C}}^{-1}\boldsymbol{s}$, one per parameter (a linear compression in the spirit of the score of Section 10.1.2), and the second half measures how t responds to the parameters; then the halves swap and the two results are averaged. Noisy compression vectors only lose information, and the second half's noise enters a two-entry vector instead of a p-entry one, so this estimate is biased low. The truth lies between the compressed and the plain values, and the corrected value is our best estimate.

12.20.3 What the simulations say

The signal-to-noise maps $\nu(\boldsymbol{\theta})$ of Section 12.20.1, the four summaries computed on them and the simulated Fisher machinery of Section 12.20.2 are all we need. `w04_beyond.py` makes 1000 fiducial lognormal maps for the covariance and 800 seeds of four shifted maps each for the derivatives; then the same again with Gaussian maps of the same spectrum (600 and 400), the control in which every gain over the power spectrum must vanish. That is 6400 maps of 256×256 pixels. Each patch is 6.4×6.4 deg; the Fisher matrices are multiplied by 24.4, the number of such patches in a 1000 deg^2 survey, treating the patches as independent.

How the summaries respond. Figure 12.48 shows, for four of the summaries, how the mean of each entry moves when $\ln \sigma_8$ or $\ln \Omega_m$ changes, in units of the scatter of that entry on one patch. More clumping (larger σ_8) moves peaks from the middle heights to the top: fewer peaks at $\nu \approx 1$, more above $\nu \approx 3$. The Euler characteristic and the island count follow the same pattern, since at high thresholds islands are almost all there is, and the lakes respond where the excursion set is mostly land, below $\nu \approx 0$. The two curves in each panel have nearly the same shape, the Ω_m curve about half the σ_8 curve.

Example: reading the degeneracy off the responses. Suppose every entry μ_a of a summary had $\partial\mu_a/\partial \ln \Omega_m = \alpha\, \partial\mu_a/\partial \ln \sigma_8$ with one number α for all a. Move the parameters by $\delta \ln \sigma_8$

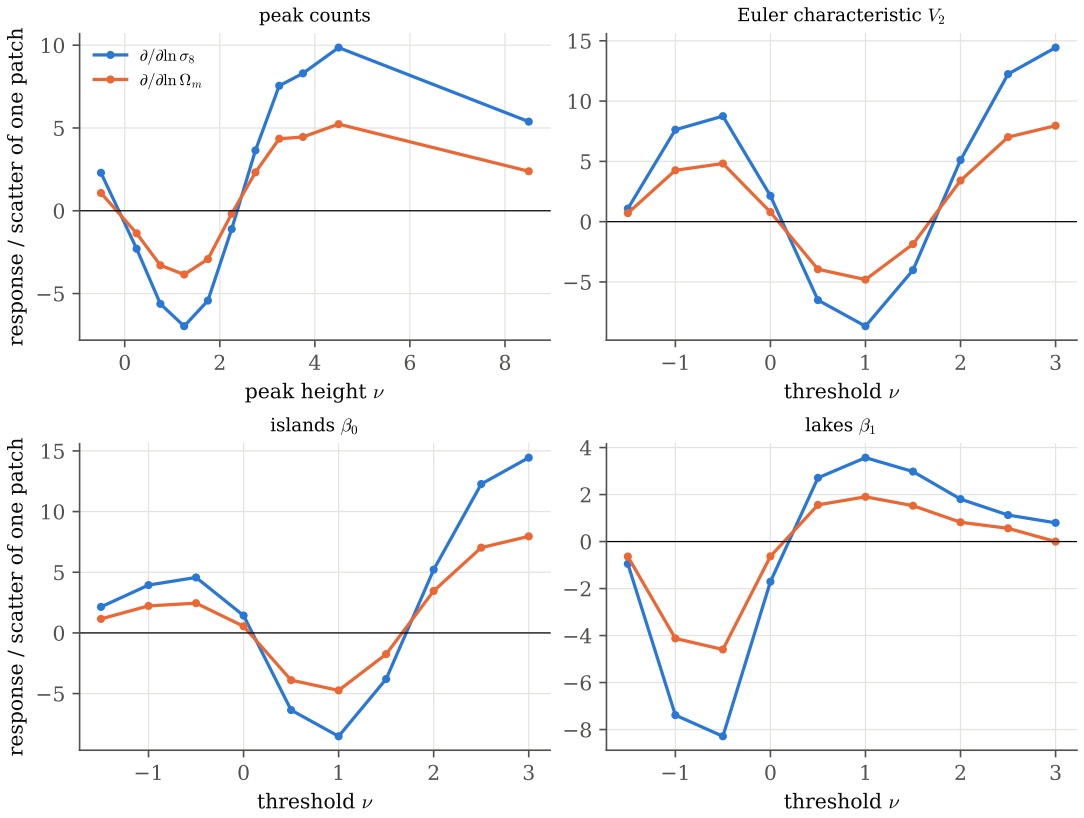


Figure 12.48: The response of four summaries of the signal-to-noise map to $\ln \sigma_8$ (blue) and $\ln \Omega_m$ (orange), each divided by the scatter of that entry on one 41 deg^2 patch. The orange curve is nearly the blue one halved: this is the S_8 degeneracy, entry by entry. The small departures from exact proportionality are what lets a summary measure the two parameters separately.

and $\delta \ln \Omega_m$:

$$\delta\mu_a \overset{(1)}{=} \frac{\partial \mu_a}{\partial \ln \sigma_8}\,\delta \ln \sigma_8 + \frac{\partial \mu_a}{\partial \ln \Omega_m}\,\delta \ln \Omega_m \overset{(2)}{=} \frac{\partial \mu_a}{\partial \ln \sigma_8}\left(\delta \ln \sigma_8 + \alpha\,\delta \ln \Omega_m\right) \overset{(3)}{=} \frac{\partial \mu_a}{\partial \ln \sigma_8}\,\delta \ln\left(\sigma_8 \Omega_m^{\alpha}\right). \tag{12.78}$$

(1) First-order Taylor expansion of the mean in the two log-parameters.
(2) The assumed proportionality of the two responses.
(3) $\delta \ln \sigma_8 + \alpha\,\delta \ln \Omega_m = \delta \ln(\sigma_8 \Omega_m^{\alpha})$.

Every entry would then move only through the single combination $\sigma_8 \Omega_m^{\alpha}$; a change along the line $\sigma_8 \Omega_m^{\alpha} = \text{const}$ would change nothing, and the Fisher matrix would have a zero eigenvalue. With the ratio of about one half read from Figure 12.48, the combination is close to $\sigma_8 \Omega_m^{0.5} \propto S_8$: the peaks, islands and lakes of the map feel the same degeneracy as the power spectrum, for the same reason (the amplitude of the convergence is set mainly by S_8, Section T1.7). Whatever a summary measures beyond S_8 comes from the entries where the ratio differs from one half, which is why the derivative noise of (12.77), which also changes that ratio, is so dangerous.

The numbers. Table 12.1 lists, for each summary and for its combination with the power spectrum, the forecast error on S_8 and the figure of merit $1/\sqrt{\det \boldsymbol{\Sigma}}$ of Section 10.4.6, where $\boldsymbol{\Sigma} = \mathbf{F}^{-1}$ is the 2×2 covariance of (Ω_m, σ_8), divided by that of the power spectrum alone. All use the corrected estimate of Section 12.20.2; the plain estimate of the Gaussian control is shown too, to make the bias visible.

	lognormal maps		Gaussian maps (control)		
summary	$\sigma(S_8)$	FoM/FoM$_{C_\ell}$	$\sigma(S_8)$	FoM/FoM$_{C_\ell}$	plain estimate
C_ℓ (8 bands)	0.0164	1	0.0164	1	1
peaks (11)	0.0130	0.77	–	–	0.23
Minkowski (30)	0.0127	1.00	0.0365	0.32	0.97
Betti (20)	0.0184	0.50	0.0226	0.39	0.63
persistent Betti (26)	0.0173	0.49	0.0357	0.21	0.49
C_ℓ + peaks	0.0112	1.55	0.0163	1.00	1.08
C_ℓ + Minkowski	0.0111	1.59	0.0148	1.10	1.48
C_ℓ + Betti	0.0132	1.29	0.0148	1.11	1.31
C_ℓ + persistent Betti	0.0131	1.30	0.0147	1.11	1.30
all four	0.0109	1.63	0.0147	1.10	1.63

Table 12.1: Forecasts for (Ω_m, σ_8) from a KiDS-like 1000 deg^2 survey, one source bin, from each summary of the signal-to-noise map and from its combination with the power spectrum. The number of entries of each data vector (lognormal case) is in brackets; entries that never change are dropped. FoM is the figure of merit $1/\sqrt{\det \boldsymbol{\Sigma}}$; for the power spectrum it is 2321 (lognormal) and 2207 (Gaussian). The last column is the FoM ratio of the Gaussian control before the derivative-noise bias is subtracted. For Gaussian maps the corrected peak matrix is not positive: its measured width across the needle is pure noise.

The control first. For Gaussian maps, adding any summary to the power spectrum must leave the figure of merit unchanged, because the spectrum holds all the information about a Gaussian field. The corrected estimates give gains between 1.00 and 1.11, above one by up to about ten per cent: that excess is the residual error of our estimator, and any lognormal gain smaller than it means nothing. The plain estimates give gains up to 1.48, and would have announced a discovery. On their own, the counts of a Gaussian map are much weaker than its spectrum (FoM ratios between 0.21 and 0.39, and for the peaks so little that the corrected matrix is not even positive): they see the amplitude through the heights in units of a fixed $\sigma_{\rm noise}$, but they lose the scale information of eight separate bands.

Now the lognormal maps. Each summary alone is roughly as strong as the power spectrum: the Minkowski functionals match its figure of merit (1.00), and peaks fall a little short of it (0.77) but give a smaller error on S_8 (0.0130 against 0.0164, 20% less). The Betti curves and the persistent Betti numbers are weaker, about half the figure of merit of the spectrum. Combined with the spectrum, every summary helps by far more than the ten per cent the control allows: peaks multiply the figure of merit by 1.55, Minkowski functionals by 1.59, Betti and persistent Betti numbers by about 1.29, and all four together by 1.63, with $\sigma(S_8)$ down by 33%. Figure 12.49 shows why the combination does so much more than the parts. All the ellipses are needles along the same S_8 line; they differ slightly in angle, and two needles at a slight angle cross in a short segment.

Where the extra information comes from. Section 12.17.2 gave the answer before any simulation: the reduced skewness of a lognormal map is $3/\lambda$, and the empty-beam depth λ depends on Ω_m but not on σ_8. Two universes on the same S_8 line have the same spectrum but differently lopsided maps. Every one of our summaries sees the lopsidedness: the peak heights through the long upper tail, the Minkowski functionals and Betti numbers through the asymmetry between high and low thresholds. That is the lever on the valley, and the Gaussian control, which has no lopsidedness, has no lever.

Example: how much a covariance from simulations costs. The Hartlap factor $(n-p-2)/(n-1)$ of (8.78) rescales the inverse covariance. For all four summaries together, $p = 65$ entries from

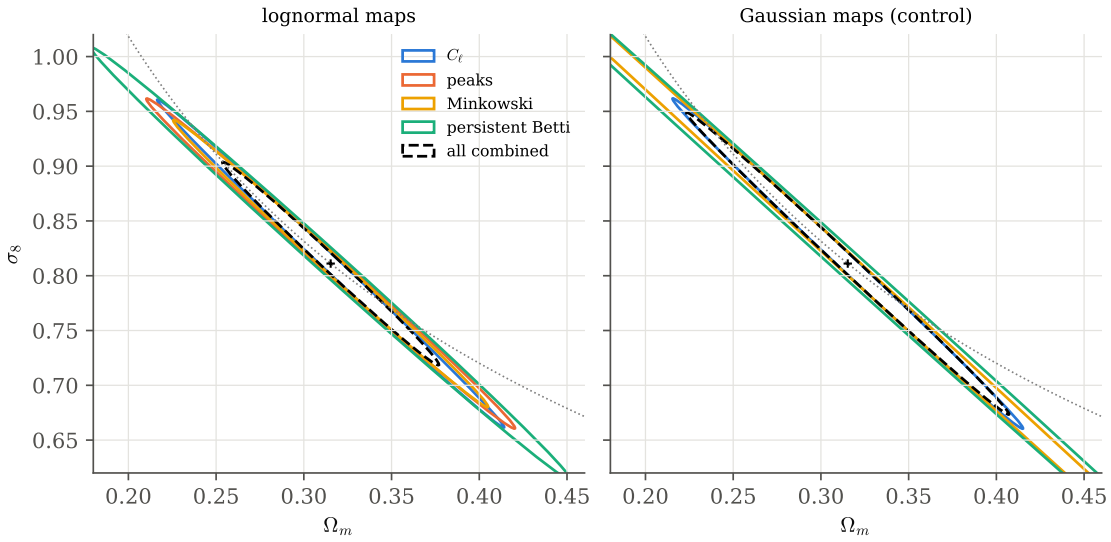


Figure 12.49: The 1σ Fisher ellipses ($\Delta\chi^2 = 2.30$) of (Ω_m, σ_8) for a $1000\,\mathrm{deg}^2$ survey, with the derivative-noise bias subtracted. Left: lognormal maps; the dashed ellipse, all four summaries together, is the shortest. Right: Gaussian maps of the same spectrum; the combination is hardly shorter than the power spectrum, as it must be. The dotted curve is the line of constant S_8. The Gaussian peak ellipse is missing because its corrected Fisher matrix is not positive.

$n = 1000$ maps:

$$\frac{n-p-2}{n-1} = \frac{1000-65-2}{999} \overset{(1)}{=} 0.934, \qquad \frac{\sigma_{\text{corrected}}}{\sigma_{\text{naive}}} \overset{(2)}{=} \left(\frac{n-p-2}{n-1}\right)^{-1/2} = 1.03.$$

(1) Arithmetic: 933/999.

(2) The Fisher matrix is proportional to $\hat{\mathbf{C}}^{-1}$, so it shrinks by the factor, and errors, which go as $\mathbf{F}^{-1/2}$, grow by its inverse square root.

A 3% correction. With the 126 simulations that Heydenreich and collaborators used for their covariance [56], the same data vector would have $(126-65-2)/125 = 0.47$: errors 46% larger, and the correction itself uncertain. That is why they kept their data vectors short and why they used the t-likelihood of Section 12.21 instead of the factor. Simulations for the covariance are the hidden price of every summary with many entries.

Example: what a figure-of-merit ratio means for the ellipse. For a 2×2 covariance $\boldsymbol{\Sigma}$ the ellipse $\Delta\chi^2 = q$ has area $\pi q\sqrt{\det\boldsymbol{\Sigma}}$ (Section 10.4.6), so the ratio of two figures of merit is the inverse ratio of the areas of their ellipses, at any confidence level. All four summaries together with the spectrum have 1.63 times the figure of merit of the spectrum alone, so their ellipse covers the fraction 1/1.63 of the area. Which dimension shrank? The area of a needle is its length times its width. The width, the error on the best-measured combination $\sigma_8\Omega_m^\alpha$, is set by the overall amplitude of the map, which the spectrum already measures well. The length is the error along the valley, which shows up as the error on Ω_m; it falls from 0.065 to 0.041. Almost all of the gain is a shorter needle. The error on S_8 falls by about the same fraction, from 0.0164 to 0.0109, because, as (12.75) showed, $\sigma(S_8)$ is mostly the small tilt between the needle and the line of constant S_8 times the needle's length.

Which peaks carry the information? Splitting the peak histogram into low ($\nu < 0.5$), medium ($0.5 \le \nu < 2.5$) and high ($\nu \ge 2.5$) peaks and computing the Fisher matrix of each part alone

gives figures of merit 147, 545 and 872 (errors on S_8 of 0.060, 0.020 and 0.014). The low peaks, which in a noisy map are mostly noise bumps in underdense regions, carry little. The rest is shared between the medium and the high peaks; in units of the convergence the high group starts at $2.5\sigma_{\rm noise} \approx 0.026$, well below the height of the peaks made by single massive clusters. The two papers that first asked the question looked at the same split, and their answer is compared with ours below.

Key idea

A summary beyond the power spectrum adds information only to the extent that the field is non-Gaussian. Measure the gain against a Gaussian control with the same spectrum, with the bias of simulated derivatives removed. For lognormal convergence maps of a KiDS-like survey every summary adds to the spectrum, all of them together multiplying the figure of merit by about 1.63; for Gaussian maps the gain is zero within the ten-per-cent error of the method.

12.20.4 Three published results, reproduced

Three papers asked, in turn, whether peaks are worth counting, whether they add to the two-point analysis, and whether persistence adds to the peaks. Each compared summaries by the area of a confidence region, so the one quantity to settle first is that area.

The yardstick. Dietrich and Hartlap define the figure of merit as one over the area of the 95% region in the (Ω_m, σ_8) plane [53]; Heydenreich and collaborators as one over the area of the 1σ region [56]; we use $1/\sqrt{\det \boldsymbol{\Sigma}}$. *How are they related?* For a Gaussian posterior the region is the ellipse $\Delta\chi^2 \le q$, of area $\pi q \sqrt{\det \boldsymbol{\Sigma}}$ (Section 10.4.6), and q is fixed by the probability it must hold. With two parameters $\Delta\chi^2$ follows the χ^2 distribution with two degrees of freedom, whose density is $\frac{1}{2}e^{-x/2}$ (Chapter 2), so

$$\mathrm{P}(\Delta\chi^2 > q) \overset{(1)}{=} \int_q^\infty \tfrac{1}{2} e^{-x/2}\,\mathrm{d}x \overset{(2)}{=} e^{-q/2},$$

$$q_{95\%} \overset{(3)}{=} -2\ln 0.05 = 5.99, \qquad q_{1\sigma} \overset{(4)}{=} -2\ln 0.317 = 2.30, \tag{12.79}$$

(1) The probability outside the ellipse is the tail of the χ^2_2 density.
(2) The antiderivative of $\frac{1}{2}e^{-x/2}$ is $-e^{-x/2}$.
(3) Set the tail to 5% and solve for q.
(4) "1σ" in two dimensions means the same 68.3% as in one, so the tail is 31.7%.

and the three figures of merit are $1/(\pi \times 5.99\sqrt{\det \boldsymbol{\Sigma}})$, $1/(\pi \times 2.30\sqrt{\det \boldsymbol{\Sigma}})$ and $1/\sqrt{\det \boldsymbol{\Sigma}}$. They differ by constant factors, so *the ratio* of the figures of merit of two summaries is the same in all three conventions, $\sqrt{\det \boldsymbol{\Sigma}_B / \det \boldsymbol{\Sigma}_A}$. That ratio is what we compare. The absolute values also depend on the survey area A: every Fisher matrix grows in proportion to A, so $\boldsymbol{\Sigma} \propto 1/A$, $\det \boldsymbol{\Sigma} \propto 1/A^2$, and the figure of merit grows in proportion to A.

Peaks, and what shape noise does to them

Their question. In 2009, cluster counts were a strong probe of dark energy, but finding clusters and weighing them brings selection effects. Kratochvil, Haiman and May asked whether one could skip the clusters and count the peaks of the lensing map itself, as a function of height, and still tell cosmological models apart [52].

Their setup. N-body simulations of three models with dark-energy equation of state $w = -0.8$, -1 and -1.2, with the same primordial amplitude, so that σ_8 differs (0.742, 0.798, 0.839). Convergence maps of 3×3 deg by ray tracing, for sources at $z_s = 2$; shape noise of 15 galaxies per arcmin^2 added pixel by pixel; Gaussian smoothing with $\theta_G = 1'$ in their convention (a standard deviation of 0.71′, Section 12.18.2); a peak is a pixel higher than its eight neighbours, as for us; peaks histogrammed in bins of κ; covariance from 200 pseudo-independent maps.

Their key equation. How well can a histogram of peak counts $\boldsymbol{N}$ separate two models? They use $\chi^2 = \Delta\boldsymbol{N}^\mathsf{T}\mathbf{C}^{-1}\Delta\boldsymbol{N}$ with $\Delta\boldsymbol{N}$ the difference of the mean histograms [52, eq. 14]. For two nearby models, separated by $\Delta\boldsymbol{\theta}$, this is our Fisher matrix:

$$\Delta\boldsymbol{N}^\mathsf{T}\mathbf{C}^{-1}\Delta\boldsymbol{N} \overset{(1)}{=} \Big(\sum_i \frac{\partial \boldsymbol{N}}{\partial\theta_i}\Delta\theta_i\Big)^\mathsf{T}\mathbf{C}^{-1}\Big(\sum_j \frac{\partial \boldsymbol{N}}{\partial\theta_j}\Delta\theta_j\Big) \overset{(2)}{=} \sum_{ij}\Delta\theta_i\, F_{ij}\,\Delta\theta_j. \tag{12.80}$$

(1) First-order Taylor expansion of the mean histogram.
(2) Collect the terms; $F_{ij} = \partial_i\boldsymbol{N}^\mathsf{T}\mathbf{C}^{-1}\partial_j\boldsymbol{N}$ is (10.20).

So "which peaks distinguish the models best" is "which peaks contribute most to $\mathbf{F}$", the question of which peaks carry the information, asked at the end of Section 12.20.3. Their noise level is the $\sigma_{\rm noise} = 0.0227$ worked out in the example of Section 12.18.2, from (12.65).

What they found, and what we compute. Shape noise raised the total number of peaks in a noise-free smoothed map from 1646 to 2432, by about 50%, and the peaks of medium height, $0 \lesssim \kappa \lesssim 0.06$, most of them not caused by any single cluster, carried most of the power to separate the models; the low peaks, $\kappa \lesssim 0$, carried little. `w04_beyond.py` counts the peaks of 50 lognormal maps with and without shape noise (Figure 12.50) and, from the full Fisher run, the information in low, medium and high peaks.

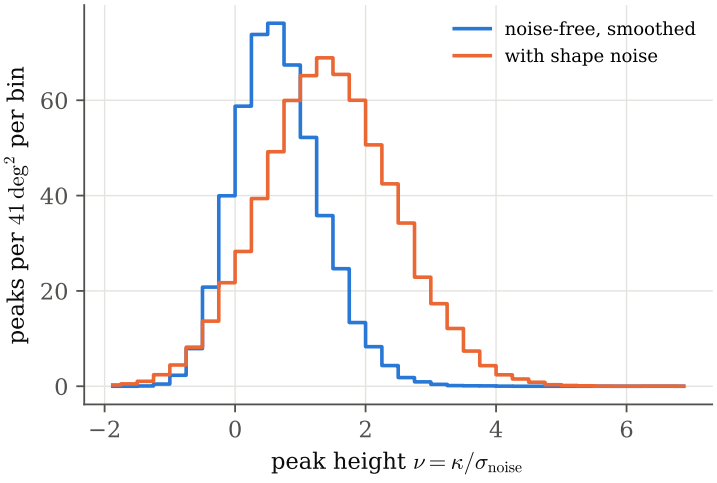


Figure 12.50: Mean histograms of peak heights on one 41 deg^2 lognormal patch, smoothed with $\theta_G = 3'$, without (blue) and with (orange) shape noise, heights in units of the smoothed noise rms. Noise adds peaks and pushes the existing ones up and down, so the histogram widens and shifts to larger ν.

Side by side.

	Kratochvil et al. (2010)	here
peaks per map, noise-free → noisy	1646 → 2432 (+48%)	490 → 685 (+40%)
smoothed noise rms $\sigma_{\rm noise}$	0.0227	0.0102
least informative peaks	low, $\kappa \lesssim 0$	low, $\nu < 0.5$ (FoM 147)
most informative peaks	medium, $0 \lesssim \kappa \lesssim 0.06$	$\nu \geq 0.5$ (FoM 545, 872)

Why they differ. The boost agrees in kind, +48% against +40%, and the difference of a few points is not something our single lognormal family can isolate: the smoothing is about four times wider in standard deviation than theirs, which changes the number of noise maxima per square degree (the Kac–Rice count of Section 12.5.2, a few θ_G^2 per maximum) and the share of the smoothed map that is noise, and the maps differ in the strength of their non-Gaussian peaks. Our split by height is in units of $\sigma_{\rm noise}$ rather than of κ; translated, our high bin starts at $\kappa \approx 0.026$, inside their medium range, and their result and ours agree that the information sits a few noise levels up, among peaks made by several structures along the line of sight, and not in the rare tall peaks of single clusters. Their maps are N-body maps of three dark-energy models, ours lognormal maps varied in (Ω_m, σ_8), so only these qualitative statements carry over.

The lesson. Shape noise does not just blur the peaks of the mass map, it makes new ones, and they sit on the signal; a peak count is only meaningful when the simulations it is compared with carry the same noise, processed the same way.

Peaks against, and with, cosmic shear

Their question. If most peaks are not clusters, are they just a noisy rewrite of the two-point function, or do they hold something else? Dietrich and Hartlap measured how well shear peaks constrain (Ω_m, σ_8) compared with cosmic shear, and how much the two do together [53].

Their setup. 158 N-body simulations spread over the (Ω_m, σ_8) plane and 35 more at the fiducial cosmology, ray-traced into 175 independent 6×6 deg fields; a CFHTLS-like survey of 180 deg^2 with 25 galaxies per arcmin^2 and total ellipticity dispersion $\sigma_\varepsilon = 0.38$ (that is 0.27 per component, the same as ours); peaks of the aperture-mass signal-to-noise map, a smoothed convergence map with a compensated filter; a Gaussian likelihood whose inverse covariance from the 175 fiducial fields carries the Hartlap factor [53, eq. 12], our (8.78).

Their key equation is the figure of merit, one over the area of the 95% region, which by (12.79) is $1/(\pi \times 5.99\sqrt{\det \boldsymbol{\Sigma}})$ for a Gaussian posterior. For a survey of 180 deg^2 instead of our 1000, our figures of merit are multiplied by 180/1000.

Side by side. Our power spectrum, peaks and their combination, in their convention and for their area:

	Dietrich & Hartlap (2010)		here	
	FoM	ratio to shear	FoM	ratio to C_ℓ
cosmic shear (C_ℓ here)	71	1	22	1
peaks	123	1.73	17	0.77
combined	173	2.44	34	1.55

The combined value is 1.41 times their peaks alone[1]; ours is 2.00 times our peaks alone.

Why they differ. *The absolute values*: their galaxy density is four times ours, so the noise power $\sigma_e^2/\bar{n}$ is four times smaller and more multipoles are signal-dominated (Section 12.19.3); their N-body maps carry the non-linear structure our lognormal maps lack; and their regions are the real, curved likelihood contours, cut by a flat prior, while ours are Fisher ellipses. *The ratios*: their peak statistic used almost exact source redshifts, while their cosmic shear used only two redshift bins, a mismatch they point out themselves; ours uses one source bin for both, and the peaks lose the scale-by-scale information of the spectrum. What survives is the shape of the result: peaks alone are comparable to the two-point analysis, and the combination beats both by a large factor, larger than either alone.

The lesson. Two summaries with the same degeneracy are not redundant: their needles differ by a small angle, and that angle, not the length of either needle, decides the gain of the combination.

Persistence against peaks

Their question. A peak count records the height of each maximum and nothing about its surroundings. Heydenreich, Brück and Harnois-Déraps asked whether the persistent Betti numbers of the signal-to-noise map, which also record how features are connected across levels (the three small maps of Figure 12.47), constrain cosmology better than peaks [56].

Their setup. Aperture-mass signal-to-noise maps of mock KiDS+VIKING-450 data (and, in a second test, of a 100 deg^2 Euclid-like survey with 30 galaxies per arcmin^2); their means emulated by a Gaussian process trained on the 26 cosmologies of the cosmo-SLICS; the covariance from 126 SLICS runs, with the t-likelihood of Sellentin and Heavens; four parameters (Ω_m, S_8, h, w_0) sampled by MCMC; persistent Betti numbers evaluated at the pairs of levels that best separate cosmologies, chosen on a dense grid.

Their key equation is the persistent Betti number of the working definition in Section 12.20.1, their eq. (10) translated into our superlevel convention, and the figure of merit, one over the area of the 1σ region, which by (12.79) has the same ratios as ours.

Side by side.

	Heydenreich et al. (2021)	here (lognormal)
FoM, persistent Betti / peaks	1.44 (KV450-like)	0.64
$\sigma(S_8)$, persistent Betti vs peaks	0.045 vs 0.048 (KV450) 0.027 vs 0.033 (Euclid)	0.0173 vs 0.0130
$\sigma(S_8)$, persistent Betti vs Betti	0.045 vs 0.048 (KV450)	0.0173 vs 0.0184

We do not reproduce their result. In our maps the persistent Betti numbers are no better than the Betti curves and weaker than the peaks.

[1]Their text says the combination has a figure of merit about 40% larger than cosmic shear alone; their Table 2 makes it 40% larger than the peaks alone and 2.4 times cosmic shear alone.

Why they differ. The main reason can be derived. Our map before noise is $\kappa = \lambda[e^{G-s^2/2} - 1]$ (12.56), an increasing function of a Gaussian field G. For any level $k > -\lambda$,

$$\{\boldsymbol{\theta} :\ \kappa(\boldsymbol{\theta}) \geq k\} \overset{(1)}{=} \left\{\boldsymbol{\theta} :\ e^{G(\boldsymbol{\theta})-s^2/2} \geq 1 + \frac{k}{\lambda}\right\} \overset{(2)}{=} \left\{\boldsymbol{\theta} :\ G(\boldsymbol{\theta}) \geq \frac{s^2}{2} + \ln\left(1 + \frac{k}{\lambda}\right)\right\}. \tag{12.81}$$

(1) Divide by $\lambda > 0$ and add 1.
(2) The logarithm is increasing, so it keeps the inequality.

Every superlevel set of the lognormal map *is* a superlevel set of the Gaussian map, at a relabelled level. So every island and lake, every merger and the pass where it happens, is the Gaussian field's; only the heights at which they happen are moved. The persistence diagram of a noise-free lognormal map is the Gaussian diagram with its axes stretched, and the stretch depends on the cosmology only through λ and s, the one-point distribution that peak heights already see. Smoothing and noise, applied after the exponential, blur this exact correspondence, but they do not create the thing persistence was designed to see. Heydenreich and collaborators trace their gain to exactly that thing: whether peaks stand alone or sit together on a larger overdensity, the environment of haloes in filaments and superclusters [56, §4], which N-body maps have and lognormal maps do not. Their gain also grew with the galaxy density, as it should if it comes from resolving that structure. Smaller differences push the same way: their levels were chosen for their power to separate cosmologies, ours form a fixed grid of five, whose diagonal repeats the Betti curves; and their persistent Betti errors include the error of their emulator, without which their gain on S_8 in the KV450-like case rises from 5% to 8%.

The lesson. A gain from topology is a property of the maps, not of the method. Lognormal maps are a fair test of the machinery (the covariance, the derivatives, the Fisher bias) and an unfair test of persistence, because they contain no structure for persistence to find that the one-point distribution does not already give.

The three reproductions together.

result	theirs	ours	main reason for the difference
noise boost of peak counts	+48%	+40%	wider smoothing
FoM, peaks / two-point	1.73	0.77	their peaks had redshifts, their shear two bins
FoM, combined / two-point	2.44	1.55	as above; density, N-body
FoM, persistent Betti / peaks	1.44	0.64	lognormal maps have no environment

Each published gain beyond the power spectrum appears in our maps exactly as far as the maps contain the structure behind it: the skewness of a lognormal field, yes; the clustering of haloes, no.

12.21 How it is done in research, and what our simplifications cost

Every number computed so far came from a chain that a real survey also runs, link for link, but with each link simplified until it fits in a few minutes on a laptop. The chain of Figure T1.8 runs from the matter field to the summaries, with nuisances entering at every step. Figure 12.51 sets our version beside the research version. For each link the surveys do something heavier, we did something cheaper, and the difference has a price in the width of the error bars.

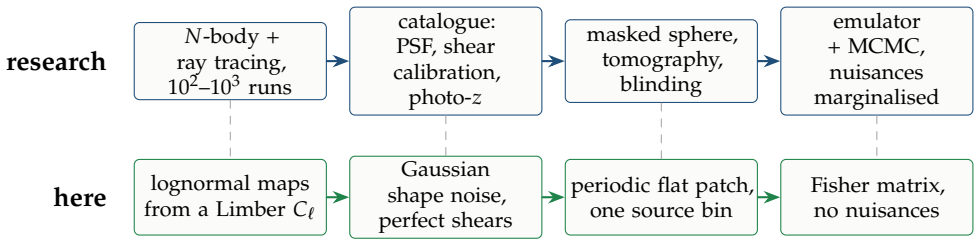


Figure 12.51: The four links of a weak-lensing analysis, as research does them (top) and as this part did them (bottom). Each dashed pair is one simplification; the text says what each costs.

The maps. The true sky is non-linear: haloes, filaments, voids, the lensing of lensing. Research analyses calibrate their summaries on ray-traced N-body simulations. Takahashi and collaborators produced 108 full-sky sets of convergence and shear maps for sources from $z = 0.05$ to 5.3, whose spectra match theory within 5% to $\ell = 3000$ [51]; the DES Y3 mass-map tests of Section 12.18.5 use one of them [54]. For summaries whose mean has no formula, a separate suite samples the parameters: Dietrich and Hartlap ran 158 cosmologies plus 35 fiducial runs, ray-traced into 175 independent fields [53]; Heydenreich and collaborators used the cosmo-SLICS, 26 cosmologies in a Latin hypercube, with a Gaussian-process emulator to interpolate between them, and 126 independent SLICS runs for the covariance [56]. Each run costs thousands of processor hours. We replaced all of this by shifted lognormal fields made from a Limber spectrum in milliseconds (Section 12.17.2). They have the right spectrum and the right kind of one-point skewness, too little of it (the empty-beam shift is about 60% too deep, Section 12.17.3), and none of the structure of real haloes. The cost is in the size of the non-Gaussian gains: ours are a statement about lognormal maps, and the real gains are expected to be larger.

The catalogue. A survey measures shapes, not shears. The telescope's point-spread function smears every image and must be modelled star by star and removed; the remaining errors are calibrated on image simulations into a multiplicative bias m per redshift bin, $\varepsilon \to (1 + m)\gamma$ (Section T1.6), whose prior is marginalised [45, 50]. The redshift distribution of each bin is estimated from photometry and calibrated against spectroscopic samples, and its uncertainty enters as shifts Δz_i or as many realisations of $n(z)$. Shape noise is measured, not assumed: HSC subtracts the noise power by rotating every galaxy's ellipticity by a random angle, which destroys the coherent shear and keeps the noise, and averaging many such rotations [49]; DES Y3 builds its mock noise the same way [54]. We assumed perfect shears, a known $n(z)$, and Gaussian noise of known amplitude. How much does the simplest of these nuisances cost? A multiplicative bias rescales the spectrum by $(1 + m)^2$, which is degenerate with the amplitude. With a Gaussian prior of width 0.02 on m, added as a third parameter to the Fisher matrix of Section 12.19.3 (Section 10.4.4), `w05_costs.py` finds $\sigma(S_8) = 0.0245$ instead of 0.0218; cut at $\ell_{\max} = 1500$, 0.0438 instead of 0.0427 (Figure 12.52).

The sky. Real footprints are masked spheres with thousands of holes; analyses either model the mask in the estimator (pseudo-C_ℓ, Section 12.19.1) or apply it to every simulation, and topological summaries treat the masked pixels explicitly, for example with relative homology [56, §2.3.3]. The sources are split into four or five tomographic bins, which shortens the S_8 needle (Section 12.19.4), and the analysis is **blinded**: the catalogue is distributed with secret shifts (in HSC, offsets in m known only to a designated person [49]; KiDS and DES use several blinded catalogues [45, 50]) so that no one can tune the analysis towards an expected S_8 before every test has been passed. Our patch is a periodic flat square without holes for the higher-order summaries, and we did no blinding because there were no data to fool us. A DES-like survey,

4143 deg^2 at $n = 5.59$, $\ell \leq 1500$, with the same m prior and one bin, would reach $\sigma(S_8) = 0.0243$ in our forecast.

The inference. Real analyses sample the posterior of cosmological and nuisance parameters together by MCMC (Chapter 9), with emulated or Boltzmann-code predictions, a covariance from simulations or an analytic model, and a likelihood that accounts for the noise of that covariance: Heydenreich and collaborators use the t-distribution of Sellentin and Heavens, $\ln \mathcal{L} = -\frac{N_s}{2} \ln[1 + \chi^2/(N_s - 1)]$ with $N_s = 126$ simulations [56, eq. 22], the exact version of the Hartlap correction. We computed Fisher matrices, which assume a Gaussian likelihood and are exact only for a linear model (Section 10.1); Figure 12.46 showed that for the power spectrum the curvature of the banana changes $\sigma(S_8)$ little. Our simulated Fisher matrices also carry the noise of their derivatives, which Section 12.20.2 measured and subtracted.

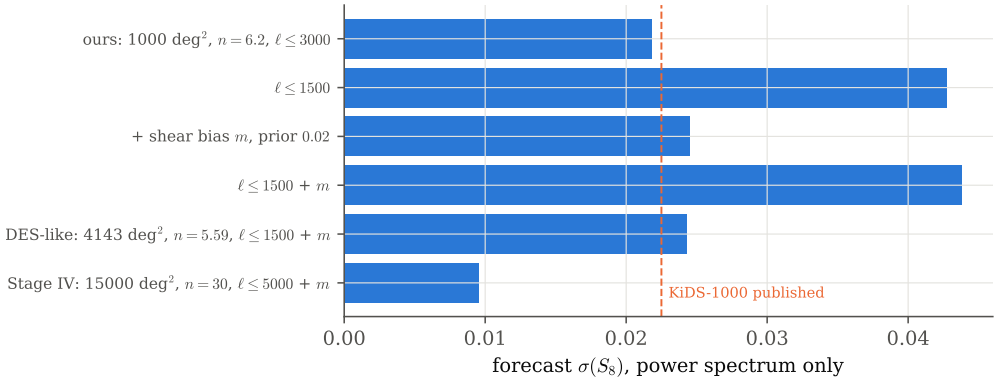


Figure 12.52: The forecast error on S_8 from the power spectrum alone, one source bin, Gaussian covariance, as the survey and the nuisances change. A multiplicative shear bias with a 2% prior, and a lower $\ell_{\max}$, each widen the error; a Stage-IV survey with 30 galaxies per arcmin^2 over 15 000 deg^2 narrows it, until the prior on m becomes the limit. The dashed line is the published KiDS-1000 precision, which uses five tomographic bins and marginalises many more nuisances.

Where the full version would go. A Stage-IV survey, 15 000 deg^2 at 30 galaxies per square arcminute and $\ell \leq 5000$, reaches $\sigma(S_8) = 0.0095$ in the same forecast, with the 2% prior on m; that number is set mostly by the prior, which is why Stage-IV shear calibration must be ten times better. For the higher-order summaries, the next step beyond our patch would be the same pipeline on HEALPix maps at $N_{\text{side}} = 256$–512 cut from the Takahashi simulations, with the survey's mask and noise, a few hundred covariance maps and an emulator over a few dozen cosmologies; at the cost of each simulation run, that is a cluster project and not a laptop one.

How this is used in real research: The higher-order lensing analyses

Peak counts have been measured in KiDS, DES and HSC data, and persistent homology, Minkowski functionals and scattering transforms are being prepared for Stage-IV surveys, always with the same structure: summaries computed identically on data and on N-body mocks that include the mask, the noise and the photometric redshifts; a covariance from hundreds of independent mocks with the Hartlap or Sellentin–Heavens correction; an emulator of the mean over a few dozen cosmologies; and an MCMC over cosmology and nuisances. The summaries themselves are the cheap part, a second per map, as here; the simulations are the expensive part [44, 56].

Seen from the pipeline that runs through the whole book, from a physical model to a field,

to a summary, to its sampling distribution and to the parameters, the weak-lensing analysis is the same chain at every scale: research spends its effort on the first link, realistic maps, and on the nuisances between the maps and the summaries, while the statistics of the last two links (covariance, Hartlap, Fisher, MCMC) is what we built here and does not change.

Recap

- A shifted lognormal map $\kappa = \lambda[e^{G-s^2/2} - 1]$ has the right spectrum and a floor at $-\lambda$, the empty beam; its reduced skewness is about $3/\lambda$, set by Ω_m and not by σ_8 (Section 12.17.2).
- Kaiser–Squires inversion is a rotation of each Fourier mode of the shear; it splits the map into κ_E and a κ_B that should vanish. Shape noise enters both with power $\sigma_e^2/\bar{n}$; smoothing with a Gaussian of width θ_G leaves $\sigma^2_{\rm noise} = \sigma_e^2/(4\pi\theta_G^2\bar{n})$, and a mask leaks E into B (Section 12.18).
- The band powers of a 1000 deg^2 KiDS-like survey measure S_8 to 0.0164 in our lognormal forecast and leave Ω_m and σ_8 separately loose: a needle along $\sigma_8\Omega_m^\alpha = \text{const}$, which the real likelihood bends into a banana (Section 12.19).
- Summaries without a formula need simulated means and covariances: the Hartlap factor for the inverse covariance, common random numbers for the derivatives, and the subtraction of the derivative-noise bias (12.77), without which a Gaussian control shows gains that cannot exist (Section 12.20.2).
- On lognormal maps, peaks, Minkowski functionals, Betti curves and persistent Betti numbers each add to the power spectrum, all together multiplying the figure of merit by 1.63; on Gaussian maps they add nothing (Section 12.20.3). Published peak gains are reproduced in kind; the persistence gain over peaks is not, because a lognormal map has the topology of a Gaussian one, relabelled (12.81) (Section 12.20.4).
- Research replaces each simplification: ray-traced N-body maps, PSF and shear calibration, photometric redshifts, masks, tomography, blinding, emulators and MCMC over the nuisances (Section 12.21).

12.22 A topological summary is a random number

In Section 12.3.1 we took one Gaussian patch, coloured everything above a threshold, and counted: so many pieces, so many holes, an Euler characteristic of so much. Run the same script with a different seed and every one of those numbers changes. Nothing went wrong. A cosmological theory never predicts the map we observe; it predicts the statistics of maps like it, and the sky hands us one draw. The counts we take from that draw are therefore draws too. Counting pieces and holes did not take us out of probability. It gave us a new random variable, and everything the earlier chapters built for random variables applies to it.

Write the parameters of the physical model as θ (for the moment one number, the amount of non-Gaussianity), the random field the model produces as $\phi(\boldsymbol{x};\theta)$, and the summary we compute from it, a Betti curve, an Euler characteristic curve or a persistence diagram, as $S[\phi]$. Repeating the experiment, new field, same recipe, gives a distribution of values,

$$P(S \mid \theta), \tag{12.82}$$

the **sampling distribution** of the summary. It is the bridge between topology and inference. Once we have it, the ordinary questions return unchanged, and they can be said in words before any symbol:

1. What values does the summary take, and how often, if the model with parameter θ is true?
2. What is its average, how much does it scatter from map to map, and how does it co-vary with another summary computed from the same map?
3. If the field were Gaussian, would a value like the one we measured, or one even further from the Gaussian average, be something we would reasonably expect?
4. Which values of θ could have produced the value we measured, and how readily would each have produced it?

The first two are prediction, the last two inference (Chapter 5). Bayes' theorem connects them exactly as before, $P(\theta \mid S) \propto P(S \mid \theta)\, P(\theta)$. Topology does not compete with Bayesian inference; it supplies a representation of the data on which Bayesian inference can act. Two separate questions are in play, and it helps to keep them apart. *What should we measure*, because it has physical meaning or is robust to what we consider nuisance? That is what Sections 12.1 and 12.9 were about. *How does the uncertainty in what we measured become a statement about θ?* That is Monte Carlo, covariance, likelihood, Fisher matrix and Markov chain, the subject of Chapters 6, 9 and 10. In the pipeline of the book, physical model → field → choose what matters → summary → sampling distribution → physics, the previous parts built the first three arrows; here we walk the last two.

Working definition 12.7 (sampling distribution of a summary)

The **sampling distribution** $P(S \mid \theta)$ of a summary S [a number or a vector of numbers computed from one map, such as β_0 at thirteen thresholds] is the distribution of the values of S over independent realisations of the field [new random phases and amplitudes, same model, same θ], each passed through the same observation and the same computation. *Operational test:* simulate n fields at θ, compute S on each, histogram; the mean, covariance and quantiles of that sample estimate those of $P(S \mid \theta)$ with errors that fall as $1/\sqrt{n}$.

12.22.1 The smallest field: nine pixels

Before any map, take a field small enough that we can list every realisation. A 3×3 grid of pixels; each pixel is "hot" (above threshold) independently with probability p, and cold with probability $q = 1 - p$. This is the excursion set of a field whose pixels are uncorrelated, at the threshold that leaves a fraction p of them hot. The hot set is the union of the closed hot squares, so two hot pixels that touch only at a corner belong to the same piece, and a hole is a group of cold pixels, joined through shared sides, that cannot reach the outside of the grid (the dual conventions of Section 12.2.2).

The question. How many pieces β_0 and holes β_1 does the hot set have, and with what probabilities? What is the average Euler characteristic $\chi = \beta_0 - \beta_1$?

The average Euler characteristic, by hand. The pieces depend on how pixels connect, which makes β_0 awkward to average directly. The Euler characteristic has a second form that never mentions connections: $\chi = V - E + F$ (12.6), where F counts hot pixels, E counts unit edges that belong to at least one hot pixel, and V counts lattice points that belong to at least one hot pixel. Each of V, E, F is a sum of yes/no indicators, one per cell of the grid, and the average of a sum is the sum of the averages whether or not the terms are independent. So we only need, for each cell, the probability that at least one of the pixels touching it is hot.

The 3×3 grid has 16 lattice points and 24 unit edges, and they come in kinds according to how many pixels touch them (Figure 12.53). Of the 16 points, the 4 corners of the grid touch one pixel, the 8 other points on the boundary touch two, and the 4 interior points touch four. Of the 24 edges, the 12 on the boundary touch one pixel and the 12 interior ones touch two. A cell touched by m pixels is present unless all m are cold, which happens with probability q^m. Then

$$\begin{aligned}
\mathbb{E}[\chi] &\overset{(1)}{=} \mathbb{E}[V] - \mathbb{E}[E] + \mathbb{E}[F] \\
&\overset{(2)}{=} \left[4(1-q) + 8(1-q^2) + 4(1-q^4)\right] - \left[12(1-q) + 12(1-q^2)\right] + 9p \\
&\overset{(3)}{=} \left[4p + 12 - 8q^2 - 4q^4\right] - \left[12p + 12 - 12q^2\right] + 9p \\
&\overset{(4)}{=} p + 4q^2 - 4q^4.
\end{aligned} \tag{12.83}$$

(1) Linearity of the expectation applied to $\chi = V - E + F$.
(2) Each kind of cell contributes (number of such cells)$\times$ P(not all of its pixels are cold); a pixel is hot with probability p, so $\mathbb{E}[F] = 9p$.
(3) $1 - q = p$ turns $4(1-q)$ into $4p$ and $12(1-q)$ into $12p$; the other brackets expand as $8(1-q^2) = 8 - 8q^2$, $4(1-q^4) = 4 - 4q^4$ and $12(1-q^2) = 12 - 12q^2$.
(4) The constants cancel, $12 - 12 = 0$; the p terms give $4p - 12p + 9p = p$; the q^2 terms give $-8q^2 + 12q^2 = 4q^2$; and $-4q^4$ is left.

Two quick checks. At $p = 1$ (all hot) it gives 1: one square, no hole. At $p = 0$ it gives $4 - 4 = 0$: nothing there.

The average number of holes, by hand. A hole needs cold pixels that cannot reach the outside through shared sides. Every pixel except the centre lies on the boundary of the grid and so touches the outside; the only candidate hole is the centre pixel. The centre is enclosed when it is cold and its four side-neighbours are hot; those four touch each other at their corners, so they form one piece that rings the centre. The corner pixels do not matter: a cold corner shares no side with the centre. Hence

$$\begin{aligned}
\mathbb{E}[\beta_1] &= \mathrm{P}(\text{centre cold})\ \mathrm{P}(\text{four side-neighbours hot}) = q\,p^4, \\
\mathbb{E}[\beta_0] &= \mathbb{E}[\chi] + \mathbb{E}[\beta_1] = p + 4q^2 - 4q^4 + p^4 q.
\end{aligned} \tag{12.84}$$

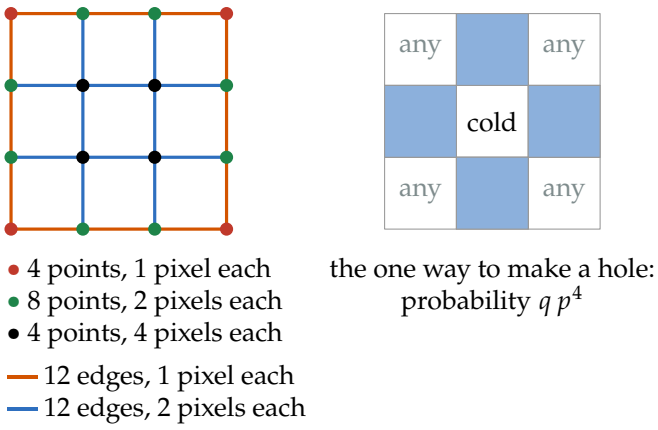


Figure 12.53: The bookkeeping behind (12.83) and (12.84). Left: the 16 lattice points and 24 edges of a 3×3 grid, sorted by how many pixels touch them; a cell is part of the hot set unless every pixel touching it is cold. Right: the centre is the only pixel that cannot reach the outside, so the only hole is a cold centre ringed by four hot side-neighbours; the corners can be anything.

Example: the numbers at $p = \frac{1}{2}$. With $q = \frac{1}{2}$, (12.83) gives $\mathbb{E}[\chi] = \frac{1}{2}+4\cdot\frac{1}{4}-4\cdot\frac{1}{16} = \frac{1}{2}+1-\frac{1}{4} = 1.25$, and (12.84) gives $\mathbb{E}[\beta_1] = \frac{1}{2}\cdot\frac{1}{16} = \frac{1}{32} = 0.03125$ and $\mathbb{E}[\beta_0] = 1.28125$. The enumeration below gives $\mathbb{E}[\chi] = 1.25000$, $\mathbb{E}[\beta_1] = 0.03125$ and $\mathbb{E}[\beta_0] = 1.28125$.

Example: when are holes most common? $\mathbb{E}[\beta_1] = p^4(1-p)$ is small at both ends: with few hot pixels there is no ring, with many there is no cold centre. Setting the derivative to zero,

$$\frac{\mathrm{d}}{\mathrm{d}p}\left[p^4(1-p)\right] \overset{(1)}{=} 4p^3 - 5p^4 \overset{(2)}{=} p^3(4-5p) = 0 \quad\Rightarrow\quad p = \tfrac{4}{5}, \qquad \mathbb{E}[\beta_1]_{\max} = \left(\tfrac{4}{5}\right)^4\tfrac{1}{5} = 0.082.$$

(1) Product rule: $\frac{\mathrm{d}}{\mathrm{d}p}p^4 = 4p^3$ and $\frac{\mathrm{d}}{\mathrm{d}p}(p^4\cdot p) = 5p^4$.

(2) Factor out p^3; the root $p = 0$ is the minimum.

The holes peak when four fifths of the pixels are hot, as in the large maps of Section 12.3, where holes dominate below the median threshold and pieces above it.

Example: why at most four pieces? Hot pixels that touch even at a corner are one piece, so separate pieces need cold pixels between them in every direction. The four corner pixels can be hot and separate if everything else is cold; a fifth hot pixel anywhere touches at least one of them or sits in the middle, which touches all four corners. So $\beta_0 \le 4$, and the enumeration confirms that the largest value is 4.

The whole distribution, and a Monte Carlo of it. The averages are not the whole story; we want $P(S \mid p)$ itself. With nine pixels there are $2^9 = 512$ configurations, and each has probability $p^k q^{9-k}$ if k of its pixels are hot. The script `code/ch12/c01_tiny.py` lists all of them, counts pieces and holes with the from-scratch routines of Section 12.2.2, adds up the probabilities, and checks every count against scikit-image:

Listing 12.16: All 512 configurations of a 3×3 grid, counted twice (`code/ch12/c01_tiny.py`)

```
configs = []
for bits in itertools.product([0, 1], repeat=9):
    m = np.array(bits, bool).reshape(3, 3)
    b0, b1 = betti_plane(m)
    chi = euler_cubical(m)
    assert chi == b0 - b1                                   # the Euler relation, checked
    512 times
    # scikit-image: 8-connected pieces, Euler number with 8-connectivity (connectivity=2)
    assert measure.label(m, connectivity=2).max() == b0
    assert measure.euler_number(np.pad(m, 1), connectivity=2) == chi
    configs.append((b0, b1, chi, m.sum()))
configs = np.array(configs)
w = p ** configs[:, 3] * q ** (9 - configs[:, 3])           # probability of each
    configuration
```

The loop builds each configuration as a 3×3 boolean array, counts it with `betti_plane` and `euler_cubical`, and asserts three things: the Euler relation $\chi = \beta_0 - \beta_1$, the number of pieces from scikit-image's `measure.label` with the 8-neighbour rule, and χ from its `measure.euler_number`. All 512×3 assertions pass. The last line stores, for each configuration, its counts and its number of hot pixels, from which its probability follows.

Then, as we would have to for a real map, the script forgets the enumeration and draws 10^5 random grids:

Listing 12.17: The same distribution by Monte Carlo (`code/ch12/c01_tiny.py`)

```
# ---- Monte Carlo: 100000 random grids
n = 100_000
grids = rng.random((n, 3, 3)) < p
mc = np.array([betti_plane(g) for g in grids])
mc_b0 = mc[:, 0].mean()
mc_b0_se = mc[:, 0].std() / np.sqrt(n)
mc_b1 = mc[:, 1].mean()
```

The Monte Carlo mean number of pieces is 1.2818 ± 0.0017 against the exact 1.28125, and the mean number of holes 0.0314 against 0.03125. Figure 12.54 shows the distribution of β_0 at $p = \frac{1}{2}$, exact and simulated side by side, and the three averages as functions of p.

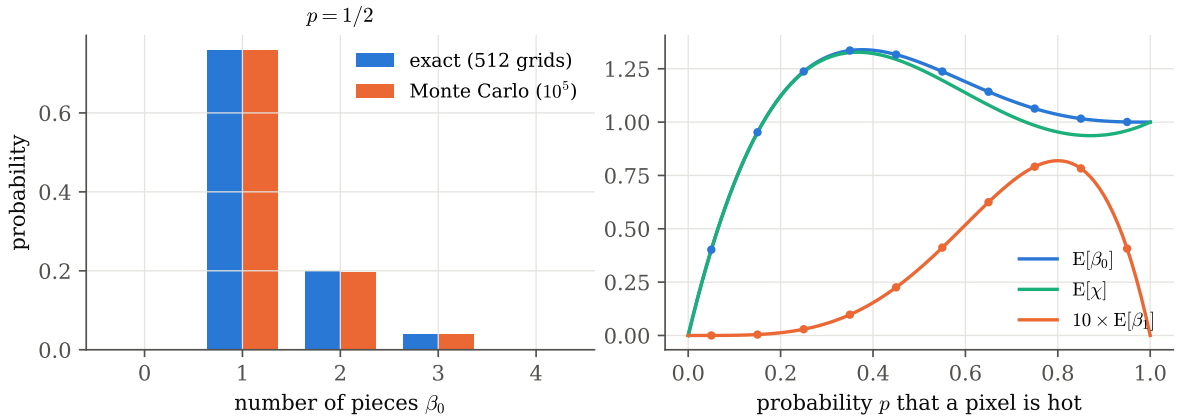


Figure 12.54: Left: the sampling distribution of the number of pieces on a 3×3 grid with $p = \frac{1}{2}$, from all 512 configurations (blue) and from 10^5 random grids (orange). One piece is by far the most common outcome; four, the largest possible, almost never happens. Right: the averages (12.83) and (12.84) (lines) against exact enumeration at ten values of p (dots); holes are multiplied by ten to be visible.

What the nine pixels already show. Three things carry over to every real map. First, a count is a random variable with a shape: at $p = \frac{1}{2}$, $\mathrm{P}(\beta_0 = 2) = 0.1992$ and the standard deviations are $\sigma(\beta_0) = 0.544$ and $\sigma(\chi) = 0.586$, comparable to the means. Second, the counts of one map are correlated with each other: $\mathrm{Cov}(\beta_0, \beta_1) = -0.0088$, negative because a hole needs the four side-pixels to be hot, and four hot side-pixels make one piece out of what could have been several. A likelihood for a vector of counts therefore needs a covariance matrix, not a list of variances. Third, enumeration stops working almost at once: a 256×256 map has 2^{65536} binary configurations, and its pixels are correlated. What remains is the second route, simulation, whose error falls as $1/\sqrt{n}$ with the number of maps whatever the size of the map (Section 6.11.2). That route, followed all the way to a posterior, is what the local model of non-Gaussianity needs next.

12.23 Two fields with one power spectrum

For the Gaussian fields of Section 7.2, specifying the mean and power spectrum fixes the distribution. Once Gaussianity is relaxed, fields with the same spectrum can differ substantially in shape, as the two maps in Section 7.7 illustrated. Primordial non-Gaussianity of the local type is the physical version of that situation: inflation with more than one field can add to the Gaussian potential a small piece of its own square ((T3.8); see the cosmology notes for the physics). The square adds a little power, but the dominant signature is in the shapes:

high peaks become a little higher and more numerous than deep troughs, or the other way round. To study that signature cleanly we want a family of fields, labelled by the amount of non-Gaussianity, whose power spectrum is *exactly* the same for every member. Then whatever separates them cannot have come from the spectrum.

The question. Can we build two random fields with the same power spectrum, one Gaussian and one not; and if so, do the Euler characteristic and the Betti curves of Sections 12.2 and 12.9.3 tell them apart on average?

12.23.1 The model, and why its spectrum does not depend on ϵ

Start from a Gaussian field g on an $N \times N$ periodic grid ($N = 256$), with zero mean, unit variance per pixel and power spectrum $S_g(k) \propto k^{-2}$: equal power in every octave of wavenumber, the flat-sky version of a scale-invariant spectrum. Its correlation function $\xi(\boldsymbol{r}) = \langle g(\boldsymbol{x}) g(\boldsymbol{x}+\boldsymbol{r})\rangle$ is the inverse Fourier transform of S_g, and $\xi(\boldsymbol{0}) = 1$. The non-Gaussian field adds a quadratic piece,

$$\phi = g + \epsilon\,(g^2 - 1), \tag{12.85}$$

which is the local model with $\epsilon = f_{\mathrm{NL}}\sigma_g$ in units where g has unit variance; subtracting $1 = \langle g^2\rangle$ keeps the mean zero. How much power does the quadratic piece add? The power spectrum is the Fourier transform of the correlation function (Wiener–Khinchin, (7.12)), so we need the correlation of ϕ at two points $\boldsymbol{x}$ and $\boldsymbol{y}$, with $g_x \equiv g(\boldsymbol{x})$ and $\xi \equiv \xi(\boldsymbol{x}-\boldsymbol{y})$:

$$\begin{aligned}
\langle \phi_x \phi_y\rangle &\overset{(1)}{=} \langle g_x g_y\rangle + \epsilon\big[\langle g_x (g_y^2-1)\rangle + \langle (g_x^2-1) g_y\rangle\big] + \epsilon^2 \langle (g_x^2-1)(g_y^2-1)\rangle \\
&\overset{(2)}{=} \xi + 0 + \epsilon^2\big[\langle g_x^2 g_y^2\rangle - \langle g_x^2\rangle - \langle g_y^2\rangle + 1\big] \\
&\overset{(3)}{=} \xi + \epsilon^2\big[(1\cdot 1 + 2\xi^2) - 1 - 1 + 1\big] \overset{(4)}{=} \xi + 2\epsilon^2\xi^2 .
\end{aligned} \tag{12.86}$$

(1) Multiply out $(g_x + \epsilon(g_x^2-1))(g_y + \epsilon(g_y^2-1))$.
(2) The middle terms are averages of an odd number of zero-mean Gaussian factors ($g_x g_y^2$ and g_x alone), which vanish by symmetry $g \to -g$.
(3) Wick's theorem (7.8) with the four factors g_x, g_x, g_y, g_y: the three pairings give $\langle g_x^2\rangle\langle g_y^2\rangle + 2\langle g_x g_y\rangle^2 = 1 + 2\xi^2$; and $\langle g^2\rangle = 1$.
(4) $1 + 2\xi^2 - 1 - 1 + 1 = 2\xi^2$.

Fourier transforming, the spectrum of ϕ is $S_\phi = S_g + \epsilon^2 S_2$, with S_2 the transform of $2\xi^2$: positive, of second order in ϵ, and with a shape of its own. We remove it with a fixed linear filter. Every Fourier mode of ϕ is multiplied by $R_\epsilon(k) = \sqrt{S_g(k)/S_\phi(k)}$, and then by a Gaussian smoothing $W(k) = e^{-k^2 s^2/2}$ of $s = 2$ pixels, which plays the role of a beam. A linear filter multiplies the power in each mode by its squared modulus (the convolution theorem, Section 7.6.1), so the observed map $u = W\,R_\epsilon\,\phi$ has the spectrum

$$S_u(k) = W^2(k)\,R_\epsilon^2(k)\,S_\phi(k) = W^2(k)\,S_g(k) \qquad \text{for every } \epsilon. \tag{12.87}$$

This is exact, because ϕ is a quadratic polynomial in g and (12.86) has no higher terms. The filter does one more thing we will need: it mixes neighbouring pixels, so u is *not* a pointwise function of a Gaussian map, and relabelling the thresholds (Section 12.3.2) will not undo ϵ completely.

Watch out: The same ensemble spectrum, not the same spectrum map by map

R_ϵ equalises the *average* power in each mode. Each individual map still has its own measured spectrum, scattered around the average by cosmic variance (Section 7.10.2), and the scatter itself depends weakly on ϵ through the four-point function. "The power spectrum cannot tell them apart" means: the estimated spectra of the two families have the same mean, so a likelihood built on the spectrum has no derivative with respect to ϵ.

Example: one map and its phase-randomised twin. A stronger version of the same statement holds for a single map. Take the Fourier transform of one $\epsilon = 0.2$ map, keep every modulus, and replace every phase by a random one (`randomise_phases` from Section 7.7). The two maps have the same $|\tilde{u}_k|$, and so the same periodogram, to machine precision: the largest relative difference of the moduli is 1.776×10^{-16}. Yet the skewness of the pixel values, $\langle u^3 \rangle / \sigma^3$, is 0.936 for the original and 0.038 for the twin. Random phases make every pixel a sum of many independent contributions, which pushes it towards Gaussian (the central limit theorem, Chapter 2); the non-Gaussianity of the original lived entirely in how its phases were aligned. Figure 12.55 shows the two maps and the spectra.

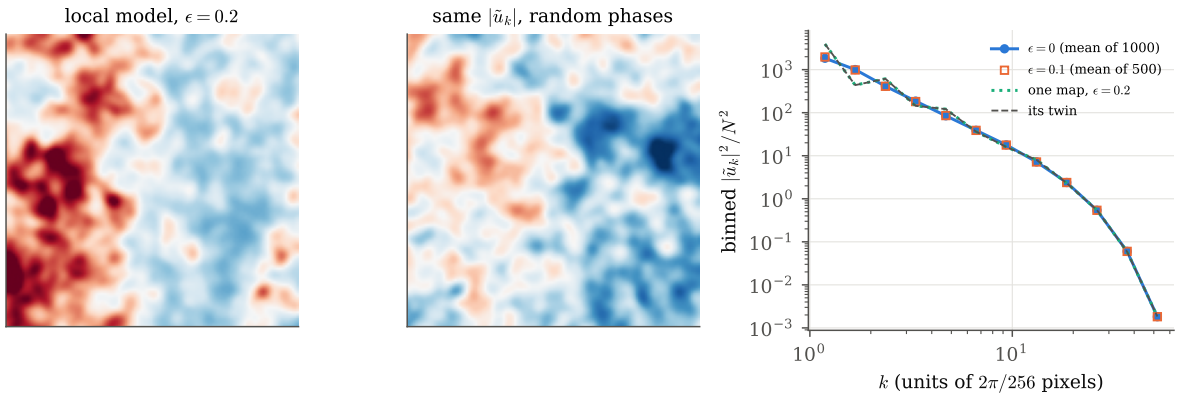


Figure 12.55: Left: a quarter of one map of the local model with $\epsilon = 0.2$ (red hot, blue cold), with sharp isolated hot spots and broad shallow cold regions. Middle: the same Fourier moduli with random phases; the hot and cold sides look alike. Right: binned spectra. The means over 1000 maps at $\epsilon = 0$ and 500 at $\epsilon = 0.1$ coincide, and the single map and its twin coincide exactly.

The ensembles. The Monte Carlo behind every result below, `code/ch12/c02_mc_fnl.py`, draws three sets of 256×256 maps: 1000 maps at $\epsilon = 0$ (the *null* set, for means and covariances), 500 more at $\epsilon = 0$ (the *check* set, kept apart for calibrating tests), and 500 seeds at each of $\epsilon \in \{0, \pm 0.025, \pm 0.05, \pm 0.1\}$ (the *grid* set). In the grid set every value of ϵ of one seed is built from the *same* Gaussian field g, which will matter for derivatives (Section 12.26). For each map it stores the binned periodogram, the skewness, the area fraction, χ, β_0 and β_1 at the thirteen thresholds $\nu = -3, -2.5, \ldots, 3$ (in units of the map's own standard deviation), the same three curves at area-fraction thresholds ν_A (12.10), and two-dimensional histograms of the persistence diagrams (Section 12.9) computed with gudhi, whose diagrams agree exactly with our union–find (Section 12.11.3). The ensemble spectra at $\epsilon = 0$ and $\epsilon = 0.1$ differ by at most 7.0% in any bin, against a Monte Carlo error of up to 3.2% on one mean. The difference of two independent means is noisier than either, and the worst of a dozen bins is expected near two standard errors: no sign of a difference.

12.23.2 What the topology sees

With the spectrum out of the way, the question is whether the average curves move. Because each grid seed has its own Gaussian g for all values of ϵ, the difference of a summary between $\epsilon = 0.05$ and $\epsilon = 0$ for the same seed contains almost no cosmic variance, and 500 seeds measure the mean shift precisely. Figure 12.56 shows the result for χ, β_0 and β_1.

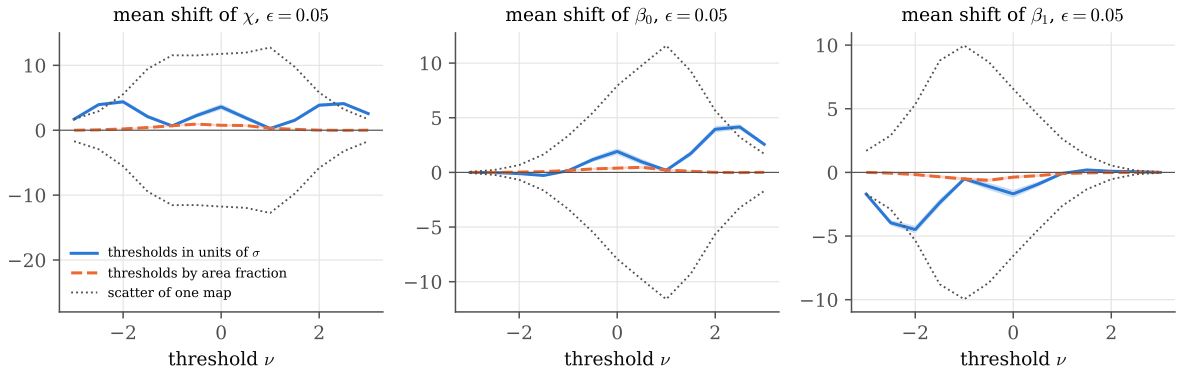


Figure 12.56: Average change of the Euler characteristic and the Betti curves when ϵ goes from 0 to 0.05, from 500 pairs of maps that share their Gaussian field (bands: ±2 standard errors). Solid: thresholds in units of the map's σ; dashed: thresholds by area fraction. Dotted: the scatter of a single map at $\epsilon = 0$. The extra hot peaks show up in β_0 at high thresholds, the shallower cold regions as fewer lakes in β_1 at low ones.

Three things are visible. The shift is real and has a shape: at $\nu = 1$ the mean Euler characteristic moves by 0.27 ± 0.08 (out of an average of 56.5), at $\nu = -1$ by 0.67, and in the hot tail at $\nu = 3$ by 2.57. It is small compared with the scatter of one map at most thresholds (12.7 at $\nu = 1$), so no single threshold will detect $\epsilon = 0.05$; but in the tails, where both the counts and their scatter are small, the shift is a sizeable fraction of the scatter, and a vector of thresholds adds these pieces of evidence together, which is the job of the covariance matrix (Section 12.24). Finally, at area-fraction thresholds the shift almost disappears. Most of what ϵ does to the counts is to move pixel values up or down, which is a relabelling of the threshold; what survives at fixed area fraction is the part that comes from the smoothing mixing neighbours.

12.23.3 Checking the shift against Matsubara's formula

For a weakly non-Gaussian field the first correction to the Euler characteristic is known ((T3.10), from [113]). In two dimensions, per unit area and with thresholds in units of σ_0,

$$\mathbb{E}[\chi(\nu)] = \frac{\lambda}{(2\pi)^{3/2}}\, e^{-\nu^2/2}\Big\{H_1(\nu) + \Big[\frac{S^{(0)}}{6}H_4(\nu) + \frac{2S^{(1)}}{3}H_2(\nu) + \frac{S^{(2)}}{3}\Big]\sigma_0\Big\}, \tag{12.88}$$

where $\lambda = \sigma_1^2/(2\sigma_0^2)$ is the gradient variance per component in units of the field variance (Section 12.5), H_n are the Hermite polynomials ($H_1 = \nu$, $H_2 = \nu^2 - 1$, $H_4 = \nu^4 - 6\nu^2 + 3$), and $S^{(0)}\sigma_0$, $S^{(1)}\sigma_0$, $S^{(2)}\sigma_0$ are the three skewness parameters (T3.9): the averages of u^3, of $u^2\nabla^2 u$ and of $|\nabla u|^2\nabla^2 u$, suitably normalised. At $\epsilon = 0$ the bracket vanishes and we are back to (12.23).

The derivative with respect to ϵ. The three skewness parameters vanish at $\epsilon = 0$ and grow in proportion to ϵ; write their slopes as $S'_a \equiv \mathrm{d}(S^{(a)}\sigma_0)/\mathrm{d}\epsilon$. For a map of area A, differentiating (12.88) at $\epsilon = 0$ gives

$$\frac{\mathrm{d}\,\mathbb{E}[\chi(\nu)]}{\mathrm{d}\epsilon} = \frac{A\lambda}{(2\pi)^{3/2}}\, e^{-\nu^2/2}\Big[\frac{S'_0}{6}H_4(\nu) + \frac{2S'_1}{3}H_2(\nu) + \frac{S'_2}{3}\Big]. \tag{12.89}$$

For a pointwise distortion $u + \epsilon(u^2 - 1)$ all three slopes would be 6 (Section T3.2.2). The Monte Carlo script `c02_mc_fnl.py` measures the three parameters on every grid map, and a straight-line fit in ϵ gives $S_0' = 4.29$, $S_1' = 4.48$, $S_2' = 5.19$: smaller than 6, because smoothing after squaring averages the square over neighbouring pixels, and unequal, because that averaging treats the field and its derivatives differently.

The same derivative at fixed area fraction. At area-fraction thresholds the term with $S^{(0)}$, the one-point skewness, should drop out. Why, and what replaces it? The area fraction above ν is, to first order ((T3.10), first line), $\Psi(\nu) + \varphi(\nu)\frac{S^{(0)}\sigma_0}{6}H_2(\nu)$, with Ψ the Gaussian tail probability (12.10) and $\varphi(\nu) = e^{-\nu^2/2}/\sqrt{2\pi}$ the Gaussian density. The threshold ν that encloses the area fraction $\Psi(\nu_A)$ therefore satisfies

$$\Psi(\nu_A) \overset{(1)}{=} \Psi(\nu) + \varphi(\nu)\frac{S^{(0)}\sigma_0}{6}H_2(\nu) \overset{(2)}{=} \Psi(\nu_A) - \varphi(\nu_A)(\nu - \nu_A) + \varphi(\nu_A)\frac{S^{(0)}\sigma_0}{6}H_2(\nu_A)$$
$$\Rightarrow \quad \nu = \nu_A + \frac{S^{(0)}\sigma_0}{6}H_2(\nu_A). \tag{12.90}$$

(1) The definition of ν_A: the measured area fraction at ν equals $\Psi(\nu_A)$.
(2) Taylor-expand $\Psi(\nu)$ around ν_A with $\Psi' = -\varphi$; in the correction term, which is already first order, replacing ν by ν_A changes only second-order terms.

Now insert $\nu = \nu_A + \delta$ with $\delta = \frac{S^{(0)}\sigma_0}{6}H_2(\nu_A)$ into the Gaussian part of (12.88), $e^{-\nu^2/2}H_1(\nu)$, and keep first order:

$$e^{-\nu^2/2}\nu \overset{(1)}{=} e^{-\nu_A^2/2}\nu_A + \delta\,(1-\nu_A^2)\,e^{-\nu_A^2/2} \overset{(2)}{=} e^{-\nu_A^2/2}\left[H_1 - \frac{S^{(0)}\sigma_0}{6}H_2^2\right]$$
$$\overset{(3)}{=} e^{-\nu_A^2/2}\left[H_1 - \frac{S^{(0)}\sigma_0}{6}\left(H_4 + 4H_2 + 2\right)\right].$$

(1) $\frac{\mathrm{d}}{\mathrm{d}\nu}(\nu e^{-\nu^2/2}) = (1-\nu^2)e^{-\nu^2/2}$.
(2) $1 - \nu_A^2 = -H_2(\nu_A)$, and δ carries another H_2.
(3) $H_2^2 = (\nu^2-1)^2 = \nu^4 - 2\nu^2 + 1 = H_4 + 4H_2 + 2$, all at ν_A.

The H_4 terms cancel against the $S^{(0)}H_4/6$ of (12.88), and the remaining $S^{(0)}$ terms combine with the others:

$$\frac{\mathrm{d}\ \mathbb{E}[\chi(\nu_A)]}{\mathrm{d}\epsilon} = \frac{A\lambda}{(2\pi)^{3/2}}\,e^{-\nu_A^2/2}\left[\frac{2(S_1' - S_0')}{3}H_2(\nu_A) + \frac{S_2' - S_0'}{3}\right], \tag{12.91}$$

Matsubara's eq. (3.54) in the variable ν_A [113]. Only the differences between the slopes of the three parameters survive: for the one-point skewness they carry no new information. For a pointwise distortion they vanish; for our smoothed field they are $4.48 - 4.29$ and $5.19 - 4.29$, small but not zero.

The comparison. The formula is per unit area of a plane without edges, so the script `code/ch12/c03_same_spectrum.py` uses periodic maps (the counting of Section 12.2.3 with wrap-around) and 400 pairs at $\epsilon = \pm 0.05$ that share g, and measures $\lambda = 0.0204$ per pixel^2 (a correlation length $1/\sqrt{2\lambda} = 4.9$ pixels) on the maps themselves:

Listing 12.18: Matsubara's first-order prediction, at σ and at area-fraction thresholds (`code/ch12/c03_same_spectrum.py`)

```
1
2 sk = z["grid_sk"]                                   # (seed, eps, [S0, S1, S2, lambda])
3 epsv = np.array(eg)
```

```
slope = np.polyfit(epsv, sk[:, :, :3].mean(0), 1)[0]   # dS^(a) sigma0 / d eps for a = 0, 1, 2
lam = sk[:, i0, 3].mean()
nuf = np.linspace(-3.2, 3.2, 200)
H = {1: nuf, 2: nuf ** 2 - 1, 4: nuf ** 4 - 6 * nuf ** 2 + 3}
amp = N * N * lam / (2 * np.pi) ** 1.5 * np.exp(-nuf ** 2 / 2)
pred = amp * (slope[0] / 6 * H[4] + 2 * slope[1] / 3 * H[2] + slope[2] / 3)
predA = amp * (2 * (slope[1] - slope[0]) / 3 * H[2] + (slope[2] - slope[0]) / 3)
```

The first lines fit straight lines in ϵ to the skewness parameters measured on all grid maps (the slopes S_a') and take λ from the $\epsilon = 0$ maps; `amp` is $A\lambda e^{-\nu^2/2}/(2\pi)^{3/2}$ with $A = N^2$ pixels; `pred` and `predA` are (12.89) and (12.91). Figure 12.57 compares them with the Monte Carlo. At $\nu = -2.0$, where the predicted derivative is largest, the simulation gives 83 ± 3 and the formula 82. At area-fraction thresholds the derivative is at most 19 (formula: 16), about 0.19 of the σ-threshold amplitude.

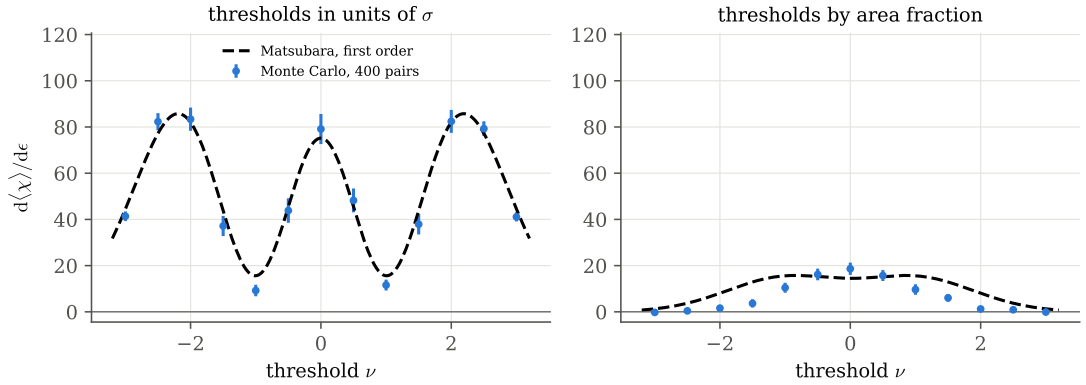


Figure 12.57: Derivative of the mean Euler characteristic of a periodic 256×256 map with respect to ϵ. Points: 400 pairs of maps at $\epsilon = \pm 0.05$ sharing their Gaussian field (± 2 standard errors). Dashed: Matsubara's first-order formula with the skewness parameters measured on the maps. Left, thresholds in units of σ: the H_4-dominated shape of a skewed histogram. Right, area-fraction thresholds: what is left is a small bump from the non-local smoothing.

The agreement says two things. The machinery is right: our counting, our maps and a perturbative formula derived by other means meet within the Monte Carlo errors, apart from the thresholds $\nu = \pm 1$ where both the Gaussian $\mathbb{E}\,\chi$ and its derivative change fastest across a pixel and the lattice effects of Section 12.2.2 are largest. And the physics is clear: most of the topological response to ϵ is a relabelling of the pixel values, and a smaller part, about a fifth here, is in how the values are arranged. Both are invisible to the power spectrum.

Key idea: Same spectrum, different shapes

For the renormalised local model the power spectrum is the same for every ϵ (12.87), while the mean Euler characteristic and Betti curves move in proportion to ϵ, with the shape predicted by Matsubara's formula. Any statistical power to detect ϵ in this family comes from beyond the two-point function.

12.24 The sampling distribution by simulation

Simulated covariances are not new to us. In Sections 8.13 and 8.13.2 we estimated the covariance of masked CMB band powers from simulated skies and derived why its inverse must be multiplied by $(n - p - 2)/(n - 1)$; in Section 12.7.4 the same recipe gave the 36-number χ^2

of *Planck*'s Minkowski-functional test on Gaussian skies; and Section 12.20.2 used it for peak counts and Betti numbers of lensing maps. What is the same here is the recipe: simulate the null model many times, average, form the covariance, invert, correct. What is new is the use we make of it. In Section 8.13.2 the question was whether the *average* of the inverse is right. To test a hypothesis we need the whole *distribution* of the χ^2 of one map, including its tail, and our data vector is made of counts, some of them small integers, which are not Gaussian at all.

The question. If the field is Gaussian, what values does the χ^2 of one map's Betti curves take, when the mean and the covariance it is measured against were themselves estimated from a finite number of simulated maps?

12.24.1 The data vector and its covariance

For each map the data vector stacks the two Betti curves of Section 12.23.2 at the thirteen thresholds, $\boldsymbol{S} = (\beta_0(\nu_1), \dots, \beta_0(\nu_{13}), \beta_1(\nu_1), \dots, \beta_1(\nu_{13}))$. One entry, the number of lakes at $\nu = 3$, is zero in all but one of the simulated maps. An entry that never varied would carry no information and would make the covariance singular, so such entries are dropped; this one varies, if only just, and stays, leaving $p = 26$ numbers. From the $n = 1000$ null maps ($\epsilon = 0$) we form the sample mean and the sample covariance,

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{S}_i, \qquad \hat{\mathbf{C}} = \frac{1}{n-1}\sum_{i=1}^{n} (\boldsymbol{S}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{S}_i - \hat{\boldsymbol{\mu}})^{\mathsf{T}}, \tag{12.92}$$

exactly as in Section 3.5.4. Figure 12.58 shows the mean curves with the scatter of one map, and the correlation matrix $\hat{C}_{ij}/\sqrt{\hat{C}_{ii}\hat{C}_{jj}}$.

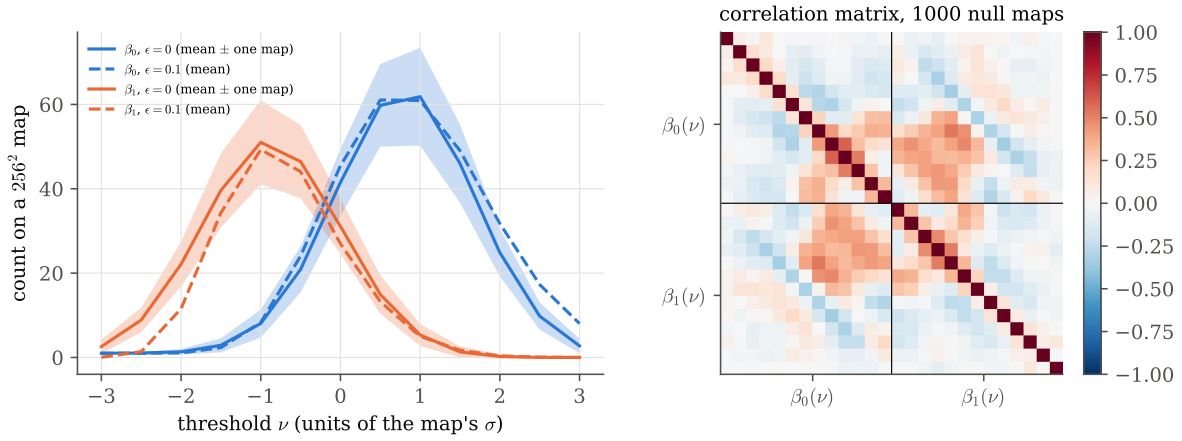


Figure 12.58: Left: mean Betti curves of 1000 Gaussian 256×256 maps (solid) with the $\pm 1\sigma$ scatter of one map (bands), and the mean at $\epsilon = 0.1$ (dashed). Pieces (β_0) dominate above the median, lakes (β_1) below. Right: the 26×26 correlation matrix of the counts. Neighbouring thresholds are positively correlated, and a broad positive block links the hot pieces above the median to the lakes below it.

Reading the correlation matrix. Two structures stand out. Neighbouring thresholds are correlated, for example 0.59 between $\beta_0(1)$ and $\beta_0(1.5)$, because they count largely the same islands: an island that is above 1.5 is also above 1. And there is a positive correlation of 0.40 between the hot pieces at $\nu = 1$ and the cold lakes at $\nu = -1$, which never share a pixel. Why

should they rise and fall together? Because both are counts of features, and the number of features per unit area of a Gaussian field is set by one number, $\lambda = \sigma_1^2/2\sigma_0^2$ (Section 12.5). Thresholds are in units of each map's own standard deviation, and with equal power per octave a good part of that standard deviation comes from the few longest waves in the box, which differ from map to map. A map whose long waves happen to be weak has a smaller σ_0, the same small-scale roughness σ_1, a larger λ, and therefore more islands *and* more lakes at every threshold. Whatever the reason in a given case, the lesson of the nine pixels returns: the counts of one map are not independent, and a likelihood must use the full matrix.

How many simulations? Each entry of $\hat{\mathbf{C}}$ is an average over n maps, so its fractional error is about $\sqrt{2/n}$, 4.5% for $n = 1000$; and the inverse is biased high by $(n-1)/(n-p-2)$, a factor $1/0.973$ here ((8.78), [6]). Hobson et al. make the same point in the language of the inverse Wishart distribution: the covariance estimated from n_s simulations is usable only if n_s is well above twice the length of the data vector [HobsonBMC §3.4.3]. With $p = 26$ and $n = 1000$ we are far into the safe regime; with a 98-bin persistence histogram and 150 maps we will not be.

12.24.2 The χ^2 of a new map when the covariance is estimated

Take one more map, independent of the n simulations, and compute

$$\chi^2 = (\boldsymbol{S} - \hat{\boldsymbol{\mu}})^{\mathsf{T}}\, \hat{\mathbf{C}}^{-1}\, (\boldsymbol{S} - \hat{\boldsymbol{\mu}}). \tag{12.93}$$

If $\hat{\boldsymbol{\mu}}$ and $\hat{\mathbf{C}}$ were the true $\boldsymbol{\mu}$ and $\mathbf{C}$, and $\boldsymbol{S}$ were Gaussian, this would be a sum of p squared independent standard normal variables after whitening, a χ^2_p variable with mean p (Section 4.8.1). Both estimates add noise. How much?

Its average. Suppose $\boldsymbol{S}$ and the simulations are Gaussian with mean $\boldsymbol{\mu}$ and covariance $\mathbf{C}$. Three facts are needed: $\boldsymbol{S}$ is independent of the simulations; for Gaussian samples $\hat{\boldsymbol{\mu}}$ and $\hat{\mathbf{C}}$ are independent of each other (Section 8.13.2); and $\hat{\boldsymbol{\mu}}$ has covariance $\mathbf{C}/n$. Then

$$\begin{aligned}\mathbb{E}[\chi^2] &\overset{(1)}{=} \mathbb{E}\Big[\mathrm{Tr}\big(\hat{\mathbf{C}}^{-1}(\boldsymbol{S}-\hat{\boldsymbol{\mu}})(\boldsymbol{S}-\hat{\boldsymbol{\mu}})^{\mathsf{T}}\big)\Big] \overset{(2)}{=} \mathrm{Tr}\Big(\mathbb{E}\big[\hat{\mathbf{C}}^{-1}\big]\ \mathbb{E}\big[(\boldsymbol{S}-\hat{\boldsymbol{\mu}})(\boldsymbol{S}-\hat{\boldsymbol{\mu}})^{\mathsf{T}}\big]\Big) \\ &\overset{(3)}{=} \mathrm{Tr}\Big(\frac{n-1}{n-p-2}\,\mathbf{C}^{-1}\,\Big(1+\frac{1}{n}\Big)\mathbf{C}\Big) \overset{(4)}{=} p\,\Big(1+\frac{1}{n}\Big)\frac{n-1}{n-p-2}.\end{aligned} \tag{12.94}$$

(1) A number is its own trace, and $\mathrm{Tr}(\boldsymbol{a}^{\mathsf{T}}\mathsf{M}\boldsymbol{a}) = \mathrm{Tr}(\mathsf{M}\boldsymbol{a}\boldsymbol{a}^{\mathsf{T}})$.
(2) $\hat{\mathbf{C}}$ is independent of both $\boldsymbol{S}$ and $\hat{\boldsymbol{\mu}}$, so the average of the product is the product of the averages; the trace is linear.
(3) The Hartlap result (8.78); and $\boldsymbol{S} - \hat{\boldsymbol{\mu}}$ has mean zero and covariance $\mathbf{C} + \mathbf{C}/n$, because the two terms are independent and their covariances add.
(4) $\mathrm{Tr}(\mathbf{C}^{-1}\mathbf{C}) = \mathrm{Tr}\,\mathbb{1} = p$.

The factor $1 + 1/n$ comes from the noise in the mean, the factor $(n-1)/(n-p-2)$ from the noise in the covariance.

Its whole distribution. The same ingredients give the shape. Whiten the problem: in coordinates where $\mathbf{C} = \mathbb{1}$, the vector $\boldsymbol{w} = (\boldsymbol{S} - \hat{\boldsymbol{\mu}})/\sqrt{1+1/n}$ is a standard normal vector, so $|\boldsymbol{w}|^2 \sim \chi^2_p$. Now look at $\hat{\mathbf{C}}^{-1}$ along the direction of $\boldsymbol{w}$. Section 8.13.2 showed, through the Schur complement (8.77), that the diagonal element of $(k\hat{\mathbf{C}})^{-1}$ along any fixed direction is one over a χ^2_{k-p+1} variable when $\mathbf{C} = \mathbb{1}$, with $k = n-1$. In whitened coordinates the simulations look the

same in every direction, and they are independent of $\boldsymbol{w}$, so we may take the direction of $\boldsymbol{w}$ as the fixed one. Therefore

$$\chi^2 \overset{(1)}{=} \left(1+\frac{1}{n}\right)|\boldsymbol{w}|^2\,\hat{\boldsymbol{e}}^{\mathsf{T}}\hat{\mathbf{C}}^{-1}\hat{\boldsymbol{e}} \overset{(2)}{=} \left(1+\frac{1}{n}\right)(n-1)\,\frac{\chi^2_p}{\chi^2_{n-p}} \overset{(3)}{=} \left(1+\frac{1}{n}\right)\frac{(n-1)\,p}{n-p}\,F_{p,n-p}, \tag{12.95}$$

(1) Write $\boldsymbol{S}-\hat{\boldsymbol{\mu}} = \sqrt{1+1/n}\,|\boldsymbol{w}|\,\hat{\boldsymbol{e}}$ with $\hat{\boldsymbol{e}}$ the unit vector along $\boldsymbol{w}$.
(2) $|\boldsymbol{w}|^2 \sim \chi^2_p$; $\hat{\mathbf{C}} = A/(n-1)$ with A the scatter matrix of Section 8.13.2, and $\hat{\boldsymbol{e}}^{\mathsf{T}}A^{-1}\hat{\boldsymbol{e}} = 1/\chi^2_{n-p}$ (the degrees of freedom $k-(p-1) = n-p$), independent of the numerator.
(3) The ratio $(\chi^2_p/p)/(\chi^2_{n-p}/(n-p))$ of two independent χ^2 variables, each divided by its degrees of freedom, is Fisher's F distribution with $(p, n-p)$ degrees of freedom.

This is Hotelling's T^2 distribution, and its mean, $(n-p)/(n-p-2)$ for the F factor, reproduces (12.94). For $n \to \infty$ the denominator $\chi^2_{n-p}/(n-p)$ tends to one and χ^2_p returns.

Key idea: A χ^2 with a simulated covariance is wider than χ^2_p

If the mean and covariance of a Gaussian data vector of length p are estimated from n simulations, the χ^2 of a new independent vector is $(1+\frac{1}{n})\frac{(n-1)p}{n-p}F_{p,n-p}$, with mean $p(1+\frac{1}{n})\frac{n-1}{n-p-2}$. The Hartlap factor fixes the mean; it does not fix the width, which comes from the χ^2_{n-p} in the denominator.

How much wider? The spread of a χ^2_k variable relative to its mean is $\sqrt{2k}/k = \sqrt{2/k}$, so the logarithm of χ^2_k/k has a variance of about $2/k$ (the delta method: $\operatorname{Var}\ln X \approx \operatorname{Var} X/(\mathbb{E}\,X)^2$). The logarithm of the ratio in (12.95) is a difference of two independent such logarithms, so their variances add:

$$\operatorname{Var}\ln\chi^2 \approx \frac{2}{p}+\frac{2}{n-p} \qquad \text{instead of} \qquad \frac{2}{p}. \tag{12.96}$$

For the Betti vector, $p = 26$ and $n = 1000$, the second term is $2/974$, negligible next to $2/26$. For a persistence histogram with $p = 94$ bins and only $n = 150$ simulations it is $2/56 = 0.036$ against $2/94 = 0.021$: the χ^2 is spread $\sqrt{0.057/0.021} \approx 1.6$ times wider, in the logarithm, than the χ^2_p table assumes, and no constant factor can repair a wrong width.

The check. The script `code/ch12/c04_test.py` computes (12.93) for the 500 maps of the check set against the mean and covariance of the 1000 null maps:

Listing 12.19: Mean, Hartlap-corrected inverse covariance, and the χ^2 of new maps (`code/ch12/c04_test.py`)

```
def chi2(X, mu, Cinv):
    D = X - mu
    return np.einsum("...i,ij,...j->...", D, Cinv, D)


def fit(N, hartlap=True, rare=0.0):
    """Mean, inverse covariance (Hartlap-corrected) and kept columns of a null sample N (n, p);
    rare > 0 also drops bins in which fewer than that fraction of the maps differ from the median."""
    ok = (N.std(0) > 1e-9) & ((N != np.median(N, 0)).mean(0) >= rare)   # drop bins that (almost) never change
    N = N[:, ok]
    n, p = N.shape
```

```
    C = np.atleast_2d(np.cov(N.T))
    Ci = np.linalg.inv(C)
    if hartlap:
        Ci = Ci * (n - p - 2) / (n - 1)
    return N.mean(0), Ci, ok, n, p
```

`fit` drops the entries that never change (and, if asked through `rare`, those that almost never change; we use that in Section 12.25), estimates the mean and the covariance, inverts, and applies the Hartlap factor if asked; `chi2` evaluates the quadratic form for a whole stack of maps at once with `einsum`. For the Betti vector ($p = 26$, $n = 1000$) the average χ^2 of the independent maps is 30.4, against 26.7 from (12.94); a single map with a rare feature can move this average by a few units, as the next paragraph shows. For the persistence histograms ($p = 98$, $n = 1000$) it is 113.1 against 108.9. With only $n = 150$ maps for the covariance ($p = 94$) it is 273 against 261; the Hartlap factor would be 0.36, and the χ^2_p table, centred at 94, is not even in the right place. Figure 12.59 shows the three distributions.

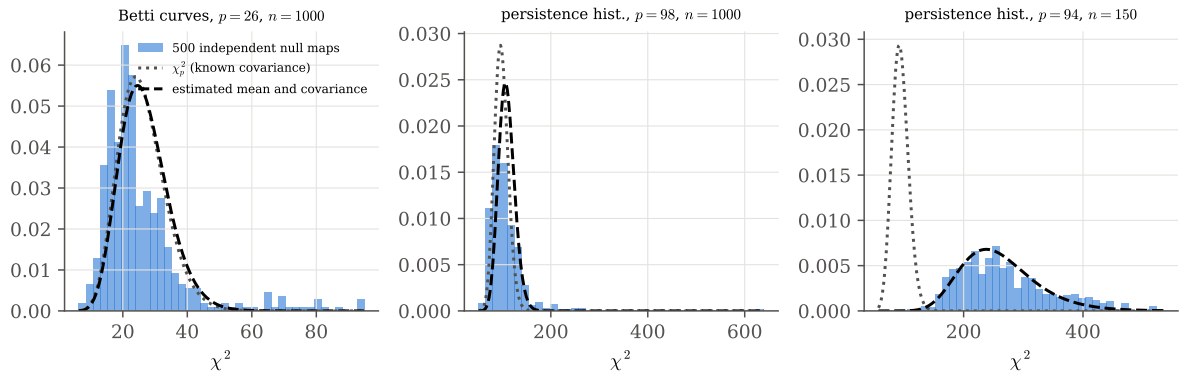


Figure 12.59: The χ^2 (12.93) of 500 Gaussian maps that were not used to estimate the mean and covariance (histograms). Dotted: χ^2_p, correct only if the covariance were known. Dashed: the Hotelling law (12.95). Left, Betti curves with 1000 simulations: the two curves nearly coincide. Middle, a 98-bin persistence histogram with 1000 simulations: shifted and widened. Right, the same with 150 simulations: the χ^2_p curve is far off, the Hotelling law still follows the histogram.

Where the Gaussian assumption fails. The derivation assumed Gaussian counts, and the bulk of the histograms agrees with it. The tail does not. Of the 500 Betti χ^2 values, 6.2% lie above the 99.9th percentile of χ^2_{26}, where 0.1% were expected, and the largest is 1099. That one map is not exotic. Its largest deviation is in $\beta_1(3)$, a count whose simulated mean is 0.00 with a scatter of only 0.03; the map had 1, a deviation of 32 standard deviations in one entry, and the square of that dominates the χ^2. Counts of rare features are small integers, closer to Poisson than to Gaussian, and a bin that is almost always the same number has a tiny variance that turns an ordinary jump of one into an enormous χ^2.

Watch out: Rare-feature bins

A bin that contains on average one or two features, or that almost never changes, makes the Gaussian χ^2 heavy-tailed and the covariance nearly singular. Either drop or merge such bins (Cole and Shiu discard bins whose fluctuations are not small compared with their mean [271]), or keep them and take the distribution of the statistic from simulations rather than from a table (Section 12.25.1).

12.25 Is this map Gaussian? A test with a simulated null

A colleague hands us one 256×256 map. Its power spectrum is exactly what a Gaussian field with spectrum $W^2 S_g$ would have, so the spectrum says nothing. Was it drawn from the Gaussian model, $\epsilon = 0$, or from one of its non-Gaussian twins of Section 12.23.1? The Gaussian model has a special status: it is the hypothesis whose sampling distribution we *know*, because we can simulate it as often as we like. That is what makes it the null hypothesis (Section 4.1.1): we do not try to prove the map is non-Gaussian; we ask whether "it is just a Gaussian map, and what we see happened by chance" is a reasonable account of it.

The question. If the map were Gaussian, would a χ^2 of its Betti curves as large as the one we measured, or larger, be something we would reasonably expect? And how often would the test catch a map that really has $\epsilon = 0.05$?

12.25.1 A p-value from simulations

Section 12.24.2 showed that the χ^2 of a Gaussian map is not distributed as χ^2_p, and that its tail is heavier still because counts are not Gaussian. Instead of a formula we use simulations a second time. The check set of Section 12.23.1, $N = 500$ Gaussian maps that played no part in estimating $\hat{\boldsymbol{\mu}}$ and $\hat{\mathbf{C}}$, gives N values $\chi^2_1, \dots, \chi^2_N$ of the statistic under the null hypothesis, computed exactly as for the observed map. With r the number of them at least as large as the observed χ^2_{obs}, define

$$p = \frac{1 + r}{1 + N}. \tag{12.97}$$

Why this is a valid p-value. If the null hypothesis is true, the observed map is just one more Gaussian map: its χ^2 and the N reference values are $N + 1$ independent draws from the same continuous distribution. Then no draw is special, and the observed one is equally likely to be the largest, the second largest, ..., the smallest. Write R for its rank counted from the top (the largest has $R = 1$); $R = 1 + r$. For any $k = 1, \dots, N + 1$,

$$\mathrm{P}\Big(p \le \frac{k}{N+1}\Big) \overset{(1)}{=} \mathrm{P}(1 + r \le k) \overset{(2)}{=} \mathrm{P}(R \le k) \overset{(3)}{=} \frac{k}{N+1}. \tag{12.98}$$

(1) Multiply both sides of $\frac{1+r}{N+1} \le \frac{k}{N+1}$ by $N + 1$.
(2) $1 + r$ is the rank from the top.
(3) Exchangeability: each of the $N + 1$ ranks has probability $1/(N + 1)$, and k of them are $\le k$.

So $\mathrm{P}(p \le \alpha) = \alpha$ whenever α is a multiple of $1/(N + 1)$, and $\mathrm{P}(p \le \alpha) \le \alpha$ otherwise: a p-value in the exact sense of (4.15), with no assumption about the shape of the distribution. The +1 in numerator and denominator is not a fudge: it counts the observed map among the draws, and it keeps p from ever being zero. The smallest attainable value is $1/(N + 1) = 1/500$ here, so with 500 reference maps no map can be called more extreme than about 1 in 500.

Working definition 12.8 (simulated p-value)

For a statistic T whose distribution under H_0 can be simulated, draw N independent realisations of H_0 through the same pipeline as the data, compute $T_1, \dots, T_N$, and report $p = (1 + \#\{T_i \ge T_{\text{obs}}\})/(1 + N)$ [the fraction of the $N + 1$ values, the observed one included, that are at least as extreme]. It is exact for any N. Its resolution is $1/(N + 1)$.

Why the reference maps must be new. It is tempting to reuse the 1000 null maps for the reference distribution. Their χ^2 values against their own mean and covariance are systematically too small:

$$\sum_{i=1}^{n}(\boldsymbol{s}_i-\hat{\boldsymbol{\mu}})^{\mathsf{T}}\hat{\mathbf{C}}^{-1}(\boldsymbol{s}_i-\hat{\boldsymbol{\mu}}) \overset{(1)}{=} \operatorname{Tr}\Big(\hat{\mathbf{C}}^{-1}\sum_{i=1}^{n}(\boldsymbol{s}_i-\hat{\boldsymbol{\mu}})(\boldsymbol{s}_i-\hat{\boldsymbol{\mu}})^{\mathsf{T}}\Big) \overset{(2)}{=} \operatorname{Tr}\big(\hat{\mathbf{C}}^{-1}(n-1)\hat{\mathbf{C}}\big) \overset{(3)}{=} (n-1)\,p,$$

(1) The trace trick of (12.94), and the trace is linear.
(2) The definition (12.92) of $\hat{\mathbf{C}}$.
(3) $\operatorname{Tr}\mathbb{1} = p$.

so their average is $p(n-1)/n$, exactly and for every set of simulations, while a new map has the larger average (12.94). The covariance was fitted to these very maps and "explains" them too well. A new map compared with the in-sample distribution looks too extreme, and the test raises false alarms. The check set exists to avoid this; with a single set of simulations one leaves each map out in turn, as the sphere analysis of Section 12.28 does.

12.25.2 Is the test honest?

Before using a test on data we run it on maps for which we know the answer. The 500 grid seeds at $\epsilon = 0$ are Gaussian maps independent of both the null and the check set. Their p-values should be uniform between 0 and 1 (Section 4.4.2).

First, a choice the pitfall of Section 12.24.2 asked for. A bin in which almost every map has the same count, such as the holes of the excursion set at $\nu = 3$ or its pieces at $\nu = -3$, has a tiny variance, and one map with a single rare feature there gets an enormous χ^2. Whether such a bin enters the vector at all then depends on whether one of the 1000 null maps happened to have the feature. For the tests we drop every bin in which fewer than 5% of the null maps differ from the median count. The Betti vector of the tests has $p = 22$ entries.

Now the check. 5.0% of the Gaussian maps get $p < 0.05$, where 5% were expected. How far off may that be? The guarantee (12.98) is for one map and a reference set drawn afresh. Our 500 maps all share the same 500 reference values, so the noise of that one set moves every p-value together. Two uncertainties add: the binomial one of 500 maps, $\sqrt{0.05 \times 0.95/500} \approx 1\%$, and the uncertainty of the 95th percentile of 500 reference values, about the same size. Together they give 1.4%, so 5.0% is 0.0 standard deviations from 5%. For the same reason a Kolmogorov–Smirnov test of the p-values against the uniform distribution would be wrong: it assumes independent draws. The honest question is whether the 500 values of χ^2 and the 500 reference values come from one distribution, and the two-sample Kolmogorov–Smirnov test (Section 4.8.4) answers it with $p = 0.11$.

Example: what goes wrong with a table. Repeat the test with the persistence histograms and only $n = 150$ simulations for the covariance, and compute the p-values of the same Gaussian maps in three ways. The χ^2_p table with the raw $\hat{\mathbf{C}}^{-1}$ rejects 100% of these Gaussian maps at the 5% level: the whole distribution sits far to the right of the table (Figure 12.59, right). Multiplying by the Hartlap factor puts the mean in the right place, and the false alarm rate drops to 20%, still about 4 times too many: the distribution is too wide (12.96), and maps pile up at both ends of the p-value histogram. The simulated p-value (12.97) gives 6%. The left panel of Figure 12.60 shows the three histograms. When the mean and covariance are estimated, the honest options are a likelihood that knows it (the t-distribution of Sellentin and Heavens, which integrates over the unknown true covariance [272]) or a reference distribution from independent simulations.

12.25.3 How often does it catch $\epsilon \neq 0$? Omnibus and aimed tests

A test that never raises false alarms is useless if it never raises true ones. The power of a test is the probability that it rejects H_0 when a specific alternative is true (Section 4.3). We measure it directly: compute the simulated p-value for 250 grid maps at each ϵ and count the fraction below 0.05.

The χ^2 test is an **omnibus** test: it asks whether the map is far from the Gaussian mean in *any* of the p directions of the data vector. Local non-Gaussianity moves the mean in one particular direction, the derivative $\boldsymbol{d} = \partial\boldsymbol{\mu}/\partial\epsilon$ that Figure 12.56 showed. A test aimed at that direction should do better.

The aimed test is the score. Model the summary as Gaussian with mean $\boldsymbol{\mu}(\epsilon)$ and a fixed covariance. The log-likelihood is $\ln \mathcal{L}(\epsilon) = -\frac{1}{2}(\boldsymbol{S} - \boldsymbol{\mu}(\epsilon))^{\mathsf{T}}\mathbf{C}^{-1}(\boldsymbol{S} - \boldsymbol{\mu}(\epsilon))$+const, and its slope at the null value $\epsilon = 0$, the score of Section 10.1.2, is

$$t \equiv \left.\frac{\partial \ln \mathcal{L}}{\partial \epsilon}\right|_{\epsilon=0} \overset{(1)}{=} -\tfrac{1}{2}\Big[-\boldsymbol{d}^{\mathsf{T}}\mathbf{C}^{-1}(\boldsymbol{S} - \boldsymbol{\mu}_0) - (\boldsymbol{S} - \boldsymbol{\mu}_0)^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{d}\Big] \overset{(2)}{=} \boldsymbol{d}^{\mathsf{T}}\mathbf{C}^{-1}(\boldsymbol{S} - \boldsymbol{\mu}_0). \tag{12.99}$$

(1) The product rule on the quadratic form; $\partial(\boldsymbol{S} - \boldsymbol{\mu})/\partial\epsilon = -\boldsymbol{d}$, with $\boldsymbol{\mu}_0 = \boldsymbol{\mu}(0)$.
(2) $\mathbf{C}^{-1}$ is symmetric, so the two terms are equal.

The score compresses the p numbers into one, weighting each by how strongly it responds to ϵ and how noisy it is. Under H_0 it has mean zero and variance $\boldsymbol{d}^{\mathsf{T}}\mathbf{C}^{-1}\mathbf{C}\mathbf{C}^{-1}\boldsymbol{d} = \boldsymbol{d}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{d} \equiv F$, the Fisher information (10.20). At a small ϵ its mean is $\boldsymbol{d}^{\mathsf{T}}\mathbf{C}^{-1}(\epsilon\boldsymbol{d}) = \epsilon F$, so $t/\sqrt{F}$ is a unit normal shifted by $\epsilon\sqrt{F}$. By the Neyman–Pearson argument of Section 4.6, the one-sided test on t is the most powerful test of $\epsilon = 0$ against a small positive ϵ; since the sign of ϵ is not known in advance we use $|t|$. In practice $\boldsymbol{d}$ comes from the grid seeds 0–249 and the power is measured on seeds 250–499, so that the direction is not tuned to the maps it is tested on.

Example: how much the aim is worth. For the omnibus test write the whitened data as $\boldsymbol{x} = \boldsymbol{z} + \epsilon\mathbf{C}^{-1/2}\boldsymbol{d}$ with $\boldsymbol{z}$ a standard normal vector; then $\chi^2 = |\boldsymbol{x}|^2$ has mean $\mathbb{E}\,|\boldsymbol{z}|^2 + 2\epsilon\,\mathbb{E}[\boldsymbol{z}]^{\mathsf{T}}\mathbf{C}^{-1/2}\boldsymbol{d} + \epsilon^2\boldsymbol{d}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{d} = p + \epsilon^2 F$. Its null scatter is $\sqrt{2p}$, so the shift has to compete with $\sqrt{2 \cdot 22} \approx 7$ instead of 1. At the 5% level the score needs $\epsilon\sqrt{F} \approx 1.96$ for a 50% chance of detection (two-sided, the far tail neglected). The omnibus test needs its mean $22 + \epsilon^2 F$ to reach the 95th percentile of χ^2_{22}, 33.9, so $\epsilon^2 F \approx 11.9$ and $\epsilon\sqrt{F} \approx 3.5$: the omnibus test needs an ϵ about 1.8 times larger. With the Fisher error of the test's Betti vector, $F^{-1/2} = 0.0141$ (Section 12.26), the score's 50% point is $\epsilon \approx 0.028$ and the measured power at $\epsilon = 0.025$ is 54%; the omnibus test's estimated 50% point is $\epsilon \approx 0.049$, and the measured powers are 42% at 0.05 and 100% at 0.1. Each measured power comes from 250 maps, so it is uncertain by a few per cent, and all of them lean on the same reference set and the same covariance. The estimate ignores that the true null distribution is wider than χ^2_{22} (Section 12.24.2), which lowers the omnibus power.

Which summary sees ϵ? Figure 12.60 answers it for eight summaries. The power spectrum never rises above the false-alarm rate (score test 5% and omnibus 4% at $\epsilon = 0.1$): by construction it cannot. The one-point summaries see ϵ, but weakly: at $\epsilon = 0.05$ the score test on the skewness rejects 22% of maps, on the area fractions 22%. The topological summaries are far stronger: Euler characteristic 95%, Betti curves 97%, persistence histograms 92%. At area-fraction thresholds the Betti curves drop to 7% at $\epsilon = 0.05$ and 10% at 0.1.

Why do counts beat the skewness? The last number says that most of what the Betti curves detect is the one-point effect, values pushed up or down. Why then do they detect it so much

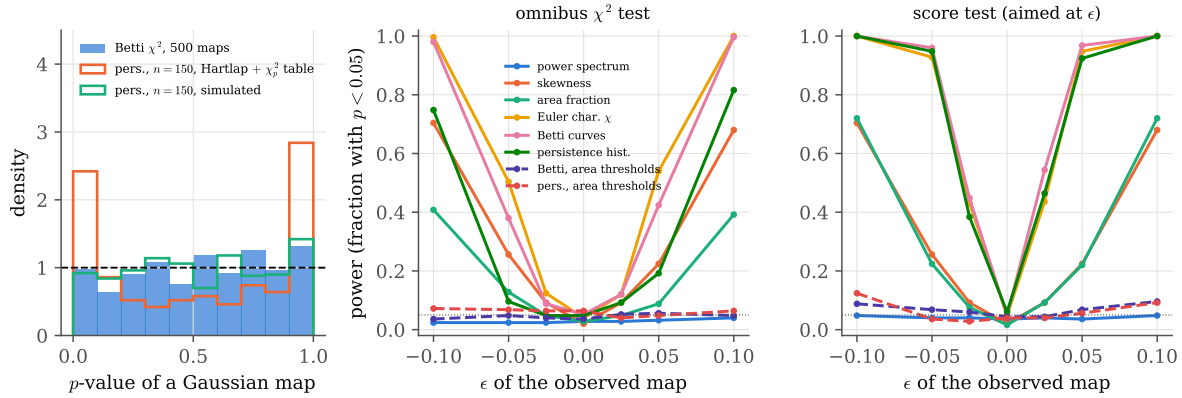


Figure 12.60: Left: p-values of Gaussian maps. Betti χ^2 with a simulated reference (filled): flat within its noise, which is larger than it looks because all 500 maps share one reference set. Persistence histogram with 150 simulations, Hartlap-corrected χ^2 referred to the χ^2_p table (orange): too many small *and* large values. The same with a simulated reference (green): flat. Middle and right: the fraction of maps rejected at $p < 0.05$ as a function of the true ϵ, for the omnibus χ^2 test and for the score test, for eight summaries. Dashed: summaries at area-fraction thresholds. The power spectrum stays at the 5% false-alarm level for every ϵ.

better than the skewness, which measures exactly that? Look at single entries. Per unit ϵ, measured in units of its own map-to-map scatter, the skewness responds by 27, the area fraction above $\nu = 2.5$ by 22, the number of pieces above 2.5 by 27 and the number of lakes below -2.5 by -26 (fewer lakes): all about the same size. The difference is in how much the entries repeat each other. The area fractions above 2.5 and above 3 have a correlation of 0.84 across maps, and the area fraction above 2.5 is correlated with the skewness at 0.79 and even with the area fraction below -2.5 at 0.41: they are moments of one histogram, pulled by the same extreme pixels and the same normalisation. The number of pieces above 2.5 and above 3 correlate at only 0.32, with the skewness at 0.36, and the hot pieces above 2.5 and the cold lakes below -2.5 at 0.09: separate peaks in separate places are nearly independent witnesses. The Betti vector collects many such witnesses, and the score test adds their evidence. The part of the signal that survives at fixed area fraction, the arrangement of values rather than the values themselves, is real (Section 12.23.3) but too weak for one map: even at $\epsilon = 0.1$ the score test on it rejects only 10% of the maps (Betti curves) and 9% (persistence histograms), against 5% for Gaussian maps.

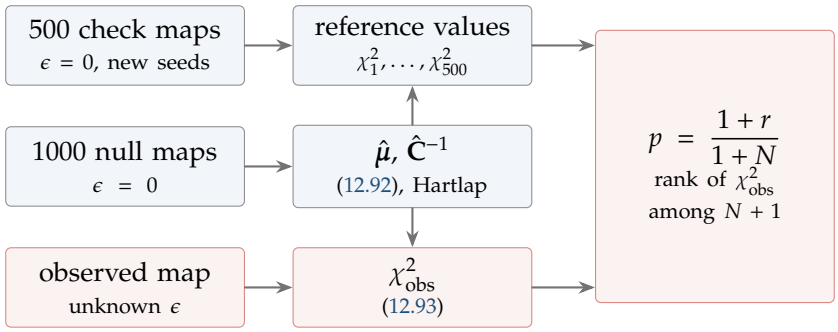


Figure 12.61: The test with a simulated null. One set of Gaussian maps (middle row) fixes the mean and the covariance, which every χ^2 uses; a second, independent set (top row) gives the distribution of the statistic under the null; the observed map (bottom row) goes through exactly the same steps, and its p-value is its rank among all of them. Every simulated quantity has its own map, so nothing is fitted to the maps it is judged by.

12.26 How well does one map measure ϵ? A Fisher forecast

A test answers yes or no. The next question is quantitative: if the map does have some ϵ, how precisely can one map measure it, and which summary measures it best? Before any data, the Fisher matrix answers this (Chapter 10). For a data vector that is Gaussian with mean $\boldsymbol{\mu}(\theta)$ and a covariance $\mathbf{C}$ that we hold fixed, the Fisher information is the mean part of (10.20), and with one parameter it is a single number:

$$F = \boldsymbol{d}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{d}, \qquad \boldsymbol{d} = \frac{\partial \boldsymbol{\mu}}{\partial \epsilon}, \qquad \sigma(\epsilon) \geq F^{-1/2}. \tag{12.100}$$

It is the same number that set the score test's variance in Section 12.25.3: the Fisher information is how much the mean moves, in units of the noise, per unit ϵ. Two features of our problem need comment. There is no formula for $\boldsymbol{d}$; it is estimated from simulations. And the full Gaussian Fisher matrix has a second term, $\frac{1}{2}\,\mathrm{Tr}[\mathbf{C}^{-1}\partial_\epsilon\mathbf{C}\,\mathbf{C}^{-1}\partial_\epsilon\mathbf{C}]$, from a covariance that changes with ϵ. Holding $\mathbf{C}$ at its $\epsilon = 0$ value, as almost every simulation-based analysis does, drops it; we return to what that costs in Section 12.27.3.

Derivatives from simulations, met before. In Section 12.20.2 we estimated derivatives of peak counts and Betti numbers of lensing maps with respect to Ω_m and σ_8 by central differences, with the same random numbers on both sides of each difference (*common random numbers*, (12.76)), and we found that the noise in estimated derivatives biases the Fisher matrix upwards by $\mathrm{Tr}[\mathbf{C}^{-1}\boldsymbol{\Sigma}_d]/N_{\mathrm{der}}$ (12.77). There, with two nearly degenerate parameters, that bias was dangerous. Here there is one parameter and no degeneracy, and the same machinery is much better behaved, which is worth seeing in numbers.

The question. What is $F^{-1/2}$ for each summary, and can we trust the estimate: does it depend on the step, on the number of simulations, on whether the random numbers are shared?

12.26.1 The derivative, seed by seed

For each grid seed k we have maps at $\epsilon = \pm h$ made from the same Gaussian field g_k. The per-seed central difference and its average over N_{der} seeds are

$$\boldsymbol{D}_k = \frac{\boldsymbol{S}_k(+h) - \boldsymbol{S}_k(-h)}{2h}, \qquad \hat{\boldsymbol{d}} = \frac{1}{N_{\mathrm{der}}}\sum_{k=1}^{N_{\mathrm{der}}} \boldsymbol{D}_k. \tag{12.101}$$

Each $\boldsymbol{S}_k(\pm h)$ contains a large seed-specific part, the particular arrangement of peaks in g_k, and a small part that responds to ϵ. The difference removes the first and keeps the second (Figure 12.62).

Example: how much the shared seed buys. For one entry, the number of pieces at $\nu = 1$, call σ^2 the variance of $S_k(\pm h)$ across seeds and ρ the correlation between $S_k(+h)$ and $S_k(-h)$ of the same seed. Then

$$\mathrm{Var}\big[S_k(+h) - S_k(-h)\big] \overset{(1)}{=} \mathrm{Var}\, S_k(+h) + \mathrm{Var}\, S_k(-h) - 2\,\mathrm{Cov}\big[S_k(+h), S_k(-h)\big] \overset{(2)}{=} 2\sigma^2(1-\rho), \tag{12.102}$$

(1) The variance of a difference (12.76).
(2) Equal variances on both sides and $\mathrm{Cov} = \rho\sigma^2$.

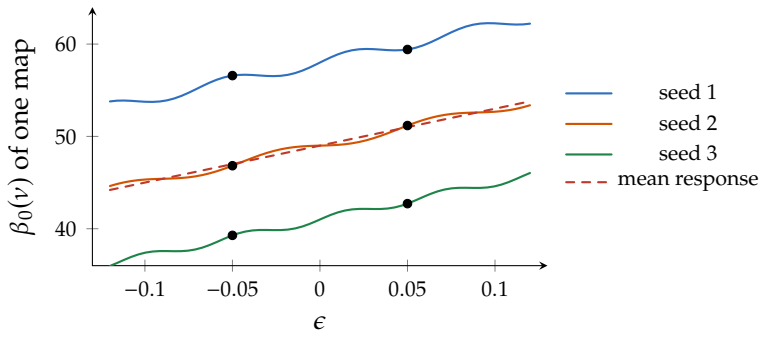


Figure 12.62: Why common random numbers work. Each seed (one Gaussian field g) gives a count that depends on ϵ; the three seeds differ by far more than ϵ moves them, but their slopes are nearly the same. A difference taken along one seed (the two dots on each curve, at $\epsilon = \pm h$) cancels the seed's offset and measures the slope; a difference between two seeds would measure mostly the offsets.

against $2\sigma^2$ for maps from different seeds. The measured correlation is $\rho = 0.983$, so the shared seed reduces the variance by $1/(1-\rho) \approx 61$; the script measures the ratio directly as 62. One pair with shared random numbers is worth about 61 pairs without.

12.26.2 Fisher errors for every summary

The script `code/ch12/c05_fisher_post.py` does this for each summary:

Listing 12.20: Fisher information with Monte Carlo derivatives, and its noise bias (`code/ch12/c05_fisher_post.py`)

```
def null_model(keys):
    N = vec("null_", keys)
    ok = N.std(0) > 1e-9
    N = N[:, ok]
    n, p = N.shape
    Ci = np.linalg.inv(np.atleast_2d(np.cov(N.T))) * (n - p - 2) / (n - 1)
    return N.mean(0), Ci, ok, n, p


def fisher(keys, step=0.05, seeds=slice(None), crn=True):
    """(F, noise bias of F, derivative, its standard error) for one summary."""
    mu, Ci, ok, n, p = null_model(keys)
    G = vec("grid_", keys)[seeds][..., ok]
    jp, jm = eg.index(step), eg.index(-step)
    if crn:
        D = (G[:, jp] - G[:, jm]) / (2 * step)            # same g on both sides
    else:
        h = len(G) // 2                                   # +step from one half, -step
    from the other
        D = (G[:h, jp] - G[h:2 * h, jm]) / (2 * step)
    d = D.mean(0)
    Sd = np.atleast_2d(np.cov(D.T)) / len(D)              # covariance of the averaged
    derivative
    F = d @ Ci @ d
    return F, np.trace(Ci @ Sd), d, np.sqrt(np.diag(Sd))
```

`null_model` gives the mean and the Hartlap-corrected inverse covariance from the 1000 null maps. `fisher` forms the per-seed differences (12.101), either along a seed (`crn=True`) or between different seeds, averages them, and returns $F = \hat{\boldsymbol{d}}^{\mathsf{T}}\hat{\mathbf{C}}^{-1}\hat{\boldsymbol{d}}$ together with the expected noise bias $\mathrm{Tr}[\hat{\mathbf{C}}^{-1}\boldsymbol{\Sigma}_d]/N_{\mathrm{der}}$, where `Sd` is the covariance of the per-seed derivatives divided by their number.

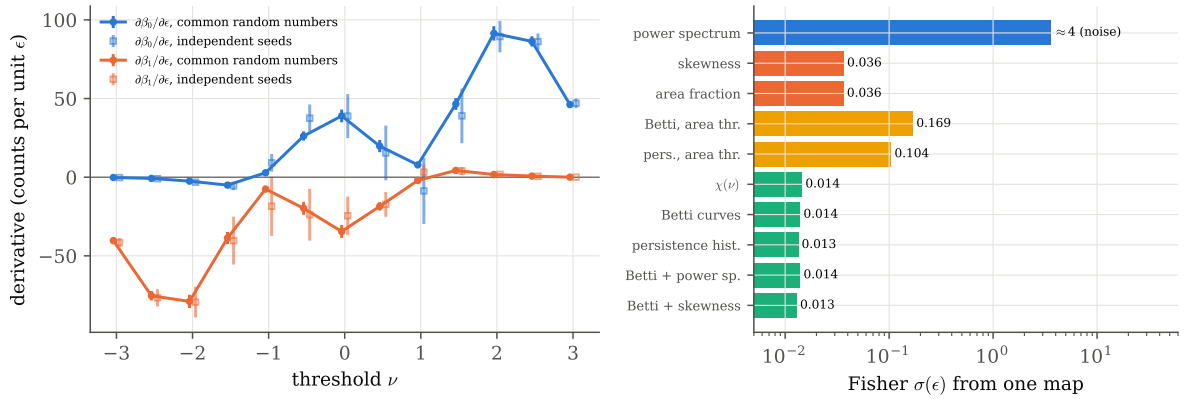


Figure 12.63: Left: the derivative of the Betti curves with respect to ϵ (±2 standard errors), from 500 seeds with shared random numbers (filled) and from 250 pairs of different seeds (open): the same curve, with error bars several times smaller. The response is concentrated in the tails, more pieces above $\nu \approx 2$ and fewer lakes below $\nu \approx -2$. Right: the Fisher error $F^{-1/2}$ on ϵ from one 256×256 map. Blue: power spectrum (no information, the bar is pure derivative noise); orange: one-point summaries; yellow: topological summaries at area-fraction thresholds; green: topological summaries at σ-thresholds and combinations.

Figure 12.63 gives the results, with steps $h = 0.05$ and 500 seeds. The Euler characteristic gives $\sigma(\epsilon) = 0.0145$, the Betti curves 0.0139, the persistence histograms 0.0133. The skewness gives 0.0365 and the area fractions 0.0364, nearly three times worse; at area-fraction thresholds the Betti curves give 0.169 and the persistence histograms 0.104. These are the same rankings the tests of Figure 12.60 showed, now as error bars. Adding the skewness to the Betti curves improves the error only to 0.0129: the Betti curves already contain most of what the histogram of values knows. Adding the power spectrum changes nothing (0.0139).

Example: a Fisher error made of noise. The power spectrum bar reads $\sigma(\epsilon) \approx 3.5$, as if one map could measure ϵ to within a few units. It cannot measure it at all: the true derivative is zero (12.87). What the bar shows is (12.77) with $\boldsymbol{d} = 0$: the estimated $\hat{\boldsymbol{d}}$ is pure noise, and $\hat{\boldsymbol{d}}^{\mathsf{T}}\hat{\mathbf{C}}^{-1}\hat{\boldsymbol{d}}$ is a positive number whose average is the noise term alone. The script's estimate of that term is 1.52 times the measured F; a measured information equal to (or, within its own noise, below) its expected noise bias is the signature of no information. For the topological summaries the same ratio is 0.1% (Betti curves), 0.3% (persistence histograms) and 3.7% (persistence at area-fraction thresholds, where the signal is weakest).

Example: why the noise bias is small here. Even without shared seeds the Betti Fisher error is 0.0138 and the bias below one per cent. Estimate it: without common random numbers, $\boldsymbol{D}_k$ has covariance about $2\mathbf{C}/(2h)^2 = 200\,\mathbf{C}$ for $h = 0.05$, so with $N_{\text{der}} = 250$ pairs

$$\frac{\mathrm{Tr}[\mathbf{C}^{-1}\boldsymbol{\Sigma}_d]}{N_{\text{der}}} \overset{(1)}{=} \frac{\mathrm{Tr}[\mathbf{C}^{-1}\,200\,\mathbf{C}]}{250} \overset{(2)}{=} \frac{200 \times 26}{250} \approx 21,$$

(1) The per-pair covariance $\boldsymbol{\Sigma}_d$: two independent maps, each with covariance $\mathbf{C}$, difference divided by $2h = 0.1$.

(2) $\mathrm{Tr}\,\mathbb{1} = p = 26$.

against $F = 1/0.0139^2 \approx 5182$: a 0.4% effect. The signal is strong, one parameter has no degeneracy to break, and noise in the derivative cannot fake a direction the data do not have. In Section 12.20.2 the true information across the degeneracy was a percent of the diagonal, and a bias of the same size changed the answer.

Checks on the derivative. Three changes should not move a trustworthy number, and they do not. The step: with $h = 0.025$, 0.05 and 0.1 the Betti error is 0.0138, 0.0139 and 0.0147; the largest step starts to feel the curvature of $\boldsymbol{\mu}(\epsilon)$ (a central difference has an error of order $h^2\mu'''$, Section 10.5). The number of seeds: with 25, 50, 100, 250 and 500 pairs it is 0.0135, 0.0135, 0.0139, 0.0140 and 0.0139. And shared versus independent seeds, 0.0139 against 0.0138.

Verification and research with the same machinery. Section 12.23.3 used simulations to *verify* a known result: Matsubara's perturbative formula for the mean Euler characteristic. The Fisher errors above are *research* numbers: no formula gives the covariance of Betti curves or the response of a persistence histogram, and the simulated derivatives and covariance are the only route. The same split appears in research. Cole and Shiu forecast $f_{\rm NL}$ from persistence diagrams of simulated CMB patches with exactly this construction, a simulated mean at a few values of $f_{\rm NL}$ and a covariance from 2000 Gaussian maps [271]; weak-lensing analyses forecast S_8 from peaks and persistent Betti numbers in the same way, with emulators of the mean in place of finite differences [56] (Section 12.20).

12.27 From one map to a posterior for ϵ

The Fisher error is a forecast: the width the posterior will have on average, if the likelihood is Gaussian and the mean is linear in the parameter. Now we take one map and compute the posterior itself. The reasoning is the one of parameter estimation in Chapter 5, applied to a new kind of data. The observed event is the Betti vector $\boldsymbol{S}_{\rm obs}$ of one map. The candidate explanations are the values of ϵ. For each candidate the likelihood asks how readily a map with that ϵ would have produced these counts; Bayes' theorem combines the answers with a prior.

The question. For one map with $\epsilon_{\rm true} = 0.05$: which values of ϵ could have produced its Betti curves, and how probable is each? And over many such maps, do the 68% intervals we report contain the truth 68% of the time?

12.27.1 The likelihood

We model the Betti vector as Gaussian, $\boldsymbol{S} \sim \mathcal{N}(\boldsymbol{\mu}(\epsilon), \mathbf{C})$. In the language of Section 3.8, for a given ϵ the model predicts the mean counts $\boldsymbol{\mu}(\epsilon)$; the observed counts differ from it by a fluctuation $\boldsymbol{S}_{\rm obs} - \boldsymbol{\mu}(\epsilon)$; and the likelihood is the probability density of exactly that fluctuation under the map-to-map scatter $\mathbf{C}$:

$$\ln \mathcal{L}(\epsilon) = -\frac{1}{2}\big(\boldsymbol{S}_{\rm obs} - \boldsymbol{\mu}(\epsilon)\big)^{\mathsf{T}}\mathbf{C}^{-1}\big(\boldsymbol{S}_{\rm obs} - \boldsymbol{\mu}(\epsilon)\big) + \text{const.} \tag{12.103}$$

A value of ϵ whose prediction requires an improbable fluctuation is disfavoured. Three ingredients have to be supplied, and each comes from simulations.

The mean. At the seven grid values of ϵ we average the Betti vectors of seeds 0–249. A straight line through them would miss the curvature that the step test of Section 12.26.2 hinted at, so each of the 26 components is fitted with a parabola in ϵ. The largest misfit of any component at any grid point is 0.03 of the scatter of one map: the parabola is as good as the simulations can tell.

The covariance. $\hat{\mathbf{C}}$ from the 1000 null maps, with the Hartlap factor, held fixed.

The prior. Uniform on $|\epsilon| < 0.15$, wide compared with the expected width and narrow enough that the parabola is never extrapolated.

Watch out: A Gaussian likelihood for counts is a model choice

(12.103) is not derived from the physics; it is an approximation to $P(\boldsymbol{S} \mid \epsilon)$ that is good when every count is large and when $\mathbf{C}$ hardly changes with ϵ. Rare-feature bins violate the first (Section 12.24.2), and the coverage test below shows how the second fails. Replacing $\mathbf{C}$ by $\mathbf{C}(\epsilon)$ inside a Gaussian likelihood is not the cure: for non-Gaussian data it adds a spurious term to the information and can make the posterior narrower than the Cramér–Rao bound allows [273].

12.27.2 One map, three ways of computing its posterior

The held-out seeds 250–499 at $\epsilon = 0.05$ serve as mock data. For the picture we take the first of them whose posterior mean lies within one posterior standard deviation of the truth, a typical map, seed 250; the coverage test below uses all 250. With one parameter the posterior can be computed on a grid of 1201 values of ϵ, which gives the reference answer. The Metropolis sampler of Section 9.10 and the ensemble sampler emcee (Section 9.16.3) are then two independent checks:

Listing 12.21: The likelihood of the Betti vector, with a uniform prior (`code/ch12/c05_fisher_post.py`)

```
def mu_of(e):
    return coef[0] * e ** 2 + coef[1] * e + coef[2]


PRIOR = 0.15


def loglike(e, S):
    if abs(e) > PRIOR:
        return -np.inf
    r = S - mu_of(e)
    return -0.5 * r @ Ci @ r
```

`mu_of` evaluates the fitted parabolas, `loglike` returns (12.103) inside the prior and $-\infty$ outside, which the samplers read as "never accept". The chain makes 20000 steps with Gaussian proposals of width 1.5 times the grid posterior's standard deviation, somewhat narrower than the $2.38/\sqrt{D} = 2.38$ standard deviations that Section 9.11.1 found optimal for a Gaussian target; the price is a little efficiency, the chain accepting more proposals than it would at the optimal width.

The grid posterior has mean 0.0364 and standard deviation 0.0148; its central 68% interval is $[0.0216, 0.0509]$, which contains the true 0.05. The Metropolis chain gives 0.0361 ± 0.0147, with acceptance fraction 0.59 and integrated autocorrelation time 5.5 steps (Section 9.12.2), so its 19000 post-burn-in steps are worth about $19000/5.5 \approx 3400$ independent draws, enough for a standard deviation good to about one per cent. emcee gives 0.0368 ± 0.0144. The Fisher forecast was 0.0139. Four routes, one answer: for this problem the likelihood is nearly Gaussian in ϵ and the forecast is the posterior.

12.27.3 Do the intervals cover?

A 68% credible interval is a statement about one map. Whether intervals constructed this way contain the truth in 68% of repeated experiments is a separate, frequentist property, *coverage* (Section 4.12), and it can be tested only by repeating the experiment, which simulations let us do. For each of the 250 held-out maps at $\epsilon_{\text{true}} = 0.05$ we compute the grid posterior and check

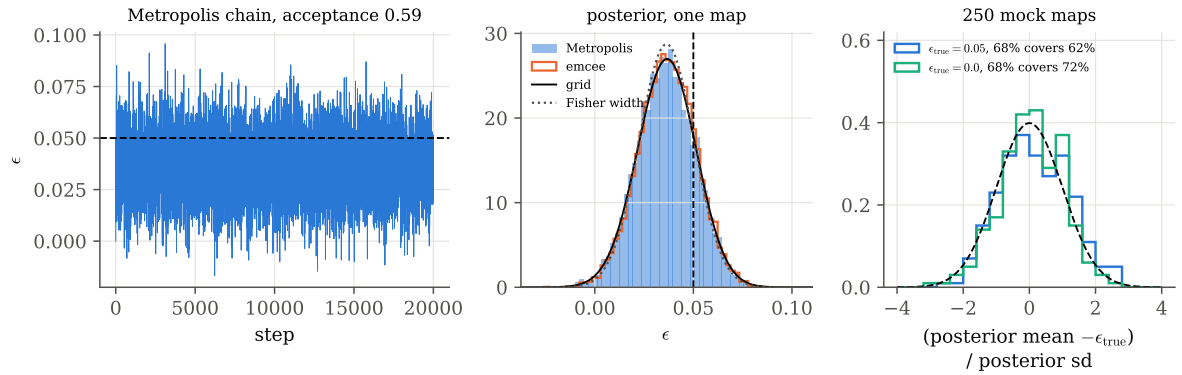


Figure 12.64: Left: the Metropolis chain for one map with $\epsilon_{\text{true}} = 0.05$ (dashed), well mixed from the first steps. Middle: its histogram, the emcee histogram, the posterior on a grid and a Gaussian of the Fisher width centred on the grid mean: four curves that lie on each other. Right: for 250 maps, the distance of the posterior mean from the truth in units of the posterior standard deviation, at $\epsilon_{\text{true}} = 0.05$ and 0; dashed: the unit normal these should follow if the intervals are honest.

whether its central 68% interval contains 0.05; then the same at $\epsilon_{\text{true}} = 0$.

At $\epsilon_{\text{true}} = 0$ the intervals contain the truth in 72% of maps; at $\epsilon_{\text{true}} = 0.05$ only in 62%. With 250 maps the binomial uncertainty on such a fraction is $\sqrt{0.68 \times 0.32/250} = 3\%$. The right panel of Figure 12.64 shows the standardised errors (posterior mean $- \epsilon_{\text{true}}$)/posterior sd at $\epsilon_{\text{true}} = 0.05$: their standard deviation is 1.05 ± 0.05 instead of 1, and their mean 0.16 ± 0.07 instead of 0. With 250 maps these numbers are themselves noisy, so they hint at a problem rather than prove one. Is there a reason the posteriors should be too narrow at $\epsilon = 0.05$?

What could make them too narrow? Go through the ingredients one at a time. The prior is wide, so it cannot be it. The mean $\boldsymbol{\mu}(\epsilon)$ was fitted at the very values of ϵ we test and fits to within 0.03 of one map's scatter, so it is not the mean. That leaves the covariance, which was measured at $\epsilon = 0$ and used at every ϵ. Is the scatter of the counts larger at $\epsilon = 0.05$? Project each map onto the direction that matters, the compressed statistic $t = \boldsymbol{d}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{S}$ of (12.99); its variance over the grid maps at $\epsilon = 0.05$ is 1.15 ± 0.09 times its variance at $\epsilon = 0$. Each seed gives both maps from the same Gaussian field, so much of the noise cancels in this ratio, but not all: it sits above one by somewhat less than two of its errors. That is a hint, not a proof, and it points the same way as the coverage. There is also a reason to expect it: a positive ϵ makes the high peaks higher and more numerous, and counts of rarer, more extreme features fluctuate more. If so, the likelihood underestimates the scatter by a factor $k = \sqrt{1.15} \approx 1.070$ in exactly the direction that determines ϵ.

Example: the coverage a wrong width produces. Suppose the estimate $\hat{\epsilon}$ scatters around the truth with standard deviation $k\sigma$ but we report $\hat{\epsilon} \pm \sigma$. The interval covers the truth when $|\hat{\epsilon} - \epsilon_{\text{true}}| < \sigma$:

$$\mathrm{P}\big(|\hat{\epsilon} - \epsilon_{\text{true}}| < \sigma\big) \overset{(1)}{=} \mathrm{P}\Big(|Z| < \frac{1}{k}\Big) \overset{(2)}{=} 2\Phi\Big(\frac{1}{k}\Big) - 1 \overset{(3)}{=} 2\Phi(0.934) - 1 \overset{(4)}{=} 0.650, \tag{12.104}$$

(1) Write $\hat{\epsilon} - \epsilon_{\text{true}} = k\sigma Z$ with Z a unit normal, and divide by $k\sigma$.
(2) Φ is the standard normal distribution function, and the normal is symmetric.
(3) $k = \sqrt{1.15} = 1.070$.
(4) $\Phi(0.934) = 0.825$ from a table.

A 7% error in the width turns a 68% interval into a 65% one. The simulations found 62 ± 3%: consistent with this cause, though 250 maps cannot rule out others, such as the offset of the mean. At $\epsilon_{\text{true}} = 0$ the covariance is the right one, and the coverage, 72%, agrees with 68% within its uncertainty.

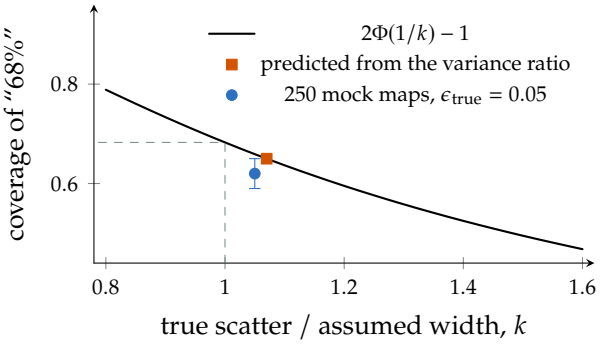


Figure 12.65: The coverage of a nominal 68% interval when the true scatter of the estimate is k times the width we report, (12.104). Honest intervals sit at $k = 1$ (dashed). The variance ratio measured for the compressed Betti statistic predicts the orange square; the blue point is the coverage actually found over 250 mock maps, placed at the measured spread of their standardised errors, with its binomial error bar.

The remedy follows from the diagnosis. Measure the covariance near the answer rather than at an arbitrary fiducial point: after a first pass, simulate the null set again at the best-fit ϵ and repeat. Or drop the Gaussian form altogether and let simulations supply the likelihood, which is the subject of Section 12.29. What we should not do is skip the test: coverage is the only check that the error bars we publish mean what they say, and for a summary without a likelihood formula, simulations are the only way to run it.

12.28 The same experiment on the CMB sky

A temperature map is a function on the celestial sphere, $T : S^2 \to \mathbb{R}$, and we never see it directly. The telescope blurs it with its beam, the detectors add noise, and the Galaxy and bright sources force us to throw away part of the sky. In Chapter 8 each of these was an operation on the map whose effect on the power-spectrum estimator we measured by passing simulated skies through it: the beam and pixel window as a multiplicative transfer function, the noise as an additive bias, the mask as a coupling of multipoles. MASTER is built on that idea, effects of the observation "calibrated in Monte Carlo simulations of the modelled observation and analysis" [Hivon §1]. Topology has exactly the same problem. A mask cuts islands in two at its edge, noise creates tiny islands that the sky does not have, a beam merges islands the sky keeps apart. There is no formula for any of this on a masked, noisy sky, but there is the same remedy: treat each observational effect as an operator on the map, push Gaussian skies through it, and measure what it does to the summary.

The questions. What do the beam, the noise and the mask each do to the counts of hot and cold spots and to their persistence? And once all three are applied, can Gaussian simulations, passed through the same operators, tell a Gaussian sky from a slightly non-Gaussian one with the same power spectrum, as the *Planck* test does?

12.28.1 Counting on the HEALPix sphere

The skies are drawn from the fiducial *Planck* 2018 temperature spectrum (Chapter T1) at $N_{\rm side} = 128$, with multipoles up to $\ell_{\rm max} = 383$ and pixels of 27′. The persistence of the hot spots is computed by the union–find algorithm of Section 12.11.1, with the grid neighbours replaced by the HEALPix neighbours that `healpy.get_all_neighbours` returns, eight per pixel (seven at a few special pixels):

Listing 12.22: Hot-spot persistence on the sphere by union–find (compiled with numba) (`code/ch12/c06_sphere.py`)

```
def h0_superlevel(u, nbr, active):
    """(birth, death) pairs of the islands of {u >= t} as t decreases (elder rule).

    nbr: (npix, k) neighbour table (-1 = none); active: pixels that are observed.
    Essential islands (one per observed region) get death = -inf.
    """
    n = u.size
    order = np.argsort(-u)
    parent = -np.ones(n, np.int64)
    birth = np.zeros(n)
    out_b = np.empty(n)
    out_d = np.empty(n)
    m = 0
    for p in order:
        if not active[p]:
            continue
        parent[p] = p
        birth[p] = u[p]
        for j in range(nbr.shape[1]):
            q = nbr[p, j]
            if q < 0 or parent[q] < 0:
                continue
            rp = _find(parent, p)
            rq = _find(parent, q)
            if rp == rq:
                continue
            if birth[rp] < birth[rq]:                    # the younger island dies
                young, old = rp, rq
            else:
                young, old = rq, rp
            if birth[young] > u[p]:                      # skip zero-persistence pairs
                out_b[m] = birth[young]
                out_d[m] = u[p]
                m += 1
            parent[young] = old
    for p in range(n):                                   # survivors: essential classes
        if active[p] and parent[p] == p:
            out_b[m] = birth[p]
            out_d[m] = -np.inf
            m += 1
    return out_b[:m], out_d[:m]
```

The pixels are visited from the hottest down (`order`). Each new pixel starts its own island, born at its own value. It is then joined to every already-active neighbour; when two different islands meet, the younger one, the one born at the lower value, dies at the current value and is recorded as a (birth, death) pair: the elder rule of Section 12.9. Pairs with zero persistence, a pixel merging at its own value, are skipped. Islands that never die get death $-\infty$. Masked pixels are never activated. The function is compiled with numba; drawing one full sky, observing it and computing both diagrams and both Betti curves takes 208 ms.

Hot spots and cold spots. Hot pixels touching at a corner belong to one island; cold pixels must share an edge, which on HEALPix means four of the eight neighbours (the diagonal directions SW, NW, NE, SE in healpy's ordering). This is the dual pair of Section 12.2.2. Cold spots are computed by the same function applied to $-T$ with the four edge-neighbours. The Betti curves are read off the diagrams: the number of islands of $\{T \geq \nu\}$ is the number of pairs alive at ν, born at or above ν and dying below it (12.41).

Example: the Euler characteristic from hot and cold spots. On the full sphere every lake of the hot set is a cold piece that does not reach "outside". On a sphere there is no outside, but there is one cold piece that plays its role. Precisely, Alexander duality on the sphere gives $\beta_1(\text{hot}) = \beta_0(\text{cold}) - 1$ when both sets are non-empty (see the topology notes). Then

$$\chi(\text{hot set}) \overset{(1)}{=} \beta_0(\text{hot}) - \beta_1(\text{hot}) \overset{(2)}{=} N_{\text{hot}} - (N_{\text{cold}} - 1) = N_{\text{hot}} - N_{\text{cold}} + 1, \tag{12.105}$$

(1) The Euler characteristic of a planar or spherical set is pieces minus holes (12.5).
(2) The duality, with N_{hot} and N_{cold} the numbers of hot and cold spots.

This is why *Planck* can compute its genus as a difference of hot-spot and cold-spot counts [114] (their eq. 15, Section T3.5.1); the +1 is the sphere's own contribution, the $2\Psi(\nu)$ term of (12.29) at low threshold.

Checks. On one noisy sky the Betti curves from the diagrams were compared with a direct count of connected components (scipy's sparse-graph labelling on the same neighbour lists, `lib_sphere.betti_sphere`) at all thirteen thresholds, for hot and cold spots: 0 disagreements in 26 comparisons. On a smaller sky ($N_{\text{side}} = 32$) the hot-spot diagram was recomputed with gudhi, by inserting every pixel as a vertex and every neighbour pair as an edge into a `SimplexTree` with the lower-star filtration of $-T$: our code finds 198 finite pairs, gudhi 198, and the largest difference in any birth or death is $0.0e + 00$.

12.28.2 Three operators, one at a time

The beam erases small structures. Figure 12.66, left, shows the mean numbers of hot and cold spots of 100 full skies for Gaussian beams of 40′, 80′ and 160′. The peak number of hot spots is 2556, 768 and 163. What should the ratios be? By the Gaussian kinematic formula, every count of features scales with $\lambda = \sigma_1^2/(2\sigma_0^2)$, the gradient variance per component in units of the field variance ((12.29)); on the sphere $\sigma_0^2 = \sum_\ell \frac{2\ell+1}{4\pi} C_\ell b_\ell^2$ and $\sigma_1^2 = \sum_\ell \frac{2\ell+1}{4\pi} \ell(\ell+1) C_\ell b_\ell^2$ ((T3.3), (T3.4)) with the beam b_ℓ. These give $\lambda = 6059$ and 307 per steradian at 40′ and 160′, a ratio of 19.8; the measured ratio of peak counts is 15.7. The same deficit shows in the Euler characteristic at $\nu = 1$, (12.105): 682 against the formula's 752 at 80′, and 2327 against 2932 at 40′. The cause is the pixel, not the formula: at 40′ the beam is only one and a half 27′ pixels wide, the smallest islands are not resolved, and the 8-neighbour rule merges what the continuous field keeps apart (Section 12.2.2). A beam is an operator that lowers λ, and with it every count, and the counting grid is part of the operator.

Noise creates short-lived features. White noise is added to every pixel of the 80′ map, with standard deviation $\sigma_N = 0.2$ and 0.5 times the rms σ_S of the beamed signal; the same skies and the same noise draws are used for every level, so differences between levels are not cosmic variance. Noise has power at every multipole up to the pixel scale, so it raises σ_1 far more than σ_0, and with it λ. The number of finite hot-spot pairs per sky goes from 3897 without noise to 8880 and 16143, and the hot spots above $\nu = 1$ from 765 to 3114 at $\sigma_N = 0.5\sigma_S$. Where do

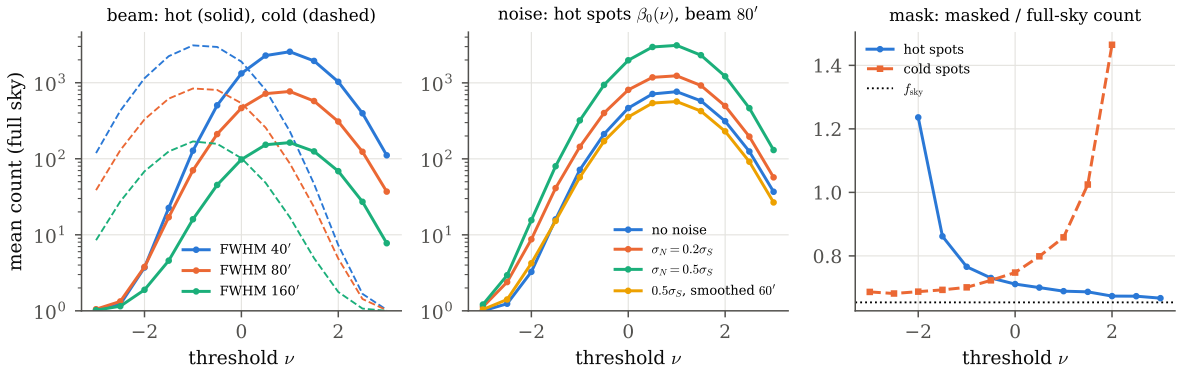


Figure 12.66: Three operators on the full-sky Betti curves, 100 skies each. Left: beams of 40′, 80′ and 160′; the counts of hot spots (solid) and cold spots (dashed) fall with the beam as λ does. Middle: white pixel noise of 0.2 and 0.5 times the signal rms on the 80′ map multiplies the hot spots; smoothing the noisy map by another 60′ removes them again. Right: the counts on a masked sky divided by the full-sky counts, against the observed sky fraction (dotted).

the new features sit in the diagram? Figure 12.67 shows it: along the diagonal. The stability theorem of Section 12.10 says that a perturbation of size η can only create pairs of persistence at most 2η, and a single pixel of noise moves the map only locally; the extra features are tiny bumps on the slopes of real hot spots, born and dead within a fraction of σ. The number of pairs with persistence below 0.1σ triples, from 1119 to 2985 and 3004. Smoothing the noisy map by a further 60′ is a second operator that undoes most of the first: 3555 pairs, 1307 short ones and 568 hot spots above $\nu = 1$, now fewer than without noise because the total beam is wider.

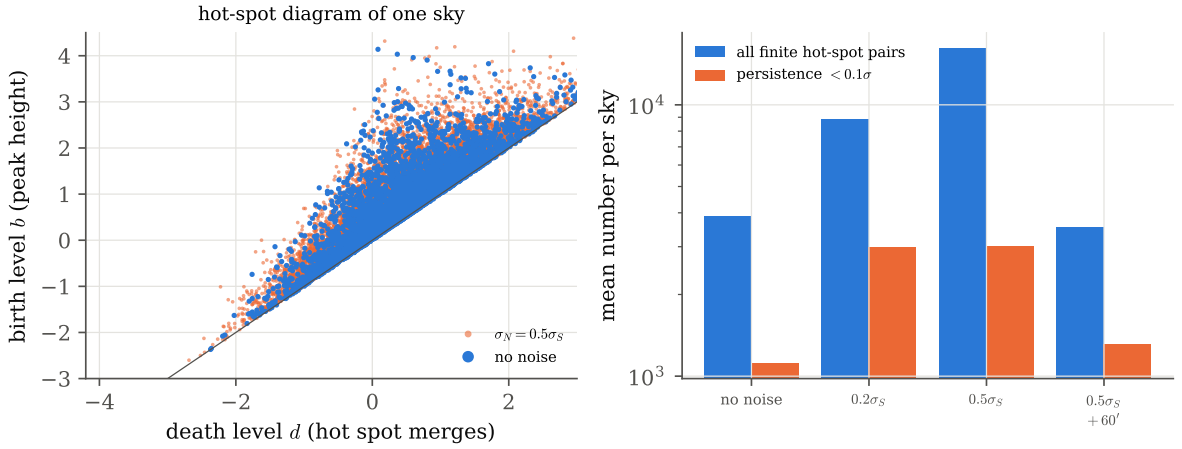


Figure 12.67: Left: the hot-spot persistence diagram of one sky with the 80′ beam, without noise (blue) and with white noise of $0.5\sigma_S$ (orange); every point is a hot spot born at height b that merges at level d. The noisy sky has about four times as many points, and the extra ones lie close to the diagonal (partly hidden under the blue points, which are drawn on top). Right: the mean number per sky of all finite pairs and of pairs with persistence below 0.1σ, for no noise, $0.2\sigma_S$, $0.5\sigma_S$, and $0.5\sigma_S$ smoothed by a further 60′.

The mask cuts and isolates. The mask removes the band $|b| < 20°$ around the Galactic equator and 100 discs of radius 1° at random positions, leaving $f_{\rm sky} = 0.654$ of the sky. If the mask simply removed area, every count would fall to $f_{\rm sky}$ of its full-sky value. At $\nu = 1.5$ the hot spots fall to 0.687 of the full-sky count: slightly more than $f_{\rm sky}$, because an island that crosses the edge of the mask is cut, and a cut island can fall apart into two. Where the excursion

set is large, the effect is much bigger (Figure 12.66, right). At high thresholds the cold set is almost the whole sky, a single piece on the full sphere; on the masked sky the band alone splits it into a northern and a southern piece, and every source hole surrounded by hot pixels can isolate another. The ratio of cold counts climbs far above one. A mask has a topology of its own, two caps with a hundred holes, and it adds that topology to every large excursion set, the topological counterpart of the mode coupling of Section 8.10.

12.28.3 A *Planck*-style Gaussianity test on masked, noisy skies

Their question. *Planck* asked whether the shapes of the observed microwave sky are those of a Gaussian, isotropic field [114]. The test (Section T3.5.1) compares the Minkowski functionals of the data with 999 end-to-end simulations through a χ^2 and reports the fraction of simulations with a larger χ^2. On the real sky the answer was "yes, Gaussian", with probabilities between 24% and 99.8% for temperature (their Table 5). We reproduce the test, on simulations only (the real sky is analysed in Chapter 13), with Betti curves in place of the functionals, and with a non-Gaussian sky to see what a failure would look like.

Their equations, in ours. Their data vector is 36 numbers, three functionals at twelve thresholds; ours is $p = 26$ numbers, hot and cold counts at thirteen thresholds, after dropping entries that never vary. Their χ^2 (their eqs. 20–22) is (12.93) with mean and covariance from simulations; their probability is the simulated p-value (12.97), whose validity we derived in (12.98). Their genus is (12.105) divided by the sky area. Nothing else is needed.

Our setup. The null model is a Gaussian sky with the 80′ beam, white noise of $0.2\sigma_S$ and the mask above: 1000 skies, of which 600 give the mean and the covariance (Hartlap factor 0.955) and 400 the reference distribution of χ^2. Calibration: we split the reference skies in two, use one half as the reference and test the other 200 Gaussian skies against it; 8.0% of them get $p < 0.05$, with an uncertainty of about 2.2%, and a two-sample Kolmogorov–Smirnov test of the two halves gives $p = 0.07$: with only 200 test skies, both are consistent with a calibrated test. (Ranking each reference sky among all the others would give 5% whatever the skies, since the ranks of a set among itself are just a reordering.) The alternative skies are made as in Section 12.23.1, now on the sphere: before the beam, the Gaussian sky s becomes $s + a(s^2 - \sigma^2)/\sigma$, and each multipole is rescaled by $R_\ell = \sqrt{C_\ell^s / C_\ell^\phi}$ so that the power spectrum is unchanged; R_ℓ is measured from the spectra of 100 skies with and without the quadratic term, built from the same Gaussian skies so that their ratio is nearly free of cosmic variance. For each of $a = 0.02$, 0.05 and 0.15, 100 skies go through the same beam, noise and mask. As a control, the same test is run with the pseudo-C_ℓ of the masked map in a dozen logarithmic bins (Section 8.9.3).

How large are these amplitudes? On large scales the temperature is $-\Phi/3$, and the local model gives $a = -3 f_{\rm NL} \sigma_t$ with $\sigma_t \approx 2 \times 10^{-5}$ (Section T3.2.3). Then $a = 0.02$ corresponds to $|f_{\rm NL}| \approx 0.02/(6 \times 10^{-5}) \approx 330$, some sixty times the *Planck* error (T3.15). Our skies are an exaggerated, pointwise version of the local model at $N_{\rm side} = 128$; they test the machinery, not the inflationary bound.

Side by side. Figure 12.68 is our version of the bracket plots: the counts of one sky, minus the simulation mean, in units of the simulation scatter. The Gaussian "data" sky stays inside the band at every threshold and has $p = 0.97$. The non-Gaussian sky with $a = 0.15$ leaves it in a characteristic pattern, too many hot spots at high thresholds and too few cold spots below -1, and its p is 0.002, the smallest value 400 reference skies allow (1/401).

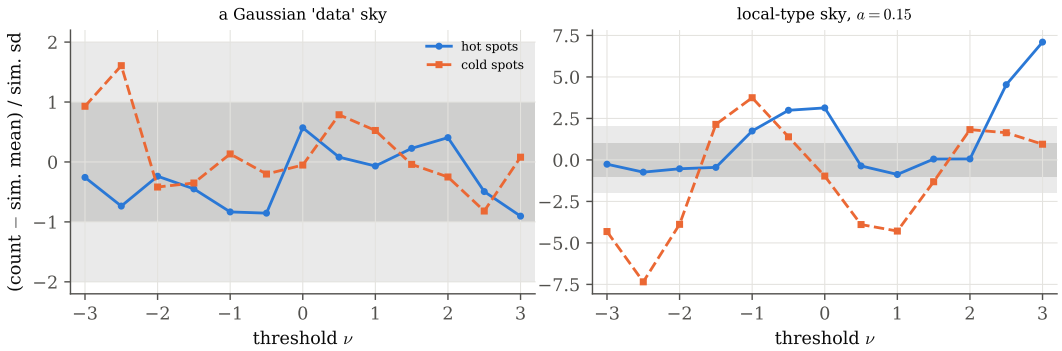


Figure 12.68: Hot-spot (solid) and cold-spot (dashed) counts of one masked, noisy, beamed sky, minus the mean of 600 Gaussian simulations processed identically, in units of their standard deviation; grey bands: ±1 and ±2. Left: a Gaussian sky, bracketed by the simulations as *Planck*'s sky was. Right: a local-type sky with $a = 0.15$ and the same power spectrum: many thresholds are several standard deviations out.

	Planck 2018 [114]	ours
summary	v_0, v_1, v_2 at 12 thresholds (36)	hot, cold counts at 13 thresholds (26)
simulations	999 FFP10, full pipeline	600 + 400 Gaussian, beam + white noise + mask
resolution	$N_{\rm side}$ 16 to 2048	$N_{\rm side} = 128$, 80′
sky fraction	≈ 0.75	0.654
Gaussian sky, probability	24% to 99.8% (real T maps)	$p = 0.97$ (one Gaussian sky); $8.0 \pm 2.2\%$ below 0.05
non-Gaussian sky	not seen	$a = 0.02$: median $p = 0.22$, 14% detected
		$a = 0.05$: median $p = 0.007$, 98% detected
		$a = 0.15$: median $p = 0.002$, 100% detected
same, pseudo-C_ℓ	—	9%, 4%, 23% detected

Why they differ, and what the control shows. The numbers in the two columns answer different questions, and the differences have stated causes. *Planck*'s probabilities are for the real sky, so they say the sky is Gaussian; ours are for skies whose answer we know, so they say whether the test works: Gaussian skies are flagged at $8.0 \pm 2.2\%$, to be compared with the nominal 5%, and non-Gaussian ones are caught with a power that rises from 14% at $a = 0.02$ to 98% at $a = 0.05$. *Planck*'s simulations contain the real noise, its anisotropy across the sky, the component-separation residuals and lensing; ours contain white noise only, so our null distribution is narrower than theirs would be, and our power is an upper bound for a sky with real systematics. The pseudo-C_ℓ control behaves as designed: its median p stays near the middle (0.42, 0.46, 0.29) and its detection rate near the false-alarm rate, except at $a = 0.15$, where it rises to 23%. The spectra have the same mean, to within 2.5 standard errors of a 100-sky average in the worst bin; what differs is their scatter: the quadratic term adds a connected four-point function, which inflates the variance of each $\hat{C}_\ell$ (the pitfall of Section 12.23.1), and an omnibus χ^2 notices a larger scatter as well as a shifted mean.

Lesson. Rebuilding the test showed what its probabilities rest on: the null hypothesis enters only through simulations processed exactly like the data, so the mask, the beam and the noise need no formula, and the test is only as good as those simulations.

12.29 When there is no likelihood: inference from the simulator alone

Every inference so far went through a formula for $P(\boldsymbol{S} \mid \epsilon)$: Gaussian, with a mean and a covariance read off simulations. Section 12.27.3 showed what that formula costs: a covariance frozen at one value of ϵ made the intervals several per cent too narrow, and counts of rare features are not Gaussian anyway. Yet the simulator itself never assumed anything Gaussian about the counts. For any ϵ it produces maps, and the counts of those maps *are* draws from the true $P(\boldsymbol{S} \mid \epsilon)$. The likelihood is defined implicitly by the simulator, even if we cannot write it down. Can we get a posterior using only the ability to simulate?

The question. Without writing $P(\boldsymbol{S} \mid \epsilon)$, which values of ϵ could have produced the map we observed, and how probable is each, using nothing but the prior and the simulator?

12.29.1 Rejection ABC

The plainest answer is a direct imitation of Bayes' theorem by sampling, **approximate Bayesian computation** (ABC). Repeat many times: draw ϵ from the prior; simulate a map with that ϵ; compute its summary t; keep ϵ if t lands within a distance δ of the observed $t_{\rm obs}$. The tolerance δ determines which posterior this procedure approximates. A proposed ϵ has density $p(\epsilon)$, and its probability of acceptance is $\mathrm{P}(|t - t_{\rm obs}| < \delta \mid \epsilon)$. Multiplying the proposal density by that acceptance probability gives the density of the retained values up to normalisation,

$$p_{\rm kept}(\epsilon) \overset{(1)}{=} \frac{p(\epsilon)\,\mathrm{P}(|t - t_{\rm obs}| < \delta \mid \epsilon)}{\int p(\epsilon')\,\mathrm{P}(|t - t_{\rm obs}| < \delta \mid \epsilon')\,\mathrm{d}\epsilon'} \overset{(2)}{=} \frac{p(\epsilon)\int_{t_{\rm obs}-\delta}^{t_{\rm obs}+\delta} p(t \mid \epsilon)\,\mathrm{d}t}{\int p(\epsilon')\int_{t_{\rm obs}-\delta}^{t_{\rm obs}+\delta} p(t \mid \epsilon')\,\mathrm{d}t\,\mathrm{d}\epsilon'}$$
$$\overset{(3)}{\approx} \frac{p(\epsilon)\,2\delta\,p(t_{\rm obs} \mid \epsilon)}{\int p(\epsilon')\,2\delta\,p(t_{\rm obs} \mid \epsilon')\,\mathrm{d}\epsilon'} \overset{(4)}{=} p(\epsilon \mid t_{\rm obs}). \tag{12.106}$$

(1) The probability of "proposed at ϵ and kept", normalised over everything kept (Bayes' theorem with the event "kept").
(2) The probability of landing in the window is the integral of the density of t over the window.
(3) For a small window the density is nearly constant across it: the integral is its width times the value at the centre.
(4) The 2δ cancels, and what remains is Bayes' theorem for the posterior given $t_{\rm obs}$.

No likelihood was evaluated. The simulator decided, by producing or not producing a t close to $t_{\rm obs}$, how readily each ϵ could have produced the observation; this is the likelihood of Chapter 5, estimated by counting. The approximation is the finite window: as $\delta \to 0$ the answer becomes exact, and the fraction of kept simulations goes to zero.

12.29.2 Why compress first

The window in (12.106) is around one number. Why not around the whole 26-number Betti vector? Because in many dimensions "close" almost never happens. Measure the distance between a simulated vector and the observed one in units of the noise, $D^2 = (\boldsymbol{S} - \boldsymbol{S}_{\rm obs})^{\mathsf{T}}\mathbf{C}^{-1}(\boldsymbol{S} - \boldsymbol{S}_{\rm obs})$. Even a simulation with exactly the right ϵ differs from the observation by two independent fluctuations, so its $\boldsymbol{S} - \boldsymbol{S}_{\rm obs}$ has covariance $2\mathbf{C}$ and

$$\mathbb{E}[D^2] = \mathrm{Tr}\big(\mathbf{C}^{-1}\,2\mathbf{C}\big) = 2p = 52, \qquad D \approx 7, \tag{12.107}$$

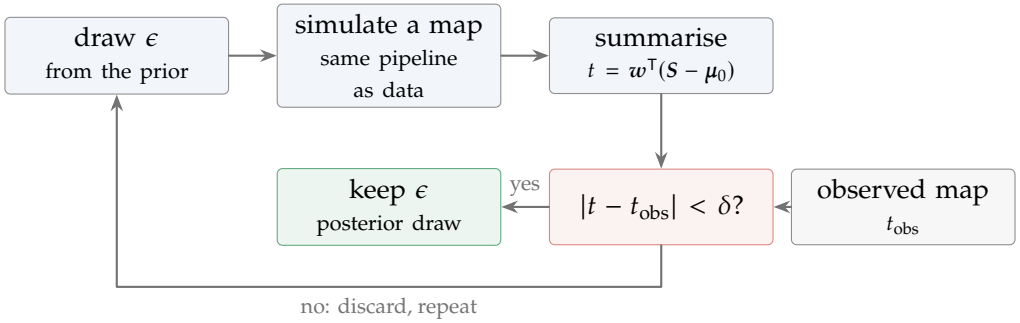


Figure 12.69: Rejection ABC. Parameters are drawn from the prior and pushed through the simulator; only those whose simulated summary lands close to the observed one survive. Surviving in proportion to how often each ϵ produces something like the data is exactly the likelihood weighting of Bayes' theorem.

while the effect of a wrong ϵ adds only $(\Delta\epsilon)^2 F$ to D^2. In 26 dimensions every simulation is far away, the right ϵ and the wrong one almost equally; keeping the nearest few per cent selects mostly on noise. This is the curse of dimensionality, and the cure is to compress the data to as few numbers as there are parameters while keeping the information.

The score as a compression. If the mean were linear, $\boldsymbol{\mu}(\epsilon) = \boldsymbol{\mu}_0 + \epsilon \boldsymbol{d}$, the Gaussian log-likelihood would be

$$\begin{aligned} \ln \mathcal{L}(\epsilon) &\overset{(1)}{=} -\tfrac{1}{2}(\boldsymbol{S} - \boldsymbol{\mu}_0 - \epsilon\boldsymbol{d})^{\mathsf{T}}\mathbf{C}^{-1}(\boldsymbol{S} - \boldsymbol{\mu}_0 - \epsilon\boldsymbol{d}) \\ &\overset{(2)}{=} -\tfrac{1}{2}(\boldsymbol{S} - \boldsymbol{\mu}_0)^{\mathsf{T}}\mathbf{C}^{-1}(\boldsymbol{S} - \boldsymbol{\mu}_0) + \epsilon\, \boldsymbol{d}^{\mathsf{T}}\mathbf{C}^{-1}(\boldsymbol{S} - \boldsymbol{\mu}_0) - \tfrac{1}{2}\epsilon^2 F, \end{aligned} \tag{12.108}$$

(1) (12.103) with the linear mean.
(2) Multiply out; the two cross terms are equal because $\mathbf{C}^{-1}$ is symmetric; $F = \boldsymbol{d}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{d}$.

The first term does not depend on ϵ and cancels in any posterior. The data enter the ϵ-dependence only through $\boldsymbol{d}^{\mathsf{T}}\mathbf{C}^{-1}(\boldsymbol{S} - \boldsymbol{\mu}_0)$, the score (12.99). Two maps with the same score have the same posterior: the score is a sufficient statistic for this model (Section 3.16.3), and dividing it by F turns it into an estimate of ϵ,

$$t = \frac{\boldsymbol{d}^{\mathsf{T}}\mathbf{C}^{-1}(\boldsymbol{S} - \boldsymbol{\mu}_0)}{\boldsymbol{d}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{d}}. \tag{12.109}$$

For a model that is only approximately linear and Gaussian, t is no longer exactly sufficient, but it keeps nearly all the Fisher information about ϵ near the fiducial point. This approximate score compression is how simulation-based analyses in cosmology reduce long data vectors to one number per parameter [274]. Note what the Gaussian model is used for here: only to choose the weights $\boldsymbol{w} = \mathbf{C}^{-1}\boldsymbol{d}/F$. If the weights are imperfect we lose some information, but the posterior is still correct for the t we use, because ABC takes the distribution of t from the simulator.

12.29.3 ABC on one map

The script `code/ch12/c07_abc.py` applies this to the same mock map as Section 12.27.2 (seed 250, $\epsilon_{\mathrm{true}} = 0.05$). It draws 20000 new maps with ϵ uniform on $(-0.15, 0.15)$:

Listing 12.23: One ABC simulation: prior draw, map, summary (`code/ch12/c07_abc.py`)

```
def one(i):
```

```
    r = rng_for("ch12", "c07_abc", i)
    eps = r.uniform(-PRIOR, PRIOR)                    # 1. draw a parameter from the
    prior
    return eps, betti_vector(model.draw(eps, r))      # 2. simulate a map, 3. summarise
    it
```

and then keeps the 2% whose compressed summary is nearest to the observed one:

Listing 12.24: Compression weights, the observed summary, and rejection on t and on all 26 numbers (`code/ch12/c07_abc.py`)

```
w = Ci @ d / F                                        # t = w . (S - mu0) estimates eps

S_obs = np.concatenate([z["grid_b0"][MOCK, eg.index(TRUE)], z["grid_b1"][MOCK,
eg.index(TRUE)]])[ok]
t_obs = (S_obs - mu0) @ w

eps, S = simulate()
S = S[:, ok]
t = (S - mu0) @ w
nkeep = int(KEEP * len(eps))
acc1 = np.argsort(np.abs(t - t_obs))[:nkeep]          # 4. keep the closest: compressed
tol1 = np.abs(t - t_obs)[acc1].max()
D = S - S_obs
dist = np.sqrt(np.einsum("ij,jk,ik->i", D, Ci, D))
acc2 = np.argsort(dist)[:nkeep]                       # 4'. keep the closest: all p
numbers
e1, e2 = eps[acc1], eps[acc2]
```

Line by line: the weights `w` are $\mathbf{C}^{-1}\boldsymbol{d}/F$ from the null maps and the derivative of Section 12.26.1; `t_obs` is the observed map's compressed summary, 0.0338; the 400 simulations with t nearest to it are kept, which sets the window at $\delta = 0.0028$; for comparison the 400 simulations nearest in the full 26-dimensional distance D are kept as well.

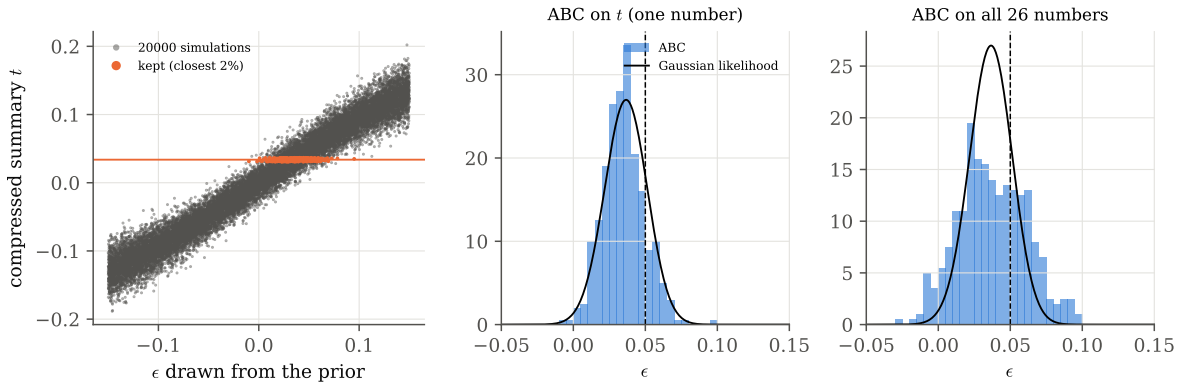


Figure 12.70: Left: the compressed summary t of 20000 maps against the ϵ they were made with. The band is the likelihood seen sideways: at each ϵ, the spread of t is its sampling distribution. The horizontal line is the observed t; the orange points, the simulations nearest to it, are the ABC posterior draws. Middle: their histogram against the Gaussian-likelihood posterior of Section 12.27.2. Right: the same with the window around all 26 numbers: much broader.

The result. ABC on t gives $\epsilon = 0.0351 \pm 0.0144$. ABC on all 26 numbers gives 0.0377 ± 0.0236, and its kept simulations are at distances up to $D = 4.9$, not far below the $D \approx 7$ that (12.107) predicts even for the right ϵ: the selection is mostly noise, as expected. The Gaussian likelihood gave 0.0364 ± 0.0148.

Is ABC more honest? On this one map ABC on t comes out a little narrower than the Gaussian likelihood (0.0144 against 0.0148). One map cannot settle the question. The width that 400 kept simulations give is itself uncertain by a few per cent, and the width of either posterior changes from map to map. But the ABC simulations do not depend on which map was observed, so the same 20000 of them serve all 250 mock maps of Section 12.27.3, at no extra cost. Over those maps the typical (median) ABC width is 0.0166, against 0.0140 for the Gaussian likelihood. The ABC interval contains the true ϵ in 65 ± 3% of the maps; the Gaussian interval does so in 62%, and both should do so in 68%.

Where does the extra width come from? The left panel of Figure 12.70 shows it. A straight-line fit of t against ϵ has slope 0.92 (an exactly linear model would give 1), and the points scatter around the line by 0.0163. The scatter grows with ϵ, from 0.0136 near $\epsilon = 0$ to 0.0143 near 0.05. A posterior from t alone is about as wide as the scatter divided by the slope, $0.0143/0.92 \approx 0.016$ near the truth. The window adds little: $\delta = 0.0028$ adds $\delta^2/3$ to the variance, about one per cent. The Gaussian likelihood instead assumed the $\epsilon = 0$ scatter everywhere, which gives 0.0140. ABC uses the scatter the simulator really produces at each ϵ, so its intervals are wider and contain the truth about as often as they should. That is the width the coverage test asked for.

What it cost. ABC used 20000 simulations to give one posterior, against about 5000 for the Gaussian-likelihood analysis, which then serves any number of maps. Rejection wastes 98% of its simulations, and shrinking δ wastes more. Modern simulation-based inference keeps every simulation: it fits a flexible model of the density $p(t \mid \epsilon)$ or $p(\epsilon \mid t)$ (a neural density estimator) to the simulated pairs, and chooses where to simulate next by where the posterior is [274]. Akeret et al. built the same rejection-and-refinement loop for cosmological forward models [275]. The logic is (12.106) in every case: the simulator defines the likelihood, and the data are compared with it through a few well-chosen numbers.

Key idea: The simulator is the likelihood

If we can simulate the data for any parameter value, we never need to write $P(\boldsymbol{S} \mid \theta)$: keeping parameter draws whose simulations resemble the data samples the posterior (12.106). "Resemble" must be measured on a few informative numbers, such as the score-compressed summary (12.109), because in many dimensions nothing resembles anything.

12.30 Reproducing the literature: topology and $f_{\rm NL}$

Two papers made the claim tested above, that topological summaries see primordial non-Gaussianity that the power spectrum does not, and turned it into numbers. Cole and Shiu forecast $f_{\rm NL}$ from persistence diagrams of CMB maps [271]; Biagetti, Cole and Shiu detected it in the topology of simulated dark-matter haloes [116]. A third, Heydenreich and collaborators, did the same for cosmological parameters from weak-lensing mass maps [56], and was reproduced in Section 12.20. We redo the key result of each of the first two on our maps, with the machinery of the previous sections, and see what survives the change of setting.

12.30.1 Cole and Shiu (2018): are persistence diagrams worth more than Betti curves?

Their question. Betti curves count the features alive at each threshold; a persistence diagram also records which birth goes with which death. The diagram therefore contains the Betti curves ((12.41)) and possibly more. Is the extra information about local non-Gaussianity worth having? In 2017 nobody had tried persistence diagrams on the CMB.

Their setup. They used public simulations of CMB temperature with local non-Gaussianity: for each of 1000 realisations, a Gaussian set of $a_{\ell m}$ and a separate non-Gaussian part, combined as $a_{\ell m} = a^{\mathrm{G}}_{\ell m} + f_{\mathrm{NL}} a^{\mathrm{NG}}_{\ell m}$ so that any f_{NL} can be dialled in. From each sky they cut two square patches of 1024×1024 pixels around the poles, and computed the persistence diagrams of the sublevel filtration (lowering the threshold from the coldest pixel up) and the Betti curves. The diagrams were binned into two-dimensional histograms (birth $\times$ persistence), and the Betti curves divided by their maximum.

resolution	$N_{\mathrm{side}} = 512$, $\ell_{\max} = 1024$, Gaussian smoothing $\theta_s = 90'$ (FWHM $212'$)
maps	2000 patches of 1024^2 pixels per value of $f_{\mathrm{NL}} \in \{0, 10, 25, 50, 100\}$
mean $\boldsymbol{\mu}(f_{\mathrm{NL}})$	average of 800 patches at $f_{\mathrm{NL}} = 0, 10, 50$
covariance	2000 patches at $f_{\mathrm{NL}} = 0$, with the Hartlap factor
statistics	β_0, β_1, $\beta_0 + \beta_1$; histograms PD0, PD1, PD0+PD1

Their key equation. Their likelihood (their eq. 4.2) is our (12.103), $\ln \mathcal{L}(f_{\mathrm{NL}}) = -\tfrac{1}{2}(\boldsymbol{d} - \boldsymbol{\mu}(f_{\mathrm{NL}}))^{\mathsf{T}} \mathbf{C}^{-1} (\boldsymbol{d} - \boldsymbol{\mu}(f_{\mathrm{NL}}))$, evaluated for "data" $\boldsymbol{d}$ equal to the simulated mean at $f_{\mathrm{NL}} = 0$, and fitted with a Gaussian in f_{NL} whose width is the quoted error. With noise-free data at the fiducial point and a mean that is linear in the parameter, $\boldsymbol{\mu}(f) = \boldsymbol{\mu}_0 + f\boldsymbol{\mu}'$,

$$\begin{aligned} \ln \mathcal{L}(f) &\overset{(1)}{=} -\tfrac{1}{2}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_0 - f\boldsymbol{\mu}')^{\mathsf{T}} \mathbf{C}^{-1} (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_0 - f\boldsymbol{\mu}') \overset{(2)}{=} -\tfrac{1}{2} f^2 \, \boldsymbol{\mu}'^{\mathsf{T}} \mathbf{C}^{-1} \boldsymbol{\mu}' \\ &\overset{(3)}{=} -\frac{f^2}{2\sigma^2}, \qquad \sigma = \left(\boldsymbol{\mu}'^{\mathsf{T}} \mathbf{C}^{-1} \boldsymbol{\mu}'\right)^{-1/2}, \end{aligned} \tag{12.110}$$

(1) Insert $\boldsymbol{d} = \boldsymbol{\mu}_0$ and the linear mean.
(2) The constant vectors cancel; f comes out of both factors.
(3) A Gaussian in f centred on zero, with variance the inverse of the quadratic coefficient.

The fitted width is the Fisher error (12.100). Symbol map: their f_{NL} is our ϵ up to a constant factor (the variance of their potential against our unit-variance g), so absolute errors cannot be compared, but *ratios* of errors between statistics can. We compare every error with that of $\beta_0 + \beta_1$.

What we compute. `code/ch12/c08_repro.py` evaluates (12.100) for the same six statistics on our 256^2 maps, with derivatives from shared seeds at $\epsilon = \pm 0.05$ and the covariance from the 1000 null maps:

Listing 12.25: Fisher error for one statistic (optionally with each curve divided by its maximum) (`code/ch12/c08_repro.py`)

```
def vec(prefix, keys, norm=False):
    parts = []
    for k in keys:
        a = z[prefix + k].astype(float)
        if norm:                                          # divide each curve by its own maximum
            a = a / np.maximum(a.max(-1, keepdims=True), 1)
        parts.append(a)
    return np.concatenate(parts, axis=-1)


def sigma_fisher(keys, norm=False):
    N = vec("null_", keys, norm)
    ok = N.std(0) > 1e-9
    N = N[:, ok]
    n, p = N.shape
```

```
    Ci = np.linalg.inv(np.atleast_2d(np.cov(N.T))) * (n - p - 2) / (n - 1)
    G = vec("grid_", keys, norm)[..., ok]
    d = ((G[:, eg.index(0.05)] - G[:, eg.index(-0.05)]) / 0.1).mean(0)
    return 1 / np.sqrt(d @ Ci @ d)
```

It also repeats the Betti rows with every curve divided by its own maximum, as they did.

Side by side.

statistic	Cole & Shiu $\Delta f_{\rm NL}$	ratio	ours $\sigma(\epsilon)$	ratio
β_0	67.4	1.11	0.0245	1.76
β_1	66.1	1.09	0.0271	1.95
$\beta_0+\beta_1$	60.6	1.00	0.0139	1.00
PD0	39.1	0.65	0.0229	1.65
PD1	37.4	0.62	0.0291	2.10
PD0+PD1	35.8	0.59	0.0133	0.96

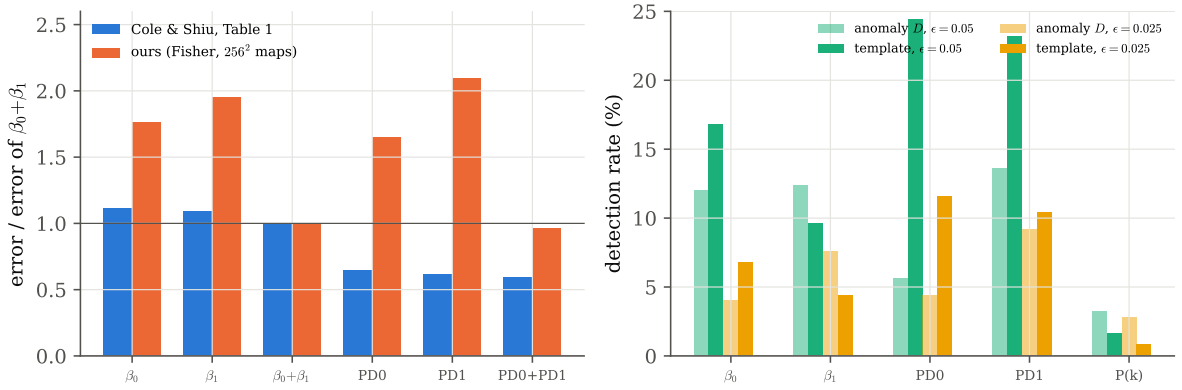


Figure 12.71: Left: errors on the non-Gaussianity amplitude relative to that of $\beta_0+\beta_1$, from Cole and Shiu's Table 1 (blue) and from our Fisher matrices (orange). They find the diagrams 40% better than the curves; on our maps the two are equal, while each curve alone is about twice as bad as the pair. Right: detection rates of $\epsilon = 0.05$ and 0.025 with Biagetti et al.'s anomaly distance (pale) and template statistic (dark), for four topological summaries and the power spectrum.

Why they differ. Two results disagree. Their diagrams beat their Betti curves by a factor 0.59; ours tie. And their β_0 and β_1 each nearly match the pair, while ours are each about twice as bad. Go through the differences in setup one at a time.

The normalisation of the Betti curves. Dividing each curve by its maximum throws away its overall amplitude. On our maps it costs something: $\beta_0+\beta_1$ normalised gives 0.0167 instead of 0.0139, so that the diagrams measured against normalised curves would come out at $0.0133/0.0167 \approx 0.80$ instead of 1. Part of their gap, not all of it.

Where the signal lives. In our model the quadratic term is applied to the field we observe, smoothed once; most of what ϵ does is move pixel values, a one-point effect (Section 12.23.2), and it shows up as more pieces in the hot tail and fewer lakes in the cold tail, two nearly independent witnesses (Section 12.25.3). Betti curves sampled out to $\pm 3\sigma$ capture that completely, and β_0 and β_1 see different tails, so each alone loses half. In their CMB maps the quadratic term acts on the primordial potential and passes through the acoustic transfer function, which mixes neighbouring points (Section T3.2.3); less of the signal is a simple shift of values, and more of it is in how features are arranged. If that is the reason, removing the one-point part from our

maps should restore their result. It does: at area-fraction thresholds, where the one-point effect is gone by construction, our persistence histograms give $\sigma(\epsilon) = 0.104$ against 0.169 for the Betti curves, a ratio of $0.104/0.169 \approx 0.62$, against their 0.59. We cannot prove from their paper that this is the whole explanation, but it is the one we can test, and it passes.

Smaller differences. Their histograms are finer than our 10×5 bins; their patches have edges, which add pieces along the boundary (the reason they normalised); their maps hold far more correlation areas than our 256^2 pixels. None of these changes the ratios we compare by as much as the two effects above.

Lesson. "Persistence diagrams carry more information than Betti curves" is not a theorem; it depends on where in the field the signal lives. When the signal is a shift of values, a good Betti curve already has it; when it is in the arrangement, birth–death pairs add information, and on our maps they add it in the same proportion as on Cole and Shiu's.

12.30.2 Biagetti, Cole and Shiu (2021): detecting f_{NL} without a likelihood

Their question. Can the topology of the dark-matter halo distribution reveal a local f_{NL} as small as 10, through the structures it affects at scales of about $10\,h^{-1}$Mpc, where the usual signature in the halo power spectrum (a bias that grows on large scales) is negligible? And how can one attach a significance to such a detection without writing a likelihood for persistence summaries?

Their setup. N-body simulations with Gaussian initial conditions ("G85L") and with $f_{\mathrm{NL}} = 10$ ("NG10L"), in boxes whose five realisations add up to $40\,(h^{-1}\mathrm{Gpc})^3$, about a Euclid-like volume; persistence diagrams of the halo catalogues with a distance-to-measure filtration; summaries built from the diagrams (distributions of births B_p, deaths D_p, persistences, and Betti curves). Significance comes from splitting the realisations into a "survey" and a "fiducial" set in all possible ways: Gaussian against Gaussian splits give the null distribution, Gaussian against non-Gaussian splits give the detection rate. Two statistics are used.

Their equations, derived. The *anomaly* distance (their eq. 4.1) is

$$D_{\mathrm{anomaly}} = \Big[\frac{1}{N}\sum_{i=1}^{N}\frac{(d_i-\mu_i)^2}{\sigma_i^2}\Big]^{1/2}, \tag{12.111}$$

with μ_i, σ_i the mean and standard deviation of entry i under the null: the χ^2 of (12.93) with the covariance replaced by its diagonal, divided by N, square-rooted. It is an omnibus test, and it ignores correlations. A detection is a D above the 95th percentile of the Gaussian splits. The *template* (their eqs. 4.2–4.3) is the mean difference between a strongly non-Gaussian ($f_{\mathrm{NL}} = 250$) and a Gaussian simulation *with the same initial seed*, $\boldsymbol{T} = \overline{\boldsymbol{S}_{\mathrm{NG}} - \boldsymbol{S}_{\mathrm{G}}}$, and the statistic is the projection of the data on it, $D_{\mathrm{template}} = (\boldsymbol{S}-\boldsymbol{\mu})\cdot\boldsymbol{T}/\sigma_T$, with σ_T its standard deviation under the null; a detection is a value above the 97.5th percentile. What is the template, in our language? To first order the shared-seed difference is $\boldsymbol{T} \approx f_{\mathrm{big}}\,\boldsymbol{d}$, the derivative of Section 12.26.1 times the large amplitude, so

$$D_{\mathrm{template}} \overset{(1)}{=} \frac{f_{\mathrm{big}}\,\boldsymbol{d}^{\mathsf{T}}(\boldsymbol{S}-\boldsymbol{\mu})}{\sqrt{\mathrm{Var}[f_{\mathrm{big}}\,\boldsymbol{d}^{\mathsf{T}}(\boldsymbol{S}-\boldsymbol{\mu})]}} \overset{(2)}{=} \frac{\boldsymbol{d}^{\mathsf{T}}(\boldsymbol{S}-\boldsymbol{\mu})}{\sqrt{\boldsymbol{d}^{\mathsf{T}}\mathbf{C}\boldsymbol{d}}}, \tag{12.112}$$

(1) $\boldsymbol{T} \approx f_{\mathrm{big}}\boldsymbol{d}$, and σ_T is the null standard deviation of the projection.

(2) $f_{\rm big}$ cancels; the variance of $\boldsymbol{a}^{\mathsf{T}}\boldsymbol{x}$ is $\boldsymbol{a}^{\mathsf{T}}\mathbf{C}\boldsymbol{a}$.

This is the score (12.99) with $\mathbf{C}^{-1}$ replaced by the identity: it aims at the right direction but does not weight the entries by their noise. How much does that cost? For any weights $\boldsymbol{w}$, the statistic $\boldsymbol{w}^{\mathsf{T}}(\boldsymbol{S}-\boldsymbol{\mu})$ has, at a small ϵ, mean $\epsilon\,\boldsymbol{w}^{\mathsf{T}}\boldsymbol{d}$ and standard deviation $\sqrt{\boldsymbol{w}^{\mathsf{T}}\mathbf{C}\boldsymbol{w}}$. Write $\boldsymbol{a}=\mathbf{C}^{1/2}\boldsymbol{w}$ and $\boldsymbol{b}=\mathbf{C}^{-1/2}\boldsymbol{d}$; then

$$\frac{\boldsymbol{w}^{\mathsf{T}}\boldsymbol{d}}{\sqrt{\boldsymbol{w}^{\mathsf{T}}\mathbf{C}\boldsymbol{w}}} \overset{(1)}{=} \frac{\boldsymbol{a}^{\mathsf{T}}\boldsymbol{b}}{|\boldsymbol{a}|} \overset{(2)}{\le} |\boldsymbol{b}| \overset{(3)}{=} \sqrt{\boldsymbol{d}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{d}} = \sqrt{F}, \tag{12.113}$$

(1) $\boldsymbol{w}^{\mathsf{T}}\boldsymbol{d}=\boldsymbol{w}^{\mathsf{T}}\mathbf{C}^{1/2}\mathbf{C}^{-1/2}\boldsymbol{d}=\boldsymbol{a}^{\mathsf{T}}\boldsymbol{b}$ and $\boldsymbol{w}^{\mathsf{T}}\mathbf{C}\boldsymbol{w}=|\boldsymbol{a}|^2$.
(2) Cauchy–Schwarz, with equality when $\boldsymbol{a}\propto\boldsymbol{b}$, that is $\boldsymbol{w}\propto\mathbf{C}^{-1}\boldsymbol{d}$: the score.
(3) $|\boldsymbol{b}|^2=\boldsymbol{d}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{d}$.

The score is the best linear statistic; the template reaches the bound only if $\boldsymbol{d}$ happens to be an eigenvector of $\mathbf{C}$. Figure 12.72 shows the geometry.

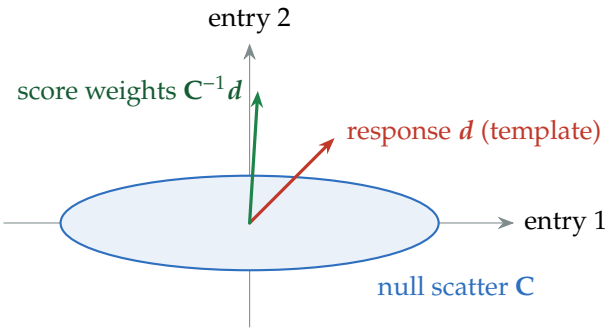


Figure 12.72: Why whitening matters. Two entries of a data vector, the first very noisy (long axis of the null ellipse) and the second quiet. A small ϵ moves the mean along $\boldsymbol{d}$ (red), equally in both. Projecting on $\boldsymbol{d}$ mixes in the large noise of entry 1; the score weights $\mathbf{C}^{-1}\boldsymbol{d}$ (green) lean on the quiet entry. For the ellipse drawn, with standard deviations 2 and 0.5, the signal-to-noise per unit ϵ is $2/\sqrt{4.25}=0.97$ along $\boldsymbol{d}$ and $\sqrt{4.25}=2.06$ along $\mathbf{C}^{-1}\boldsymbol{d}$.

What we compute. The same two statistics on our maps, in `c08_repro.py`: null mean and standard deviations from the 1000 null maps, the null distributions of $D_{\rm anomaly}$ and $D_{\rm template}$ from the 500 check maps, the template from the shared-seed differences between $\epsilon=0.1$ and $\epsilon=0$ for grid seeds 0–249 (twice the tested amplitude, where theirs was twenty-five times), and detection rates for seeds 250–499 at $\epsilon=0.05$ and 0.025, for β_0, β_1, PD0, PD1 and the power spectrum.

Side by side. Biagetti et al. detect $f_{\rm NL}=10$ in about 85% of survey volumes with the template on B_1 (births of loops) and 74% on D_1 (deaths of loops), at the 97.5% level, and several summaries reach detection rates above 60% with the anomaly distance at the 95% level; the halo power spectrum, on large scales, does about as well as the topology. On our maps, at $\epsilon=0.05$:

	β_0	β_1	PD0	PD1	$P(k)$
anomaly D, 95% (%)	12	12	6	14	3
template, 97.5% (%)	17	10	24	23	2

and at $\epsilon=0.025$ the template rates are 7, 4, 12, 10 and 1%.

Why they differ. Each rate comes from 250 maps, so a rate near 15% is uncertain by $\sqrt{0.15 \times 0.85/250} \approx 2$ percentage points, and both statistics are judged against the same 500 check maps. Within that precision the qualitative pattern is theirs: where the two statistics differ by several such errors, the template beats the anomaly distance, because it aims (Section 12.25.3); the anomaly distance, an unweighted omnibus statistic, is the weakest. The power spectrum is where our setting departs on purpose: our maps have the same spectrum by construction, so it detects nothing beyond the false-alarm rate, while in a halo catalogue local $f_{\rm NL}$ changes the large-scale bias and the spectrum does respond. The absolute rates differ because the experiments differ in size, one 256^2 map against $40\,(h^{-1}\mathrm{Gpc})^3$ of haloes. The rates are also low compared with what our data can do, and (12.113) says why: the template is unwhitened. For β_0 alone the Fisher error is 0.0245, so the score at $\epsilon = 0.05$ is shifted by $0.05/0.0245 \approx 2.0$ standard deviations, and a one-sided test at 97.5% (threshold 1.96) would detect it about half the time; the template detects 17%. Our β_0 curve has entries with very different noise, large counts near the median and small, quiet counts in the tail where the signal is, exactly the situation of Figure 12.72.

Lesson. Their two statistics are the omnibus test and an unwhitened score. Reproducing them showed that the step from anomaly to template, aiming at the expected signal, is what raises the detection rate, and that the next step, weighting by the inverse covariance, is worth as much again whenever the entries of the summary are not equally noisy.

Heydenreich et al. (2021), recalled. The weak-lensing analogue was reproduced in Section 12.20: persistent Betti numbers of aperture-mass maps, with a simulated covariance and an emulated mean, constrain S_8 better than peak counts, by 44% in figure of merit in their KV450-like setup and 18% in error for a Euclid-like survey [56]. It is the same pipeline as for ϵ, with a cosmological parameter in place of ϵ and a non-Gaussian field produced by gravity rather than by inflation.

12.31 How it is done in research

Every step from the nine-pixel grid to the ABC posterior was a scaled-down version of an analysis that is run on real data. The steps are the same; the size of the maps, the realism of the simulations, the number of parameters and the care taken at each step are not. Figure 12.73 lays out the pipeline as a real analysis runs it, for a CMB Gaussianity test or an $f_{\rm NL}$ forecast from topology, and for its large-scale-structure and lensing relatives.

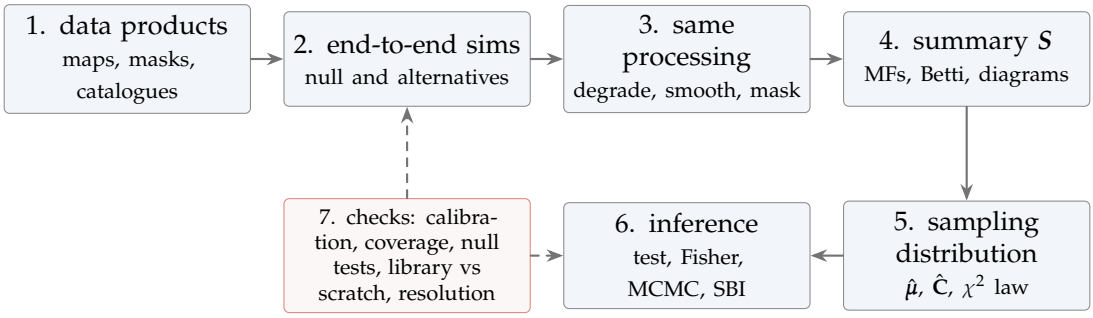


Figure 12.73: The pipeline of a simulation-based topological analysis. The data and every simulation pass through the same processing and the same summary; the simulations then supply the mean, the covariance and the distribution of the test statistic, and every inference is checked on simulations whose answer is known before it is applied to the sky.

1. Data products. For the CMB: the component-separated temperature and E-mode maps of *Planck* at $N_{\text{side}} = 2048$, the common mask (about 75% of the sky kept) and the noise and beam models [114]. For large-scale structure: halo or galaxy catalogues, or N-body snapshots [116]; for lensing, shear catalogues and mass maps (Section 12.21). *Ours:* no data; flat 256^2 maps and full skies at $N_{\text{side}} = 128$. *Cost:* everything here is a demonstration on simulations; the real sky is analysed in Chapter 13.

2. Simulations. *Planck* uses 999 FFP10 end-to-end simulations, which contain the lensed CMB, the instrument noise with its anisotropy and correlations, the scanning and the component-separation pipeline [114]. Non-Gaussian alternatives are made by adding a separately computed non-Gaussian part to Gaussian $a_{\ell m}$, so that any f_{NL} can be dialled in, with the transfer functions of the full physics [271]. *Ours:* Gaussian fields with white noise; a quadratic term applied pointwise and then smoothed, renormalised to the same spectrum. *Cost:* our null distributions are narrower than real ones (no correlated noise, no foreground residuals), so our detection rates are upper bounds, and our non-Gaussian signal is mostly a shift of values, which favours Betti curves over persistence diagrams (Section 12.30.1).

3. Processing. The real maps are degraded to several resolutions, smoothed with a beam that grows with the pixel size, and masked, and every simulation is processed in exactly the same way; analyses at several resolutions or needlet scales probe several physical scales. *Ours:* one resolution and one beam per study. *Cost:* we test one scale at a time, and do not face the look-elsewhere effect that many resolutions and statistics create (Section 4.14).

4. Summaries. Area, boundary length and genus at a dozen thresholds for *Planck*; Betti curves and binned persistence diagrams for the topology papers, computed with gudhi, PHAT, ripser or dedicated cubical codes. *Ours:* our own union–find and cell counts, checked against gudhi and scikit-image (Sections 12.22.1 and 12.28.1). *Cost:* none in principle; the check against a library is the step a real analysis must also do once.

5. The sampling distribution. Mean and covariance from hundreds to thousands of simulations, the Hartlap factor or the Sellentin–Heavens t-likelihood for the noise in the covariance [6, 272], and the distribution of the test statistic from simulations rather than from tables. *Ours:* the same, with 1000 maps for the covariance and 500 for the reference distribution. *Cost:* with $p = 26$ and $n = 1000$ the noise in the covariance is small (Hartlap factor 0.973); with longer summaries, 98-bin persistence histograms, it is not, and Section 12.25.2 showed that the table then fails.

6. Inference. For a Gaussianity test, a χ^2 and a simulated probability. For f_{NL} itself, *Planck* uses the bispectrum, the optimal statistic for a known three-point shape, and keeps the Minkowski functionals as a model-independent check (Section T3.5.2). Forecasts from topology use Fisher matrices from simulated derivatives; parameter inference with several cosmological and nuisance parameters uses an emulator of $\boldsymbol{\mu}(\boldsymbol{\theta})$ over a few dozen simulated cosmologies and an MCMC [56], or simulation-based inference with neural density estimators [274]. *Ours:* one parameter, finite differences at seven values, a grid posterior checked by Metropolis and emcee, and rejection ABC. *Cost:* no degeneracies with other parameters (Section 12.20.2 showed how much they matter), and no nuisance parameters to marginalise.

7. Checks. The checks we ran are the ones a real analysis runs. *Calibration*: p-values of null simulations must be uniform (Section 12.25.2). *Coverage*: intervals must contain the truth at the

stated rate over mock data sets (Section 12.27.3); ours did not at $\epsilon = 0.05$, and the diagnosis, a covariance fixed at the wrong parameter, is a real and common failure. *Null tests*: statistics that should see nothing must see nothing, as the power spectrum did here. *Implementation*: a from-scratch code against a library, and a convergence test in resolution and in the number of simulations.

How this is used in real research: how the tests are run on the real sky

Planck tested the Gaussianity and isotropy of the temperature and polarization maps with the three Minkowski functionals at resolutions from $N_{\rm side} = 16$ to 2048, using 999 simulations for the mean, the covariance and the distribution of χ^2; the temperature maps were consistent with Gaussian skies, with probabilities from 24% to 99.8%, and the bispectrum bounded the local $f_{\rm NL}$ to -0.9 ± 5.1 [112, 114] (Section T3.5). Persistent homology has entered the same pipelines as a summary: Betti curves and persistence diagrams for $f_{\rm NL}$ in CMB simulations [271] and in halo catalogues [116], and persistent Betti numbers of weak-lensing mass maps, which tightened S_8 beyond peak counts in survey-like analyses [56] (Section 12.20). In every case the topology is the cheap part, seconds per map; the simulations, processed exactly like the data, are the expensive part and the scientific core.

Each reduction, what the full analysis does instead, and what the reduction costs:

reduction	full research version	what the reduction costs
map size	full sky at $N_{\rm side} = 2048$; 1024^2-pixel patches	our 256^2 maps hold fewer correlation areas: errors on ϵ larger by roughly the square root of the area ratio
sphere resolution	$N_{\rm side} = 2048$, beams down to 5′	at $N_{\rm side} = 128$ the 40′ beam is barely resolved and counts fall short of the formula (Section 12.28.2)
non-Gaussian model	local $f_{\rm NL}$ through the radiation transfer functions	our signal is mostly one-point; amplitudes are ∼100 times the *Planck* bound
noise	anisotropic, correlated, with foreground residuals	white noise only; null scatter underestimated
simulations	999 end-to-end FFP10 skies; thousands of N-body volumes	1000 to 5000 Gaussian maps per study; p-values resolved to 1/501 or 1/401
covariance	t-likelihood or covariance at the best fit	fixed at $\epsilon = 0$; 68% intervals covered 62%
parameters	cosmology plus nuisances, emulated means	one parameter, finite differences

Recap

- A topological summary is a random variable with a sampling distribution $P(S \mid \theta)$ (12.82); on a 3×3 grid it can be enumerated ($\mathbb{E}\,\chi = p + 4q^2 - 4q^4$, $\mathbb{E}\,\beta_1 = p^4 q$), on a real map only simulated (Section 12.22).
- The local model $g + \epsilon(g^2 - 1)$, renormalised by $\sqrt{S_g/S_\phi}$ with $S_\phi = S_g + \epsilon^2\,\mathcal{F}[2\xi^2]$, has the same spectrum for every ϵ; its Euler characteristic moves as Matsubara's formula predicts, mostly through the one-point term, which area-fraction thresholds remove (Section 12.23).
- With $\hat{\boldsymbol{\mu}}$ and $\hat{\mathbf{C}}$ from n simulations, the χ^2 of a new map is $(1 + \frac{1}{n})\frac{(n-1)p}{n-p} F_{p,n-p}$: Hartlap fixes its mean, not its width; rare-feature bins give it a heavy tail (Section 12.24).

- The simulated p-value $(1 + r)/(1 + N)$ is exact by exchangeability; reference maps must be independent of those used for the covariance. Aimed (score) tests beat omnibus χ^2 tests by a factor ≈ 1.8 in detectable ϵ for $p = 22$ (Section 12.25).
- Fisher errors from shared-seed derivatives: Betti 0.0139, persistence 0.0133, skewness 0.0365, power spectrum none; a Fisher number equal to its own noise bias means no information (Section 12.26).
- Grid, Metropolis and emcee posteriors agree with the Fisher width; 68% intervals covered $62 \pm 3\%$ at $\epsilon = 0.05$, consistent with a covariance that grows with ϵ by 1.15, which predicts $2\Phi(1/\sqrt{1.15}) - 1 \approx 0.650$ (Section 12.27).
- On the sphere: beams lower λ and every count; noise adds short-lived features on the diagonal; masks cut and isolate. A *Planck*-style test with these operators is calibrated and detects same-spectrum non-Gaussian skies that pseudo-C_ℓ misses (Section 12.28).
- ABC samples the posterior from the simulator alone (12.106); it needs compression to one number per parameter. Over 250 maps its intervals are wider than the Gaussian ones and contain the truth about as often as they should (Section 12.29).
- Reproduced: Cole and Shiu's diagram-to-curve ratio 0.59 appears on our maps once the one-point signal is removed (0.62); Biagetti et al.'s template statistic beats their anomaly distance, and whitening would add as much again (Section 12.30).

Reading the literature

- **Excursion sets and the geometry of random fields.** Rice on counting crossings and maxima [264]; Adler, and Adler and Taylor, for the Gaussian kinematic formula [265, 266]; Tomita on curvature invariants [267]; Hamilton, Gott and Weinberg on the genus of large-scale structure [115]; Schmalzing and Górski on Minkowski functionals on the sphere [110]; Matsubara on weakly non-Gaussian fields [113].
- **Persistent homology.** Edelsbrunner, Letscher and Zomorodian on persistence [268]; Cohen-Steiner, Edelsbrunner and Harer on stability [270]; Otter et al. for a roadmap of algorithms and software [269]; Pranav et al. and Wilding et al. on the topology of the cosmic web [4, 5].
- **Polarization singularities and lensing.** Vachaspati and Lue, and Huterer and Vachaspati, on the singularities of CMB polarization [1, 2]; Challinor and Chon, and Lewis and Challinor, on lensing of the CMB [3, 47].
- **The CMB tests.** *Planck* 2018 on isotropy and statistics, and on primordial non-Gaussianity [112, 114]; Cole and Shiu on persistence diagrams of CMB maps [271]; Biagetti, Cole and Shiu on persistent homology and $f_{\rm NL}$ in halo catalogues [116].
- **Weak lensing.** Bartelmann and Schneider, and Kilbinger, for reviews [43, 44]; Kaiser and Squires on mass reconstruction [46]; Xavier et al. on lognormal fields [55]; Jeffrey et al. on the DES mass maps [54]; Kratochvil et al. and Dietrich and Hartlap on peak counts [52, 53]; Heydenreich et al. on persistent Betti numbers [56]; the HSC, KiDS and DES cosmic-shear analyses [45, 49, 50].
- **Inference from simulations.** Hobson et al. on covariances from simulations [HobsonBMC §3.4.3]; Hivon et al. on Monte Carlo calibration of an estimator [Hivon §1]; Hartlap, Simon and Schneider, Taylor, Joachimi and Kitching, and Sellentin and Heavens on estimated covariances [6, 165, 272]; Carron on parameter-dependent covariances [273]; Akeret et al. on approximate Bayesian computation in cosmology [275]; Alsing et al. on density-estimation likelihood-free inference and score compression [274].

Chapter summary

What should we measure in a field when the power spectrum is not enough? The answer is the shape of its excursion sets. Above a threshold a field becomes a region; its pieces and holes are counted by Betti numbers, their difference is the Euler characteristic, and on pixels the counting needs dual neighbour rules and can be done by $V - E + F$. These counts survive smooth one-to-one remappings and monotone changes of the values, which is what makes them robust; for a Gaussian field their averages follow from counting critical points and depend on the spectrum only through λ. Sweeping the threshold turns the excursion sets into a filtration, and recording when each piece and hole is born and dies gives persistence diagrams, stable under small perturbations and computable by union–find. On weak-lensing mass maps the same ideas met a field made non-Gaussian by gravity: shear converted to convergence, noise and masks modelled, and peaks, Minkowski functionals and persistent Betti numbers compared with the two-point function. Finally the summaries became random observables: sampling distributions by simulation, covariances and their noise, tests with simulated null distributions, Fisher forecasts with simulated derivatives, posteriors, coverage, and inference from the simulator alone. In the pipeline of the book,

$$\text{physical model} \to \text{field or data} \to \text{what matters} \to \text{summary } S \to \boxed{P(S \mid \theta) \to \text{physics}},$$

topological summaries now have every arrow: a reason to choose them, a construction, and a sampling distribution that turns them into tests, forecasts and posteriors. Chapter 13 applies them to the real *Planck* sky, and Chapter 14 collects what the whole book's machinery assumes and where it breaks.

Excursion sets and Minkowski functionals (part 12a)

concept	operational meaning	used later in
nuisance and protection	operator $\phi \to \mathcal{N}[\phi]$; T protected if $T[\mathcal{N}\phi] = T[\phi]$	Chapters 13 and 14
excursion set, β_0, β_1, χ	pixels above ν; pieces, holes, $\chi = \beta_0 - \beta_1 = V - E + F$	Chapter 13
dual connectivity	8 neighbours for the set, 4 for its complement	Chapter 13
area-fraction threshold	$\Psi(\nu_A)$ = measured area; removes monotone (one-point) distortions	Chapter 13
Minkowski functionals	area, boundary length, χ; the additive shape measures	Chapter 13
Gaussian kinematic formula	$\mathbb{E}\,\chi/A = \lambda \nu e^{-\nu^2/2}/(2\pi)^{3/2}$, $\lambda = \sigma_1^2/2\sigma_0^2$; sphere adds $2\Psi(\nu)$	Chapters 13 and 14
Matsubara correction	first-order shift from $S^{(0)}$, $S^{(1)}$, $S^{(2)}$; only differences survive at fixed area	Chapter 13

Persistent homology (part 12b)

concept	operational meaning	used later in
filtration, birth, death	nested excursion sets as ν falls; elder rule pairs each death with the younger feature	Chapter 13
persistence diagram, barcode	the set of (b, d) pairs; Betti curve = pairs alive at ν	Chapter 13
stability	$d_B(\mathrm{Dgm} f, \mathrm{Dgm} g) \le \|f - g\|_\infty$; noise adds points near the diagonal	Chapters 13 and 14
union–find, duality	islands by merging; lakes as islands of $-\phi$ with a sea around the map	Chapter 13
Vietoris–Rips	complex of a point cloud; its H_0 is the minimum spanning tree	Chapter 11
cosmic web	clusters, tunnels and voids as H_0, H_1, H_2 of a density field; gravity moves the diagram away from the Gaussian one	Chapter 11
polarization singularities	zeros of $Q + iU$, index $\pm\frac{1}{2}$; kept by one-to-one remappings	Chapter 14

Weak-lensing maps (part 12w)

concept	operational meaning	used later in
lognormal convergence	shifted lognormal $e^Z - \lambda$ matching the mean, variance and skewness of κ	Chapter 14
Kaiser–Squires	$\tilde{\kappa} = \tilde{\gamma}\, e^{-2i\varphi_\ell}$ in Fourier space; E and B modes; mass-sheet constant	Chapter 14
mask on the mass map	leaks E into B in band powers; cuts and isolates features in counts	Chapter 14
shape noise	$\sigma_e^2/(n\, A_{\mathrm{pix}})$ per pixel; must be smoothed before counting	Chapter 14
band powers and S_8	pseudo-C_ℓ on a masked patch; Fisher banana in (Ω_m, σ_8)	Chapter 14
beyond two points	peaks, MFs, Betti and persistent Betti numbers with simulated covariance and derivatives	Chapter 14
Fisher noise bias	$\mathrm{Tr}[\mathbf{C}^{-1}\boldsymbol{\Sigma}_d]/N_{\mathrm{der}}$; dangerous along degeneracies	Chapter 14

Topological summaries as statistics (part 12c)

concept	operational meaning	used later in
sampling distribution	$P(S \mid \theta)$ from simulations through the data's pipeline	Chapters 13 and 14
same-spectrum twins	renormalised local model; only beyond-two-point statistics separate them	Chapter 13
χ^2 with simulated $\hat{\mathbf{C}}$	Hotelling law $(1 + \frac{1}{n})\frac{(n-1)p}{n-p} F_{p,n-p}$; Hartlap fixes the mean only	Chapters 13 and 14
simulated p-value	$(1 + r)/(1 + N)$ with independent reference simulations; exact	Chapter 13
score vs omnibus	$t = \boldsymbol{d}^{\mathsf{T}}\mathbf{C}^{-1}(\boldsymbol{S} - \boldsymbol{\mu}_0)$ aims; χ^2 spreads over p directions	Chapters 13 and 14
Fisher from simulations	shared-seed central differences; compare F with its noise bias	Chapter 14
coverage test	fraction of mock intervals containing the truth; $2\Phi(1/k) - 1$ for a width off by k	Chapter 14
operators on the sky	beam lowers λ, noise adds diagonal features, mask cuts and isolates	Chapter 13
ABC and compression	keep prior draws whose simulated t is near t_{obs}; compress to one number per parameter	Chapter 14

Chapter 13

Capstone: The Analysis on Real Data

Where this chapter is going

Every method in the book so far was tried on data whose truth we had chosen. The skies of Chapter 8 were drawn from a spectrum we fixed, the parameters of Section 9.22 were recovered from a mock spectrum generated with known values, and the galaxies of Chapter 11 were points we had scattered ourselves. Because we chose the inputs, we could check how well each method recovered them. With observations we do not know the true spectrum or the true distance to a galaxy, and a hot spot may be primordial or may come from residual Galactic emission. We can still test predictions that follow from our description of the measurement. These *checks* give us ways to detect a failure even when the quantity we ultimately want is unknown.

The chapter takes three public data sets through the pipeline the book has been building,

$$\text{physical model} \to \text{field} \to \text{what matters} \to \text{summary} \to \text{its sampling distribution} \to \text{inference},$$

and each time the same three ingredients do the work: a summary computed from the data, a sampling distribution for that summary obtained by simulating the observation, and a comparison between the two. The *Planck* temperature map comes first. Its summary is the angular power spectrum, its sampling distribution comes from Gaussian skies passed through the same mask, beam and noise, and the inference is a fit of two cosmological parameters and a set of consistency tests. The same map comes back in Section 13.9, where the summary is the shape of its hot and cold regions (Minkowski functionals, Betti numbers, persistence) and the inference is a test of whether the sky is Gaussian. The last data set is a catalogue of galaxies from the BOSS survey (Section 13.15). Its summary is the pair correlation function, its sampling distribution comes from the data themselves (the jackknife), and the inference is a distance to the galaxies, read off the acoustic peak (Figure 13.1).

13.1 One sky, and the questions we put to it

In Section 9.22 we fitted cosmological parameters to a spectrum that looked like *Planck*'s but was not. At every multipole we drew a number from the law that a masked sky's spectrum follows, a scaled χ^2 with $(2\ell+1)f_{\mathrm{sky}}$ degrees of freedom around the fiducial C_ℓ plus the noise of the 143 GHz channel, and then asked the sampler to find the parameters again. It did, and the widths of the posteriors came out close to the published ones. That exercise tested the likelihood and the sampler. It said nothing about the step before them: going from a map of the sky to a spectrum. Here we take that step on the real sky. The tools are those of Chapters 7 to 9, imported unchanged, and the input is a map that *Planck* measured.

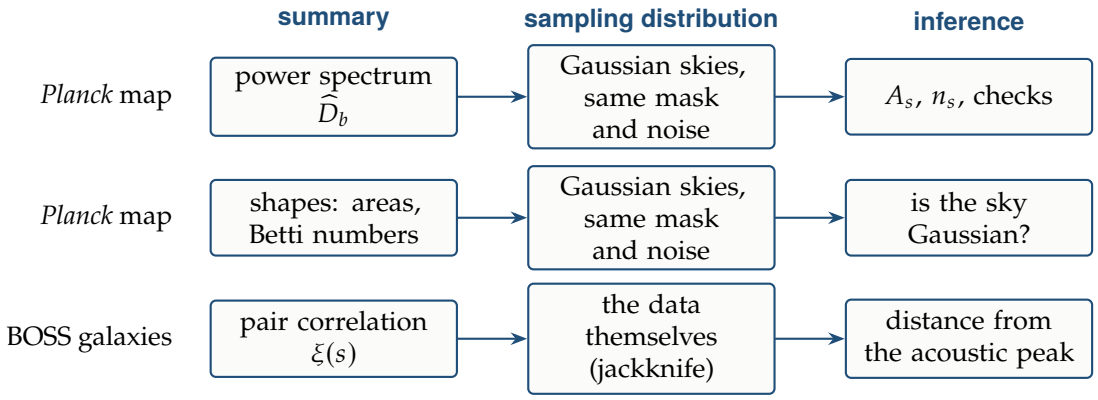


Figure 13.1: Three real data sets, one pipeline. Each row computes a summary from the data, finds how that summary would scatter if our description of the data were right, and draws a conclusion by comparing the two. The first row is the temperature power spectrum and two primordial parameters; the second row puts the same map's hot and cold regions to a Gaussianity test; the third measures a distance from the clustering of galaxies.

The event. The data are the **SMICA** temperature map of the 2018 *Planck* release [186]. It is a map of the CMB alone: the nine frequency maps of the satellite were combined, multipole by multipole, with weights chosen so that emission with the spectrum of the CMB passes and emission from our Galaxy and from other galaxies is suppressed. It comes at HEALPix resolution $N_{\text{side}} = 2048$ [$12N_{\text{side}}^2 \approx 5 \times 10^7$ pixels of 1.7′] (Section 7.11.2, [7]), smoothed by an effective beam that is a 5′ Gaussian. It also comes as two **half-mission maps**, made from the first and the second half of the observing time: the same sky, with independent noise. Our event is this map, and the summary we compute from it is its power spectrum, binned exactly as *Planck* bins its own.

What could have produced it? Before computing anything, we list the situations that could have made the map, one at a time, each with its reason.

1. *A Gaussian, statistically isotropic CMB with a spectrum $C_\ell(\boldsymbol{\theta})$, seen through the beam and the pixels, plus detector noise.* This is the standard model's prediction (Section 7.9). If it is the whole story, the spectrum carries all the information in the map, and the parameters $\boldsymbol{\theta}$ are what we want.
2. *Residual emission from our Galaxy or from other galaxies.* Component separation is not perfect. Dust and synchrotron residuals concentrate near the Galactic plane. Unresolved radio and infrared galaxies add power that grows with ℓ.
3. *Something the processing did.* The beam may not be exactly the one we assume, the noise of the two halves may be slightly correlated, and the filling of the masked regions may leak into the kept ones.

Only the first explanation comes with a sampling distribution we can produce: we can draw as many Gaussian skies as we like and observe each of them the way *Planck* observed the real one. For the second and third we cannot simulate the effect without saying how large it is, which we do not know. But we know where each would show up. Galactic residuals would make the spectrum depend on how much sky near the plane we keep. A mismatch between the two halves would show up in their difference. So the second and third candidates are tested by *checks*: quantities that the first candidate predicts and the other two would spoil.

The questions, in words.

(a) What is the temperature power spectrum of this map, in *Planck*'s bins, once everything the observation did to it (mask, beam, pixels, noise) has been undone on average?

(b) If the sky were a Gaussian realisation of the *Planck* best-fit spectrum, how much would those bandpowers scatter, and how would their errors be correlated?

(c) Is the best-fit spectrum a reasonable explanation of our bandpowers, and do they agree with the spectrum *Planck* published from the same sky?

(d) With the other four parameters held fixed, which amplitude A_s and tilt n_s of the primordial spectrum does the map prefer?

(e) Does anything in the map fail a test that a Gaussian CMB would pass?

The inputs, and what would break them. Every answer below rests on a short list of assumptions, and it helps to separate them. Two are *physical*. (i) Outside the mask, the map is a Gaussian isotropic CMB plus noise; residual foregrounds would break this, and the checks of Section 13.6 would see it as a spectrum that depends on where we look. (ii) The noise in the two half-mission maps is independent; correlated noise would bias the cross-spectrum upward at high ℓ, and the spectrum of the half-difference would disagree with its simulations. Two are *observational*: (iii) the beam is the one *Planck* supplies, and (iv) the mask removes every region we do not trust. A wrong beam would tilt the spectrum at high ℓ, which a fit would turn into a wrong n_s; a mask too small would let Galactic power in. One is *mathematical*: (v) the map is band-limited, with no power above $\ell = 3N_{\rm side} - 1$, the highest multipole a HEALPix map of resolution $N_{\rm side}$ can represent (Section 7.11.2); we enforce this when we change resolution.

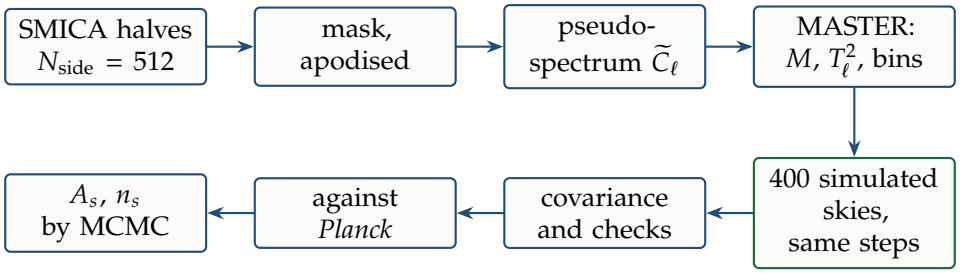


Figure 13.2: The pipeline on the real sky. The top row turns the map into bandpowers; it is the estimator of Section 8.11, applied once to the data. The simulated skies (green) go through exactly the same steps, and their scatter supplies the error bars, the checks and the likelihood used in the bottom row.

The top row of Figure 13.2 is one pass of the MASTER estimator of Section 8.11 over the data, and it yields a single set of bandpowers with no error bar. The error bar, the comparison, the checks and the fit all come from the bottom row: the same estimator applied to simulated skies gives the sampling distribution of the bandpowers, and that distribution is what lets us compare, check and fit.

13.1.1 The files, and a map small enough for a laptop

The *Planck* products are public. We downloaded them once with `curl` from the *Planck* archive at IRSA (Table 13.1). The SMICA maps sit under `all-sky-maps/maps/component-maps/cmb/` of the third release, one level deeper than the path often quoted. Each map file holds several fields: the temperature, the two polarisation components, an *inpainted* temperature map in which the masked regions are filled with a CMB-like constrained realisation, and two masks.

file	content	size
`COM_CMB_IQU-smica_2048_R3.00_full.fits`	SMICA map, full mission	2.0 GB
`COM_CMB_IQU-smica_2048_R3.00_hm1.fits`	first half-mission map	0.6 GB
`COM_CMB_IQU-smica_2048_R3.00_hm2.fits`	second half-mission map	0.6 GB
`COM_Mask_CMB-common-Mask-Int_2048_R3.00.fits`	common temperature mask	0.2 GB
`COM_PowerSpect_CMB-TT-binned_R3.01.txt`	TT bandpowers, $\Delta\ell = 30$	7 kB
`COM_PowerSpect_CMB-TT-full_R3.01.txt`	the same, unbinned	0.2 MB

Table 13.1: The public *Planck* 2018 products used for the power spectrum and, in Section 13.10, for the shapes of the sky.

Why we cannot use the map as it is. A map at $N_{\mathrm{side}} = 2048$ has

$$N_{\mathrm{pix}} = 12 \times 2048^2 = 50\,331\,648$$

pixels, 201 MB per map in single precision. Every simulated sky needs a few spherical-harmonic transforms of a map that size, and the cost of one transform grows as $N_{\mathrm{pix}}^{3/2}$ on the HEALPix grid (Figure 8.2). Four hundred simulations, each with a dozen transforms, at this resolution would take days. At $N_{\mathrm{side}} = 512$ the map has $12 \times 512^2 = 3\,145\,728$ pixels, sixteen times fewer, and each transform is $16^{3/2} = 64$ times cheaper. The price is the band limit: a map at $N_{\mathrm{side}} = 512$ holds multipoles only up to $3 \times 512 - 1 = 1535$, and we analyse only $\ell < 1020$, where the pixel window is still mild (below). *Planck*'s own temperature analysis runs to $\ell = 2508$ [75, Table 13]; the damping tail between 1000 and 2500 is lost to us, and with it most of the leverage on n_s that a long lever arm in ℓ provides (Section 13.7 measures how much).

How to change the resolution honestly. The question is: what map at $N_{\mathrm{side}} = 512$ would *Planck* have made of the same sky, with the same beam? Not the average of the 16 small pixels in each large one, although that is the first idea. Averaging inside a pixel is the pixel window (Section 8.2.1), so a pixel average applied to a map that already carries the $N_{\mathrm{side}} = 2048$ window would carry *both* windows, and the second one would only approximately be a pixel window of $N_{\mathrm{side}} = 512$. It is cleaner in harmonic space, where every effect of the instrument is a multiplication. Write the coefficients of the downloaded map, using the transfer function $T_\ell = B_\ell p_\ell$ of (8.12) with B_ℓ the beam and p_ℓ the pixel window, as

$$\begin{aligned} a_{\ell m}^{(2048)} &\overset{(1)}{=} B_\ell \, p_\ell^{(2048)} \, a_{\ell m}^{\mathrm{sky}} + n_{\ell m}^{(2048)}, \\ \frac{p_\ell^{(512)}}{p_\ell^{(2048)}} \, a_{\ell m}^{(2048)} &\overset{(2)}{=} B_\ell \, p_\ell^{(512)} \, a_{\ell m}^{\mathrm{sky}} + \frac{p_\ell^{(512)}}{p_\ell^{(2048)}} \, n_{\ell m}^{(2048)} \qquad (\ell \le 1535). \end{aligned} \tag{13.1}$$

(1) The observation model of Section 8.5: the sky's coefficients multiplied by the beam and the pixel window, plus noise.

(2) Multiply both sides by the ratio of the two pixel windows and keep only $\ell \le 3 \times 512 - 1$. The sky term now carries $B_\ell p_\ell^{(512)}$, exactly the transfer function of a map made at $N_{\mathrm{side}} = 512$ with the same beam.

Synthesising a map at $N_{\mathrm{side}} = 512$ from the right-hand side of (13.1) gives a map whose sky part is described by $T_\ell = B_\ell p_\ell^{(512)}$, the same form as every simulated map of Chapter 8. The noise term is changed by the same factor, which is why the noise of the degraded map is measured from the degraded maps themselves (Section 13.4.1).

Example: how large is the change? The pixel window of $N_{\mathrm{side}} = 2048$ is close to one over our range: $p_{1019}^{(2048)} = 0.9885$. The window of $N_{\mathrm{side}} = 512$ is not. Its small-ℓ form (8.11) with

$\Omega_{\rm pix} = 4\pi/3\,145\,728 = 3.995 \times 10^{-6}$ sr gives at $\ell = 1019$

$$p_{1019}^{(512)} \approx 1 - \frac{1019 \times 1020 \times 3.995 \times 10^{-6}}{24} = 1 - 0.173 = 0.827,$$

against the tabulated 0.829. The beam at the same multipole is $B_{1019} = 0.820$, so the power of the degraded map at our highest multipole is multiplied by $T_{1019}^2 = 0.462$ (Figure 13.3). Dividing it out is not a small correction, and neither is forgetting it.

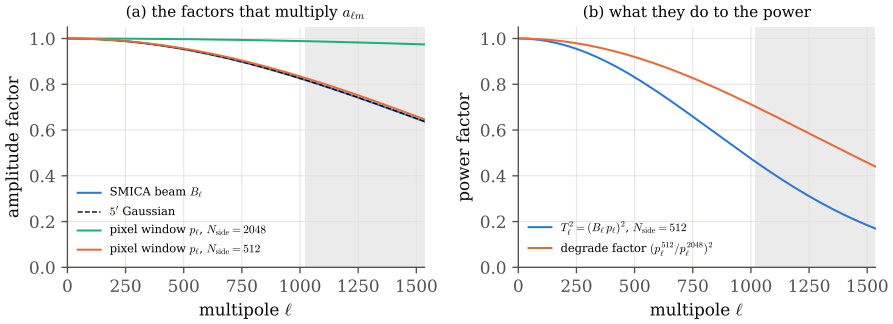


Figure 13.3: (a) The factors that multiply each $a_{\ell m}$ of the map: the SMICA effective beam (blue, on top of a 5′ Gaussian, dashed), and the pixel windows of the downloaded resolution (green) and of the working resolution (orange). (b) Their effect on the power: the squared transfer function of the degraded map (blue) and the squared ratio of pixel windows that the degrade applies (orange). Multipoles above $\ell = 1019$ (grey) are not analysed.

The holes have to be filled first. One more step is needed before the transform, and it is easy to miss. The SMICA map is not zero inside the mask; it contains the residual emission of the Galaxy and of bright sources, which the mask exists to hide. In the full-mission map the largest value inside the holes is 7899 μK, against 501 μK in the kept sky. Truncating a map at $\ell = 1535$ is a low-pass filter, and a low-pass filter applied to a sharp spike produces ripples around it (the Gibbs phenomenon of Fourier series): a hot point source of a few thousand μK would spread rings of tens of μK into the kept pixels around its hole. So before the transform we replace the pixels inside the mask by the inpainted CMB that *Planck* provides. Those pixels stay masked in everything that follows; the inpainting only guarantees that what is hidden in them is no brighter than the CMB, so that the band limit cannot carry it out of the holes.

Example: the mask at the lower resolution. The common mask is a 0/1 map at $N_{\rm side} = 2048$ that keeps a fraction 0.779 of the sky. Each pixel at $N_{\rm side} = 512$ contains $(2048/512)^2 = 16$ small pixels. We keep a large pixel only if all 16 are kept; a pixel that is partly in a hole is discarded, so a hole can only grow. The kept fraction drops to 0.775: the holes grow by at most one large pixel (6.9′) along their edges.

Listing 13.1: The harmonic degrade: fill the holes, transform, swap the pixel windows, synthesise (`code/ch13/01_planck_data.py`)

```
1
2  # ------------------------------------------------------------------ the maps
3  p2048 = np.asarray(hp.pixwin(2048, lmax=LMAX))
4  full_src = lp.raw_path("full")
5  mask2048 = hp.read_map(lp.raw_path("mask"), dtype=np.float64) < 0.5      # True in the holes
6  inpainted = hp.read_map(full_src, field=5, dtype=np.float64) * 1e6                    #
       I_STOKES_INP
7  for kind in ("full", "hm1", "hm2"):
8      if all(lp.degraded_path(kind, n).exists() for n in NSIDES):
```

```
        continue
    src = lp.raw_path(kind)
    m = hp.read_map(src, field=0, dtype=np.float64) * 1e6            # K_CMB -> muK, NESTED -> RING
    # the holes hold bright Galactic and source residuals; a sharp band limit would ring them out
    # over the kept sky, so fill the holes with Planck's inpainted CMB first (they stay masked)
    tag = {"full": "Full", "hm1": "HmOne", "hm2": "HmTwo"}[kind]
    nums[f"TwelveAHoleMax{tag}"] = f"{np.abs(m[mask2048]).max():.0f}"
    nums[f"TwelveAKeptMax{tag}"] = f"{np.abs(m[~mask2048]).max():.0f}"
    m[mask2048] = inpainted[mask2048]
    alm = hp.map2alm(m, lmax=LMAX, iter=1)
    del m
    for n in NSIDES:
        L = 3 * n - 1
        a = hp.resize_alm(alm, LMAX, LMAX, L, L)
        fl = np.asarray(hp.pixwin(n, lmax=L)) / p2048[: L + 1]          # swap the pixel windows
        hp.write_map(lp.degraded_path(kind, n), hp.alm2map(hp.almxfl(a, fl), n, lmax=L),
```

The loop reads each map (`full`, `hm1`, `hm2`) and converts it from K to μK. The two lines before the transform record the largest values inside and outside the holes and then overwrite the holes with the inpainted CMB. `hp.map2alm` computes the $a_{\ell m}$ up to $\ell = 3071$, `hp.resize_alm` keeps those up to $\ell = 1535$, `hp.almxfl` multiplies them by $p_\ell^{(512)}/p_\ell^{(2048)}$ as in (13.1), and `hp.alm2map` synthesises the new map. The whole preparation took 52 s; the 3.4 GB of downloads were then moved out of the working directory, and the degraded maps occupy 13 MB each.

The full analysis does none of this: *Planck* measures its spectrum at $N_{\rm side} = 2048$ from the frequency maps, not from SMICA, up to $\ell = 2508$. Our reduction to $N_{\rm side} = 512$ and $\ell < 1020$ costs the damping tail. The number of independent modes in a map grows as $\ell_{\rm max}^2$, so cutting at 1019 instead of 2508 keeps $(1019/2508)^2 \approx 17\%$ of them, and discards about four fifths of the modes *Planck* analyses (fewer in terms of information, since the noise and the foregrounds grow with ℓ).

13.2 The mask: which sky to trust, and how to let go of the rest

In Section 8.10.4 we met the trade that every mask forces on us. A mask with sharp edges keeps every observed pixel at full weight, but its window spectrum falls slowly, so the coupling matrix carries power from the acoustic peaks into multipoles far away; a field of small holes was the worst case, with a leakage nearly flat in ℓ. Tapering the edges, *apodisation*, cleans the kernel by orders of magnitude, and pays for it in modes: the effective sky fraction $f_{\rm sky} w_2^2/w_4$ drops (in the example of that section, from 802 modes to 600 in one bin). Those masks were caps and bands of our own design. The *Planck* common mask was designed by people who wanted to hide our Galaxy and every bright source in the sky, and its geometry is whatever the sky made it.

What the mask looks like. The common temperature mask at $N_{\rm side} = 512$ keeps a fraction 0.775 of the sky (Figure 13.5a). The masked region has two very different parts. A broad band along the Galactic plane, wider towards the Galactic centre, hides the dust, synchrotron and free-free emission of the Milky Way. Scattered over the rest of the sky are thousands of small holes, each hiding a radio or infrared source bright enough to be detected on its own.

The question. How should each kind of hole be tapered? A wide taper around the Galactic band costs little sky in proportion and cleans a long edge. A wide taper around each of thousands of small holes might cost a great deal. To taper them differently we first have to tell them apart.

13.2.1 Telling the holes apart

A **hole** is a connected set of masked pixels [masked pixels that can be joined by steps between neighbouring masked pixels, without crossing a kept one]. On the HEALPix grid every pixel has eight neighbours (four sharing an edge, four sharing a corner), which `hp.get_all_neighbours` lists. So the holes are the connected components of a graph whose nodes are the masked pixels and whose links join masked neighbours, the same idea as the clusters of Chapter 12. We find them with `scipy.sparse.csgraph.connected_components`, then measure each hole's area as its number of pixels times the pixel area. A hole larger than $20\,\mathrm{deg}^2$ is *Galactic*, the rest are *point-source holes*.

Example: what $20\,\mathrm{deg}^2$ means in pixels. The sky has $41\,253\,\mathrm{deg}^2$, so at $N_{\mathrm{side}} = 512$ one pixel covers $41\,253/3\,145\,728 = 0.0131\,\mathrm{deg}^2$, a square of 6.9′ on a side. A hole of $20\,\mathrm{deg}^2$ is about 1500 pixels, a disc of radius 2.5°. A point-source hole is a disc of a few tens of arcminutes, so tens of pixels: the two populations are separated by two orders of magnitude, and the threshold does not need to be chosen carefully.

The result: 3307 holes, of which 3 are larger than $20\,\mathrm{deg}^2$. The Galactic holes alone keep a fraction 0.796 of the sky; the point-source holes alone remove a fraction 0.021. Together they keep $0.775 = 0.796 - 0.021$, where the source holes that sit inside the Galactic band are not counted twice.

13.2.2 Two tapers

Planck apodises its Galactic masks with a Gaussian of 4.71° FWHM ($\sigma = 2°$) and its point-source masks with one of 30′ FWHM, and multiplies the two [75, §3.2.2]. The apodisation is computed from the distance of each pixel to the nearest mask border, which guarantees that the window is zero inside every hole. We follow the same logic with the cosine taper of Section 8.10.4: for a kept pixel at angular distance d from the nearest hole of a given kind,

$$W(d) = \begin{cases} \tfrac{1}{2}\bigl(1 - \cos(\pi d/\theta_{\mathrm{ap}})\bigr), & d < \theta_{\mathrm{ap}}, \\ 1, & d \geq \theta_{\mathrm{ap}}, \end{cases} \tag{13.2}$$

with $\theta_{\mathrm{ap}} = 2°$ for the Galactic holes and $\theta_{\mathrm{ap}} = 0.5°$ for the source holes, and the window of the analysis is the product of the two tapered masks (Figure 13.4). The window rises smoothly from 0 at the edge, through $\tfrac{1}{2}$ at $d = \theta_{\mathrm{ap}}/2$, to 1, with zero slope at both ends.

Why not one taper for all? Take the wider taper and put it around every source. The ring between the edge of a small hole and a distance of 2° has an area of nearly $\pi(2°)^2 = 12.6\,\mathrm{deg}^2$, and there are about 3300 such holes:

$$3300 \times 12.6\,\mathrm{deg}^2 \approx 41\,500\,\mathrm{deg}^2,$$

the whole sky. The rings overlap, so the real loss is smaller, but most of the sky would sit at reduced weight. With 0.5° the same count gives $3300 \times \pi \times 0.25\,\mathrm{deg}^2 \approx 2600\,\mathrm{deg}^2$, about 6% of the sky, and in fact less, because inside each ring the window is on average one half. The

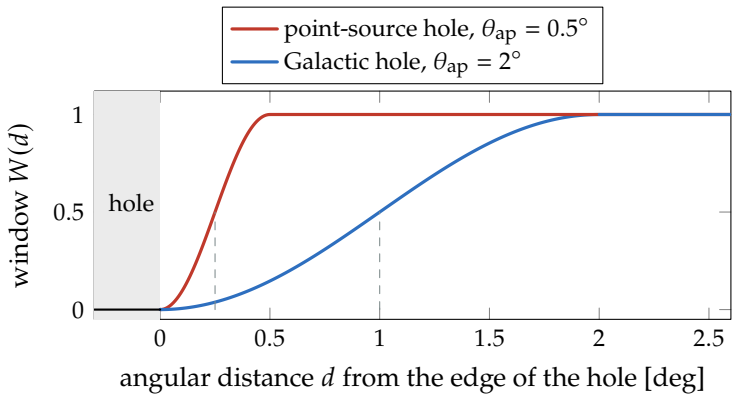


Figure 13.4: The two tapers of (13.2). Inside a hole (grey) the window is zero. Moving away from the edge it rises as half a cosine, reaching one half at half the taper width (dashed) and one at the full width: within 0.5° of a point source (red), within 2° of the Galactic band (blue).

Galactic band is one long edge, and a 2° taper along it costs a thin strip. So the widths are set by geometry: wide where there is a little edge, narrow where there is a lot of it. *Planck* reports the same rule of thumb: its large Galactic apodisation reduces the effective sky fraction by about 10% [75, §3.2.2].

Example: what the taper costs us. The moments of the window (Section 8.9: w_i is the average of W^i over the pixels where $W > 0$) are $f_{\rm sky} = 0.775$ and $w_2 = 0.894$, so the mean of W^2 over the whole sky is $f_{\rm sky} w_2 = 0.692$. The effective sky fraction that counts modes (8.81) is $f_{\rm sky} w_2^2 / w_4 = 0.708$, against 0.775 for the binary mask. Where the error is cosmic variance, it scales as one over the square root of the number of modes, so the taper makes the error bars larger by

$$\sqrt{0.775/0.708} \approx 1.05,$$

five per cent. What it buys is visible in Section 13.3: without it, the correction for mode coupling has to undo leakage of ten per cent and more.

13.2.3 Four more windows, for the checks

A real analysis does not trust a spectrum measured through one window. If the map outside the mask is a Gaussian isotropic CMB, the spectrum must not depend on which part of that sky we use, up to the scatter that using less sky brings. So we build four more windows, each designed to expose one of the explanations listed in Section 13.1 (Figure 13.5c,d).

- The *binary* common mask, with no taper ($f_{\rm sky} = 0.775$). Same sky, different weights: a difference tests the correction for mode coupling.
- The main window times a cut $|b| > 30°$, with a 2° cosine edge ($f_{\rm sky} = 0.488$). Galactic residuals would be stronger in the main window, which goes closer to the plane.
- The main window restricted to the northern or to the southern Galactic hemisphere ($f_{\rm sky} = 0.387$ and 0.387). A statistically isotropic sky gives the same spectrum in both, up to the scatter.

The coupling matrices. For each window we compute the spectrum $\mathcal{W}_\ell$ of the window with `hp.anafast` and the coupling matrix $M_{\ell\ell'}$ of (8.61) with the Wigner-3j code of Section 8.10.2, imported from `code/ch08/lib_masks.py`. We compute it up to $\ell = 1199$, not just to the last

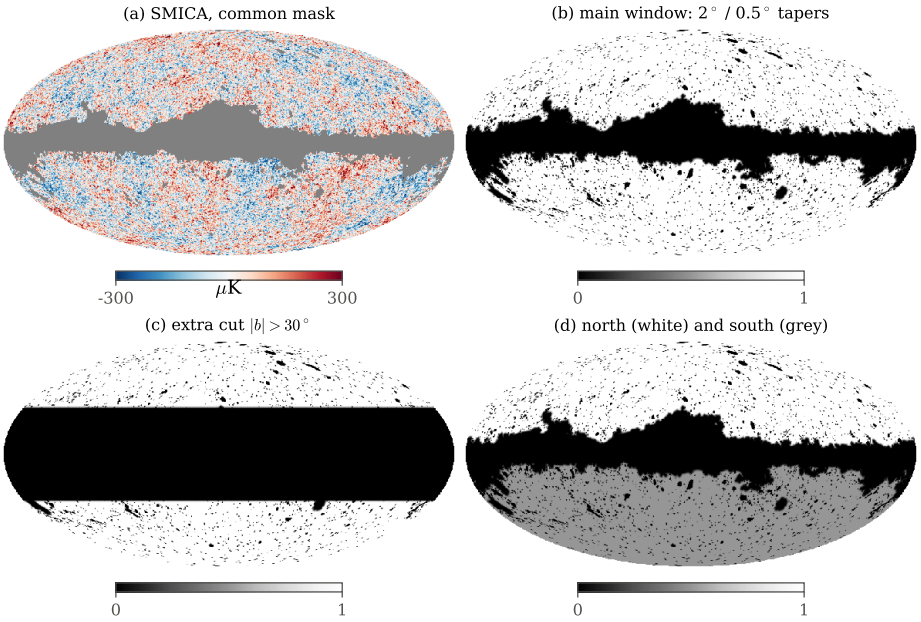


Figure 13.5: (a) The degraded SMICA map outside the common mask (grey), in Galactic coordinates. (b) The window of the analysis: the common mask with its Galactic holes tapered over 2° and its point-source holes over 0.5°. (c) The same window with everything at $|b| < 30°$ removed. (d) The northern (white) and southern (grey) Galactic hemispheres of the window. Panels (c) and (d) are used only by the checks.

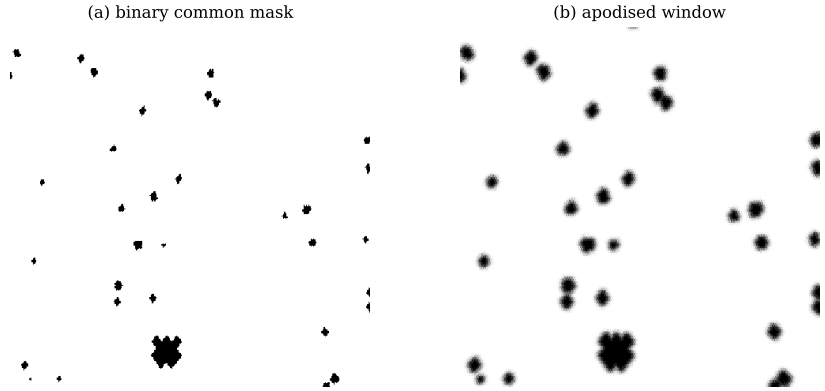


Figure 13.6: A 20° patch at high Galactic latitude, centred on $(l, b) = (-60°, 45°)$. (a) The binary common mask: a scatter of point-source holes. (b) The window: each hole is now surrounded by a grey ring where the weight rises from zero to one over 0.5°.

multipole we report, 1019: a true multipole slightly above 1019 leaks into our last bins, and the correction can remove that leakage only if the matrix knows about it. Each matrix is 1200×1200 and is cached in `data/ch13/`.

Listing 13.2: Splitting the holes, tapering each kind, and the four windows for the checks (`code/ch13/02_mask.py`)

```
binary = hp.read_map(lp.PLANCK / f"common_mask_n{NSIDE}.fits", dtype=np.float64)
gal, ps, ncomp, nbig = lp.mask_components(binary, NSIDE)
nums.update({"TwelveAHoles": ncomp, "TwelveAHolesBig": nbig,
             "TwelveAFskyGal": f"{gal.mean():.3f}", "TwelveAFskyPS": f"{1 - ps.mean():.3f}"})
print(f"{ncomp} holes, {nbig} extended; Galactic part removes {1 - gal.mean():.3f}, sources {1 - ps.mean():.3f}")
```

```
W_main = lp.apodise(gal, NSIDE, 2.0) * lp.apodise(ps, NSIDE, 0.5)
masks = {
    "binary": binary,
    "main": W_main,
    "cut30": W_main * lp.latitude_taper(NSIDE, "both", 30.0, 2.0),
    "north": W_main * lp.latitude_taper(NSIDE, "north", 0.0, 2.0),
    "south": W_main * lp.latitude_taper(NSIDE, "south", 0.0, 2.0),
}
```

`lp.mask_components` returns two 0/1 masks, one with only the Galactic holes and one with only the point-source holes. `lp.apodise` computes, for every kept pixel, the angle to the nearest masked pixel (a nearest-neighbour search on the unit vectors of the pixel centres, with `scipy.spatial.cKDTree`) and applies (13.2). The latitude cuts use the same cosine edge, as a function of Galactic latitude b instead of the distance to a hole.

Planck's own likelihood uses a different mask at each frequency (keeping 66%, 57% and 47% of the sky at 100, 143 and 217 GHz) with the point-source holes of each channel [75, §3.2.2, Fig. 29]; we use the single common mask of the SMICA map, which keeps more sky (0.775) because SMICA has already removed most of the Galactic emission. The cost of the substitution is that our spectrum is more exposed to any residual of SMICA's foreground cleaning near the plane, which is exactly what the $|b| > 30^\circ$ check is for.

13.3 From the map to bandpowers, one correction at a time

In Sections 8.11 and 8.11.1 three recipes for undoing a mask were tried on simulated skies. Dividing the pseudo-spectrum by the sky fraction was biased wherever the spectrum had structure, because the mask moves power between multipoles. Inverting the coupling matrix multipole by multipole was exact but noisy and, for a small patch, impossible. Inverting it between bins, the MASTER recipe, was unbiased for any spectrum that is flat inside each bin. On simulations we could see this because we knew the spectrum we had put in. On the real sky we do not, so we do the next best thing: apply the corrections one at a time and watch what each does to the numbers, with *Planck*'s published spectrum as a reference line (Figure 13.8). Before any of that, two small things must be removed from the map, and the spectrum must be computed in a way that has no noise bias.

13.3.1 The monopole and the dipole

The SMICA map has had the monopole of the CMB (its mean temperature) and the large dipole caused by our motion removed. What remains of $\ell = 0$ and $\ell = 1$ is an arbitrary offset and a residual of that subtraction, neither of which is part of the anisotropy we model. On the full sky they would sit in $\ell = 0, 1$ and never touch the multipoles we analyse. Behind a mask they do: Y_{00} and Y_{1m} are not orthogonal to the higher harmonics over the cut sky, so a pure dipole on the kept pixels has pseudo-coefficients at every ℓ (the leakage of Section 8.9.1). So we fit $c + \boldsymbol{d} \cdot \hat{\boldsymbol{n}}_p$ to the kept pixels by least squares, four numbers (c and the three components of $\boldsymbol{d}$), and subtract the fit from the whole map. The fit gives $c = 0.6\,\mu$K and $|\boldsymbol{d}| = 5.2\,\mu$K, small next to the $100\,\mu$K fluctuations of the map. Because a mask breaks the orthogonality, this fit also absorbs the small part of the real CMB anisotropy that happens to look like a dipole on the kept sky; it is a few μK, and it affects mostly the buffer bin $2 \le \ell < 30$ that we do not report.

13.3.2 Two halves, one sky

The half-mission maps of Section 13.1 let us avoid the noise bias altogether. Recall (8.26): if the two maps have coefficients $d^{(i)}_{\ell m} = T_\ell a_{\ell m} + n^{(i)}_{\ell m}$ with independent noise, the cross-spectrum $\widehat{C}^{12}_\ell = \frac{1}{2\ell+1}\sum_m \mathrm{Re}\, d^{(1)}_{\ell m} d^{(2)*}_{\ell m}$ has mean $T^2_\ell C_\ell$: the noise of one half has nothing to multiply in the other, so it averages away instead of adding power. Behind a mask the same holds for the pseudo-spectra, since the mask multiplies both maps by the same W.

The two halves give a second map for free. Their half-difference (hm1 − hm2)/2 contains no sky, since the sky is the same in both, and its noise has the same level as the noise in their average (hm1 + hm2)/2. It is a **null map**: a map of what the noise looks like. The three spectra are tied by an identity. Write a and b for the masked coefficients of the two halves at some (ℓ, m):

$$\left|\frac{a+b}{2}\right|^2 - \left|\frac{a-b}{2}\right|^2 \overset{(1)}{=} \frac{|a|^2 + |b|^2 + 2\mathrm{Re}\, ab^*}{4} - \frac{|a|^2 + |b|^2 - 2\mathrm{Re}\, ab^*}{4} \overset{(2)}{=} \mathrm{Re}\, ab^*, \tag{13.3}$$

(1) $|a \pm b|^2 = (a \pm b)(a^* \pm b^*) = |a|^2 + |b|^2 \pm (ab^* + a^*b)$, and $ab^* + a^*b = 2\mathrm{Re}\, ab^*$.

(2) The squared moduli cancel.

Averaging over m: the spectrum of the average map minus the spectrum of the half-difference is exactly the cross-spectrum. The SMICA full-mission map is not exactly the average of the two halves (the halves are separated with their own weights), so on the data the identity holds only approximately, and how well it holds is one of the checks of Section 13.6.

13.3.3 *Planck*'s bins

To compare our spectrum with *Planck*'s we must bin exactly as *Planck* bins. The published binned spectrum uses bins of $\Delta\ell = 30$ starting at $\ell = 30$, and inside each bin it weights C_ℓ by $\ell(\ell+1)$ [75, (22)]:

$$C_b = \sum_{\ell \in b} w_{b\ell} C_\ell, \qquad w_{b\ell} = \frac{\ell(\ell+1)}{\sum_{\ell' \in b} \ell'(\ell'+1)}, \qquad \ell^{\mathrm{eff}}_b = \sum_{\ell \in b} w_{b\ell}\, \ell, \qquad D_b = \frac{\ell^{\mathrm{eff}}_b(\ell^{\mathrm{eff}}_b + 1)}{2\pi} C_b. \tag{13.4}$$

The weight $\ell(\ell+1)$ is there because $D_\ell = \ell(\ell+1)C_\ell/2\pi$ varies slowly across a bin while C_ℓ falls steeply: weighting C_ℓ by $\ell(\ell+1)$ averages the slowly varying quantity. The file quotes D_b at the multipole ℓ^{eff}_b.

In the notation of the MASTER estimator (Section 8.11.3), binning is a matrix $P_{b\ell}$ that turns a spectrum into bandpowers, and $Q_{\ell b}$ is a matrix that spreads a bandpower back over the multipoles of its bin. Reading them off (13.4),

$$P_{b\ell} = \frac{\ell^{\mathrm{eff}}_b(\ell^{\mathrm{eff}}_b + 1)}{2\pi} w_{b\ell}, \qquad Q_{\ell b} = \frac{2\pi}{\ell^{\mathrm{eff}}_b(\ell^{\mathrm{eff}}_b + 1)} \quad (\ell \in b), \tag{13.5}$$

and both are zero for ℓ outside bin b. Q is the spectrum that the MASTER recipe assumes inside each bin: a C_ℓ that is constant across the bin, at the value whose D at ℓ^{eff}_b is the bandpower. Binning that spectrum must give the bandpower back:

$$\sum_\ell P_{b\ell} Q_{\ell b'} \overset{(1)}{=} \delta_{bb'} \sum_{\ell \in b} \frac{\ell^{\mathrm{eff}}_b(\ell^{\mathrm{eff}}_b + 1)}{2\pi} w_{b\ell} \frac{2\pi}{\ell^{\mathrm{eff}}_b(\ell^{\mathrm{eff}}_b + 1)} \overset{(2)}{=} \delta_{bb'} \sum_{\ell \in b} w_{b\ell} \overset{(3)}{=} \delta_{bb'}. \tag{13.6}$$

(1) $P_{b\ell}$ is zero unless $\ell \in b$ and $Q_{\ell b'}$ is zero unless $\ell \in b'$; bins do not overlap, so the product vanishes unless $b = b'$.

(2) The two prefactors cancel.
(3) The weights of a bin sum to one by construction.

So $PQ = \mathbb{1}$, the property the derivation of (8.67) needs.

Example: the first bin. For $30 \le \ell \le 59$, $\ell^{\rm eff} = \sum \ell^2(\ell+1)/\sum \ell(\ell+1) = 47.71$; the *Planck* file quotes 47.711, and our binning code reproduces every $\ell_b^{\rm eff}$ in the file to better than one part in 10^5 (the largest difference is 3.1e-06). Now take the fiducial spectrum and bin it. The plain average of D_ℓ over the bin is 1322 μK^2, and D_ℓ at $\ell = 48$ is 1384 μK^2, but (13.4) gives $D_b = 1463\,\mu\text{K}^2$. A bandpower is neither the spectrum at $\ell^{\rm eff}$ nor its average. To see why, take a $D_\ell = D$ that is flat across the bin, so that $C_\ell = 2\pi D/\ell(\ell+1)$:

$$D_b \overset{(1)}{=} \frac{\ell^{\rm eff}(\ell^{\rm eff}+1)}{2\pi} \sum_{\ell \in b} \frac{\ell(\ell+1)}{\sum_{\ell' \in b} \ell'(\ell'+1)} \frac{2\pi D}{\ell(\ell+1)} \overset{(2)}{=} \frac{\ell^{\rm eff}(\ell^{\rm eff}+1)\,\Delta\ell}{\sum_{\ell' \in b} \ell'(\ell'+1)} D \overset{(3)}{=} \frac{\ell^{\rm eff}(\ell^{\rm eff}+1)}{\overline{\ell(\ell+1)}} D. \tag{13.7}$$

(1) (13.4) with $C_\ell = 2\pi D/\ell(\ell+1)$.
(2) The factors $\ell(\ell+1)$ and 2π cancel, and the sum of 1 over the bin is its width $\Delta\ell$.
(3) $\overline{\ell(\ell+1)}$ is the plain average of $\ell(\ell+1)$ over the bin.

In the first bin the factor is 2324/2100 = 1.1069: the effective multipole is a weighted mean of ℓ, and squaring a mean is not the mean of the squares. In the last bin, $990 \le \ell \le 1019$, the same factor is 1.0002, because the bin is narrow compared with its ℓ. The lesson is a rule we will use repeatedly: a theory curve is compared with bandpowers only after it has been passed through *the same* binning (and, below, the same window).

We add a buffer bin $2 \le \ell < 30$ below the reported range and six bins $1020 \le \ell < 1200$ above it. They are solved for together with the reported bins, so that the power at those multipoles, which the mask spreads into our range, is accounted for, and then they are discarded. We report 33 bins, $30 \le \ell \le 1019$.

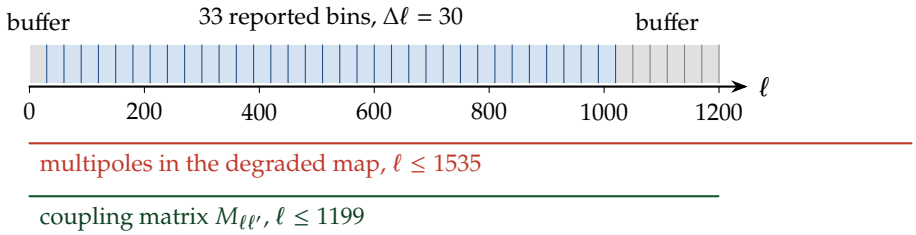


Figure 13.7: The 33 reported bins (blue) each contain 30 multipoles and reproduce the *Planck* binning from $\ell = 30$ to 1019. Masking couples these bins to multipoles outside the reported range, so we also solve for the grey buffer bins on either side before discarding them. The degraded map holds multipoles up to 1535; the coupling matrix is computed up to 1199.

13.3.4 MASTER with our own coupling matrix

With the coupling matrix M of the window (Section 13.2.3) and the squared transfer function $T_\ell^2 = (B_\ell p_\ell)^2$ of Section 13.1.1, the binned master equation (8.67) [Hivon (20)–(26)] reads, for the cross-spectrum (no noise term to subtract, and no time-domain filtering, $F_\ell = 1$),

$$\langle P\widetilde{C}^{12} \rangle = K\,D, \qquad K_{bb'} = \sum_{\ell\ell'} P_{b\ell} M_{\ell\ell'} T_{\ell'}^2 Q_{\ell' b'}, \qquad \widehat{D}_b = \sum_{b'} K^{-1}_{bb'} \sum_{\ell} P_{b'\ell} \widetilde{C}_\ell^{12}. \tag{13.8}$$

K is 40×40 (the 33 reported bins and the 7 buffer bins), and it is inverted once. As derived in (8.68), the mean of $\widehat{D}_b$ for the true spectrum is

$$\langle \widehat{D}_b \rangle = \sum_\ell \mathcal{F}_{b\ell} C_\ell, \qquad \mathcal{F}_{b\ell} = \sum_{b'} K^{-1}_{bb'} \sum_{\ell'} P_{b'\ell'} M_{\ell'\ell} T^2_\ell, \tag{13.9}$$

the **bandpower window functions**. They are the honest way to put a theory on the plot: a spectrum that is not flat inside each bin is turned into bandpowers by $\mathcal{F}$, not by P. NaMaster, the public code that implements MASTER for any mask, writes the same two equations, the binned coupling matrix and the filter that must be applied to any theory before it is compared with the data [169, (16)–(18)].

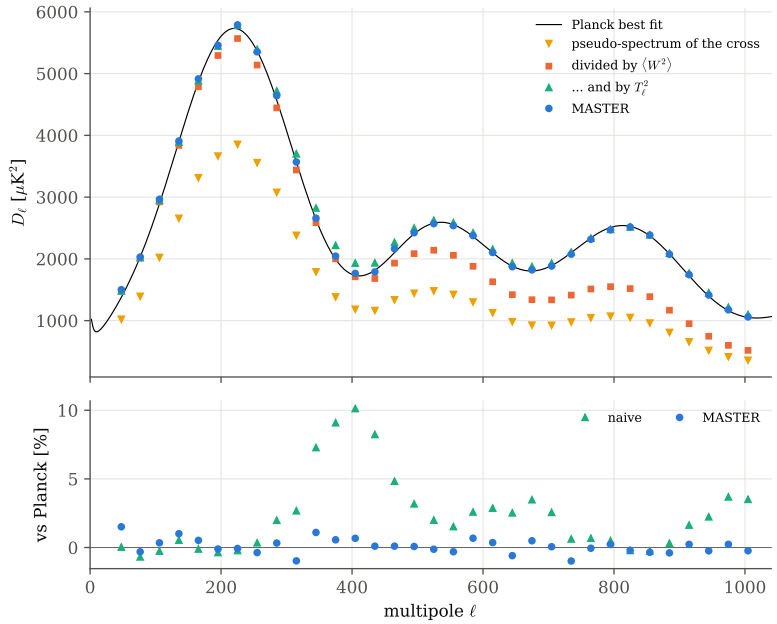


Figure 13.8: The SMICA half-mission cross-spectrum, binned in *Planck*'s bins, after each correction. Top: the binned pseudo-spectrum (yellow triangles); divided by the mean of W^2 (orange squares); then also by T^2_ℓ (green triangles, the naive estimator); and the MASTER bandpowers (blue). The black line is the *Planck* best fit. Bottom: the naive and the MASTER bandpowers relative to *Planck*'s published bandpowers.

What each step does, on the data. Figure 13.8 follows the cross-spectrum through the corrections.

1. *The pseudo-spectrum* is low everywhere. In the first bin it is 1023 μK^2, against 1479 in *Planck*'s spectrum.
2. *Dividing by the mean of* W^2 over the sky, $f_{\rm sky} w_2 = 0.692$, fixes the low multipoles ($1023/0.692 \approx 1478$), but the high ones stay low, and more so the higher ℓ.
3. *Dividing by* T^2_ℓ undoes the beam and the pixels; at $\ell = 1019$ that is a factor $1/0.462$. The result is the naive estimator (8.64), and it is off by up to 10%, with a characteristic pattern: too high in the troughs, especially the first one near $\ell = 400$, nearly right at the peaks.
4. *MASTER* removes the pattern. Its bandpowers differ from *Planck*'s by at most 1.5%, 0.5% in the root mean square.

Why the naive estimator fails in the troughs. Dividing by $f_{\rm sky} w_2$ assumes that the mask removes a fixed share of the power at each multipole and moves none of it: $M_{\ell\ell'} = f_{\rm sky} w_2 \, \delta_{\ell\ell'}$. The true coupling matrix has a narrow core around $\ell' = \ell$, set by the size of the large sky regions, and a broad low floor that comes from the thousands of small holes, whose window spectrum is nearly flat (Section 8.10.4). The floor moves a small fraction of every multipole's power to multipoles far away. At a peak the power that leaves and the power that arrives nearly balance. In the trough at $\ell \approx 400$, where $D_\ell \approx 1750\,\mu\mathrm{K}^2$, the arriving share of the first peak, where $D_\ell \approx 5800\,\mu\mathrm{K}^2$, is not balanced by the little that leaves, so the trough fills up by about a tenth. The 0.5° taper of the source holes weakens the floor but cannot remove it, because the holes themselves are small. MASTER does not need the floor to be small; it needs it to be known, and M knows it.

Example: the residual of MASTER, in perspective. The bottom panel of Figure 13.8 shows MASTER within 1.5% of *Planck*'s bandpowers. Is that agreement good, or suspiciously good? Neither question can be answered yet. Both bandpowers are measurements of nearly the same sky, so they share most of their cosmic variance, and their difference is much smaller than either one's error bar. How much smaller is a question about the joint sampling distribution of two estimators, and it needs simulated skies (Sections 13.4 and 13.5.4).

Listing 13.3: The spectra of every window: pseudo-spectra of the halves, their cross, their half-difference, and the MASTER correction (`code/ch13/03_bandpowers.py`)

```
# ------------------------------------------------------------- 1-4. spectra for every window
out = {"edges": EDGES, "rep": REP}
for name, W in masks.items():
    if name in ("gal", "ps"):
        continue
    W = W.astype(np.float64)
    M = lp.coupling(W, LC, name=name)
    mst = lp.Master(M, T2, EDGES)
    a1, a2, af = lp.pseudo_cl(W, [maps["hm1"], maps["hm2"], maps["full"]], LC)
    pcl = {"cross": hp.alm2cl(a1, a2), "null": hp.alm2cl(0.5 * (a1 - a2)), "auto": hp.alm2cl(af)}
    for k, c in pcl.items():
        out[f"{name}_{k}"] = mst(c)
        out[f"{name}_pcl_{k}"] = c
    out[f"{name}_F"] = mst.window()
    out[f"{name}_leff"] = mst.leff
    out[f"{name}_fw2"] = np.mean(W ** 2)
```

For each window the loop loads the cached coupling matrix and builds the MASTER operator (`lp.Master`: it forms K of (13.8) once and inverts it). `lp.pseudo_cl` multiplies each map by the window and transforms it. From the three sets of coefficients come the cross-spectrum of the halves, the spectrum of their half-difference and the auto-spectrum of the full map, each corrected by the same operator. `mst.window()` returns the bandpower windows $\mathcal{F}_{b\ell}$ of (13.9), which the comparison with theory uses.

13.4 How much would the bandpowers scatter? Simulated skies

We now have 33 numbers, one per bin, measured from one sky. Each of them is a draw from a sampling distribution that we cannot observe: the distribution of the bandpowers over all the skies the standard model could have produced, each seen through *Planck*'s beam, pixels, noise and our mask. In Chapter 8 that distribution was the research problem, and simulation was the answer: for a masked, noisy sky no formula gives the covariance of

MASTER bandpowers exactly, and the scatter of the estimator over many simulated skies is the measurement (Section 8.15). Here the same machine runs for real. Two things must be decided before it can: what noise the simulated maps carry, and how many of them we can afford.

13.4.1 A noise model built from the data

Planck simulates its noise end to end: the time-ordered data of every detector are simulated with the measured noise properties, made into maps by the same map-making code, and passed through the component separation, giving a few hundred noise realisations of each map (the FFP10 simulations). We have only the data. But the data contain a map of their own noise: the half-difference $d = (\mathrm{hm1} - \mathrm{hm2})/2$ of Section 13.3.2, in which the sky cancels.

The question. What random map, added to a simulated sky, would make a half-mission map whose noise looks like ours? Two features of d matter. Its *amplitude varies over the sky*, because *Planck*'s scanning covers some regions many more times than others and the component separation weights the frequencies differently in different places. And its *power varies with scale*, because SMICA's weights and the deconvolution of each channel's beam to the common 5′ make the noise anything but white. A model that captures both, and nothing more, is

$$n_{\mathrm{half}}(p) = \sqrt{2}\,\sqrt{v(p)}\;g(p), \tag{13.10}$$

where $v(p)$ is the local variance of d relative to its mean (the map d^2 smoothed with a 2° Gaussian and divided by its average over the kept sky), and g is a statistically isotropic Gaussian field whose spectrum N_ℓ^g is that of $d/\sqrt{v}$, measured through the window and divided by the mean of W^2 (Section 8.11.1), then averaged over 31 neighbouring multipoles to remove its own scatter.

Where the $\sqrt{2}$ comes from. Suppose the noise of each half has variance σ^2 in some pixel, and the two halves are independent. Then

$$\mathrm{Var}(d) \overset{(1)}{=} \mathrm{Var}\Big(\frac{n_1 - n_2}{2}\Big) \overset{(2)}{=} \frac{\mathrm{Var}\,n_1 + \mathrm{Var}\,n_2}{4} \overset{(3)}{=} \frac{\sigma^2}{2} \quad\Longrightarrow\quad \sigma^2 = 2\,\mathrm{Var}(d). \tag{13.11}$$

(1) The sky is identical in the two halves and cancels in d.
(2) The variance of a sum of independent variables is the sum of the variances, and a constant factor comes out squared: $(1/2)^2$, with $(-1)^2 = 1$ for the second term.
(3) Both halves have the same noise level.

So a half's noise has twice the variance of d, a factor $\sqrt{2}$ in amplitude. By the same steps, the average of the halves, $(n_1 + n_2)/2$, has variance $\sigma^2/2 = \mathrm{Var}(d)$: the noise of the full map has the spectrum of d itself, and the noise of one half has twice that spectrum.

Example: how much noise is there? Very little, at these multipoles. At $\ell = 1000$ the noise power of the full map is 2% of the beamed sky power $T_\ell^2 C_\ell$, and the noise of a single half stays below the beamed sky at every $\ell \le 1199$ (Figure 13.9a). In temperature, below $\ell \approx 1000$, *Planck* is limited by cosmic variance, not by its detectors, which is why the noise model can be approximate without changing the error bars much. Where the noise is, however, varies a lot: v reaches 11.8 times the mean in the worst regions (Figure 13.9b). Close to the mask edges v is artificially small, because the holes were filled with the same inpainted map in both halves, so $d = 0$ there and the smoothing pulls v down; a floor of 0.05 keeps $d/\sqrt{v}$ finite.

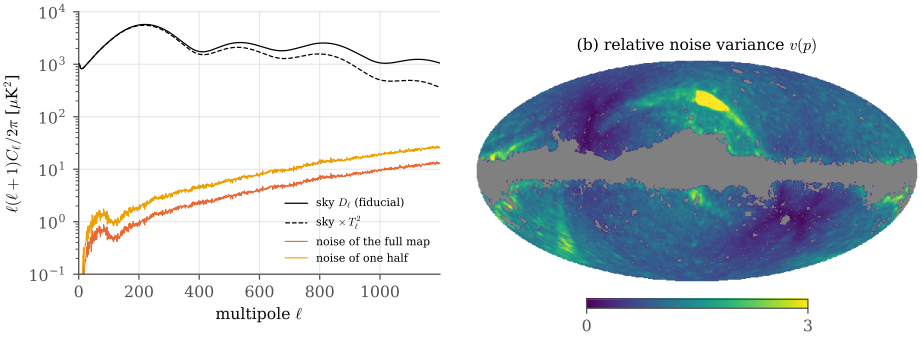


Figure 13.9: (a) The fiducial sky spectrum (solid black), the same after the beam and pixel window (dashed), and the noise spectra measured from the half-difference map: the noise of the full map (red) and of one half (orange). Below $\ell = 1200$ the sky dominates everywhere. (b) The relative noise variance $v(p)$ outside the mask: the noise level changes by more than a factor of ten over the sky.

What the model leaves out is equally important to state. The noise is drawn Gaussian, while residual glitches and foreground leftovers are not. Its correlations are those of a modulated isotropic field, while real scanning noise has stripes along the scan direction. The two halves are drawn independent, while *Planck* found small correlations between half-missions at high multipoles [75, §3.2.1]. And v and N_ℓ^g were estimated from the very map whose noise they describe, so the half-difference test of Section 13.6 is a consistency check of the model, not an independent test of it.

13.4.2 The simulated skies

The rule that makes simulations useful is simple and absolute: *whatever is done to the data is done to every simulation*. The same beam, the same pixel window, the same noise model, the same five windows, the same binning and the same MASTER operator.

Recipe for one simulated sky.

1. Draw $a_{\ell m}$ from the fiducial lensed spectrum C_ℓ (CAMB at the *Planck* 2018 best fit [8], Section 7.11.1), for $\ell \le 1535$.
2. Multiply by $T_\ell = B_\ell p_\ell^{(512)}$ and synthesise the sky map s at $N_{\rm side} = 512$.
3. Make two half-mission maps $s + n_1$ and $s + n_2$, with n_1, n_2 independent draws of (13.10).
4. For each of the five windows, compute the pseudo-coefficients of both halves, their cross-spectrum and the spectrum of their half-difference, and correct both with that window's MASTER operator (13.8).

The output is, for each sky and each window, 33 cross bandpowers and 33 null bandpowers.

One sky takes 1.7 s on four cores: three map syntheses and ten masked transforms at $N_{\rm side} = 512$. We ran 400 skies, in chunks of 50 cached in `data/ch13/` so that an interrupted run resumes where it stopped. The skies are Gaussian, while the real sky is slightly non-Gaussian through gravitational lensing; the lensed spectrum gets the mean right, and the extra covariance that lensing induces between multipoles is negligible at our error bars below $\ell \approx 1000$.

Listing 13.4: One simulated sky, observed and analysed like the real one (`code/ch13/04_sims.py`)

```
def one_sky(rng):
    a = cs.synalm(clf, LS, rng)                          # the sky
```

```
    s = hp.alm2map(hp.almxfl(a, T), NSIDE, lmax=LS)          # as seen through beam and
    pixels
    halves = []
    for _ in range(2):
        g = hp.alm2map(cs.synalm(ng, LS, rng), NSIDE, lmax=LS)
        halves.append(s + np.sqrt(2.0) * sqv * g)
    res = {}
    for k in WINDOWS:
        a1, a2 = lp.pseudo_cl(Ws[k], halves, LC)
        res[f"{k}_cross"] = msts[k](hp.alm2cl(a1, a2))
        res[f"{k}_null"] = msts[k](hp.alm2cl(0.5 * (a1 - a2)))
    return res
```

`cs.synalm` is the Gaussian $a_{\ell m}$ generator of Chapter 8 (`code/ch08/lib_cmbsim.py`). The sky map is made once and shared by the two halves; each half gets its own noise, a Gaussian field with spectrum N_ℓ^g multiplied pixel by pixel by $\sqrt{2v}$. The inner loop is Section 13.3.4 again, line for line.

13.4.3 What the simulations say

Is the estimator unbiased? If MASTER is unbiased, the average of the simulated bandpowers must equal the fiducial spectrum passed through the bandpower windows, $t_b = \sum_\ell \mathcal{F}_{b\ell} C_\ell$ (13.9). If the theory were right, would an average of n simulated bandpower vectors as far from $\boldsymbol{t}$ as ours, or farther, be something we would reasonably expect? The average $\bar{\boldsymbol{x}}$ of n independent vectors with covariance $\mathbf{C}$ has covariance $\mathbf{C}/n$. For Gaussian vectors with mean $\boldsymbol{t}$,

$$\chi^2_{\text{bias}} = (\bar{\boldsymbol{x}} - \boldsymbol{t})^{\mathsf{T}} \left(\frac{\mathbf{C}}{n}\right)^{-1} (\bar{\boldsymbol{x}} - \boldsymbol{t}) = n\,(\bar{\boldsymbol{x}} - \boldsymbol{t})^{\mathsf{T}} \mathbf{C}^{-1} (\bar{\boldsymbol{x}} - \boldsymbol{t}) \tag{13.12}$$

is a χ^2 with $p = 33$ degrees of freedom: a quadratic form in a Gaussian vector, weighted by the inverse of its own covariance, is a sum of p squared independent unit Gaussians (Section 2.13.3). In practice $\mathbf{C}$ is itself estimated from the same simulations, and its inverse is multiplied by the Hartlap factor $(n - p - 2)/(n - 1) = 0.915$ derived in Section 8.13.2, which removes the bias of the inverse of a noisy matrix.

With the exact windows, $\chi^2_{\text{bias}} = 43.7$ for 33 degrees of freedom, a probability to exceed of 0.10: no sign of bias, at the precision of 400 simulations, which is $1/\sqrt{400}$ of an error bar. If instead the theory is binned with the plain operator P, as if it were flat inside each bin, $\chi^2_{\text{bias}} = 1099$ (Figure 13.10b). The difference between the two theory vectors is at most 1.2%, which is 0.46 of a single bandpower's error bar: small for one sky, but the average of 400 skies is twenty times more precise and sees it. The same is true of any analysis that averages many bandpowers into a few parameters, which is why the NaMaster authors insist on the exact windows [169, §2.1, Fig. 10].

How large are the error bars? The square roots of the diagonal of $\mathbf{C}$ are our error bars. Two estimates give them a size. The cosmic variance of a bin on the *full* sky, with no noise, is the variance of the average of $\Delta\ell$ independent $\widehat{D}_\ell$, each of variance $2D_\ell^2/(2\ell + 1)$ (7.77). Our error bars are larger by a factor 1.47 in the first bin, 1.27 (median) over the first fifteen, and 1.30 in the last. The mode counting of (8.81) predicts $1/\sqrt{f_{\text{sky}} w_2^2/w_4} = 1.19$ where the noise is negligible, and more where it is not. The bins in the middle of the range, where the noise is still negligible, sit about 7% above that prediction, and the first bin sits $1.47/1.19 \approx 1.24$ times above it. Mode counting with one effective sky fraction describes the covariance of a masked spectrum only to leading order. What it leaves out (the coupling of neighbouring multipoles through the small

holes, and the amplification of fluctuations when MASTER inverts the coupling, which in the first bin is done together with the buffer bin below it) is exactly what the simulations contain.

Example: Knox's formula against the simulations. In the first bin, $\ell^{\mathrm{eff}} = 47.7$ and $\Delta\ell = 30$, Knox's formula with the effective sky fraction, $\sigma(D_b) \approx \sqrt{2/[(2\ell^{\mathrm{eff}}+1)\Delta\ell\, f_{\mathrm{sky}} w_2^2/w_4]}\; D_b$ (Section 8.4.2), gives 46.3 $\mu\mathrm{K}^2$; the simulations give 53.0 $\mu\mathrm{K}^2$. In the last bin the formula, which ignores the noise, gives 7.2 against 7.9. The formula is a good first guess, and the simulations are the answer.

Are neighbouring bins correlated? The mask couples multipoles, so the bandpowers of neighbouring bins share some of their fluctuations (Section 8.12.2). With bins of 30 multipoles, wide compared with the coupling kernel's core, the correlation is weak: on average -0.016 between neighbours, and at most -0.112 in magnitude in the other direction (Figure 13.10a). MASTER's deconvolution tends to make neighbours slightly *anti*-correlated: when it undoes the coupling, an upward fluctuation that the mask spread into the next bin is subtracted from it.

How good is the covariance itself? Every element of $\mathbf{C}$ is estimated from 400 draws. A diagonal element, a sample variance, has a relative error of $\sqrt{2/(n-1)}$ (Section 8.13.1); the error bar is its square root, so its relative error is half of that, $1/\sqrt{2(n-1)} = 3.5\%$. Our error bars are uncertain by about 3.5%, and the probabilities to exceed computed below are good to a few per cent of their value.

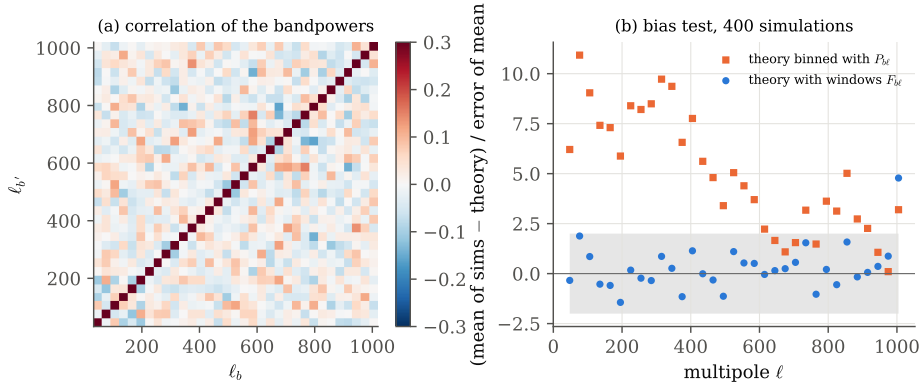


Figure 13.10: (a) The correlation matrix of the 33 bandpowers from the simulations: almost diagonal, with weak correlations between neighbours. (b) The bias test: the average of the simulated bandpowers minus the theory, in units of the error of the average. The theory passed through the exact bandpower windows (blue) is consistent with zero; the theory binned as if it were flat in each bin (orange) is not.

How research codes get the covariance. Here the simulations are the research instrument, not a check of a formula. *Planck*'s high-ℓ likelihood uses an analytic covariance instead, computed from the fiducial spectrum, the masks and a noise model, with corrections for the point-source holes and the anisotropic noise calibrated on simulations [75, §3.1]. Efstathiou showed why this is possible at high ℓ: for apodised masks the pseudo-C_ℓ covariance is close to an analytic expression, and the estimator is nearly as precise as the maximum-likelihood one [134]. The end-to-end FFP10 simulations then validate the analytic covariance rather than replace it. Both routes need the same thing we built: a model of the observation good enough that simulated and real skies are statistically indistinguishable.

13.5 Our spectrum against *Planck*'s

Their question. The *Planck* collaboration wanted the temperature power spectrum of the CMB over $30 \le \ell \le 2508$ with errors small enough that the six parameters of ΛCDM would be limited by the sky, not by the instrument, and a likelihood that turns that spectrum into parameters without bias [75, §1, §3]. The binned spectrum in the file we downloaded is one product of that work: the CMB spectrum after the foreground model has been subtracted and the frequency channels have been averaged.

Their setup, in plain terms. The high-ℓ likelihood code of the release is called Plik. It did not use a component-separated map. It measured cross-spectra between the half-mission maps of three frequency channels, 100, 143 and 217 GHz, each through its own mask (keeping 66%, 57% and 47% of the sky) with apodised Galactic and point-source holes, over a multipole range chosen per channel [75, §3.2.1–3.2.4, Table 13]. The residual emission of dust, of unresolved galaxies and of galaxy clusters was not masked away but *modelled*: each foreground has a spectrum with a few free amplitudes, fitted together with the cosmology. The covariance of the spectra is computed analytically from the fiducial model, the masks and a noise model. Our setup differs in almost every item (Table 13.2), and the point of the comparison is to see whether the two routes land on the same spectrum.

	Planck 2018 (Plik)	ours
maps	100, 143, 217 GHz half-missions	SMICA half-missions
resolution	$N_{\rm side} = 2048$	$N_{\rm side} = 512$
masks	one per frequency, 47–66% of the sky	common mask, 0.775 of the sky
foregrounds	modelled, amplitudes fitted	removed by SMICA, the rest masked
multipoles	$30 \le \ell \le 2508$	$30 \le \ell \le 1019$
covariance	analytic, validated on simulations	400 Gaussian simulations
binning (this file)	$\Delta\ell = 30$, weights $\ell(\ell+1)$	identical

Table 13.2: The two analyses compared. Only the binning is the same.

13.5.1 Their key equation: a Gaussian likelihood for bandpowers

Planck's high-ℓ likelihood is [75, (20)]

$$-\ln \mathcal{L}(\boldsymbol{\theta}) = \tfrac{1}{2}\big[\widehat{\boldsymbol{C}} - \boldsymbol{C}(\boldsymbol{\theta})\big]^{\mathsf{T}} \Sigma^{-1} \big[\widehat{\boldsymbol{C}} - \boldsymbol{C}(\boldsymbol{\theta})\big] + \text{const}, \tag{13.13}$$

with $\widehat{\boldsymbol{C}}$ the vector of measured bandpowers, $\boldsymbol{C}(\boldsymbol{\theta})$ the model passed through the same binning, and Σ their covariance. Where does it come from? Ask the likelihood's question (Section 3.8): for given parameters, how plausible is the fluctuation that turns the model prediction into the bandpowers we observed? Answering it needs the sampling distribution of the bandpowers. Each bandpower is an average over many modes, $(2\ell+1)\,\Delta\ell\, f_{\rm sky} w_2^2/w_4 \approx 2000$ already in our first bin, and an average of many independent pieces is close to Gaussian (the central limit theorem, Section 2.10). Hamimeche and Lewis checked how close: at high ℓ the Gaussian form, with a covariance computed at the fiducial model, gives the same parameters as the exact likelihood [16], while at $\ell \lesssim 30$, where the exact distribution is a skewed scaled χ^2, it does not (Section 9.19.4). That is why *Planck* splits its likelihood at $\ell = 30$, and why we start there. With

a Gaussian sampling distribution,

$$p\left(\widehat{\boldsymbol{D}} \mid \boldsymbol{\theta}\right) \overset{(1)}{=} \frac{1}{\sqrt{(2\pi)^p \det \mathbf{C}}} \exp\left[-\tfrac{1}{2}\left(\widehat{\boldsymbol{D}} - \boldsymbol{t}(\boldsymbol{\theta})\right)^{\mathsf{T}} \mathbf{C}^{-1} \left(\widehat{\boldsymbol{D}} - \boldsymbol{t}(\boldsymbol{\theta})\right)\right],$$

$$-2 \ln \mathcal{L}(\boldsymbol{\theta}) \overset{(2)}{=} \left(\widehat{\boldsymbol{D}} - \boldsymbol{t}(\boldsymbol{\theta})\right)^{\mathsf{T}} \mathbf{C}^{-1} \left(\widehat{\boldsymbol{D}} - \boldsymbol{t}(\boldsymbol{\theta})\right) + \text{const} \equiv \chi^2(\boldsymbol{\theta}) + \text{const}, \tag{13.14}$$

(1) The multivariate Gaussian density of p bandpowers with mean $\boldsymbol{t}(\boldsymbol{\theta})$, the model through the bandpower windows (13.9), and covariance $\mathbf{C}$ (Section 2.12).
(2) Read as a function of $\boldsymbol{\theta}$ with the data fixed; $\mathbf{C}$ is held at its fiducial value, so $\det \mathbf{C}$ is a constant and goes into it.

This is (13.13) in our symbols: $\widehat{C} \to \widehat{\boldsymbol{D}}$ (we work in D_b), $\boldsymbol{C}(\boldsymbol{\theta}) \to \boldsymbol{t}(\boldsymbol{\theta})$, $\Sigma \to \mathbf{C}$. In our analysis $\mathbf{C}$ is the simulated covariance of Section 13.4.3 and its inverse carries the Hartlap factor.

13.5.2 What we compute

Three things, all in `code/ch13/05_compare.py`. First, the value of χ^2 at the *Planck* best fit, $\boldsymbol{t} = \mathcal{F}\boldsymbol{C}^{\text{fid}}$: a goodness-of-fit test of the best-fit spectrum on our bandpowers. Second, the simplest possible fit, one overall amplitude. Third, the difference between our bandpowers and *Planck*'s, bin by bin.

Is the best fit a reasonable explanation? The null hypothesis is that our bandpowers are a draw from the Gaussian of (13.14) with $\boldsymbol{t}$ at the best fit: "the differences are chance", with chance fully specified. Under it the χ^2 is a sum of $p = 33$ squared unit Gaussians (Section 2.13.3), with mean 33 and standard deviation $\sqrt{66} \approx 8$. If the best fit were right, would a χ^2 as large as ours, or larger, be something we would reasonably expect? We find $\chi^2 = 33.5$. The χ^2_{33} tail above it is 0.44, and among the 400 simulations, whose χ^2 is computed in the same way, a fraction 0.30 has a larger one. (The two differ because each simulation is also one of the skies that estimated $\mathbf{C}$, which pulls its own χ^2 down a little; for the data, which are not in that set, the χ^2 tail is the better reference.) Either way the answer is yes: the best fit explains our spectrum as well as it explains a typical simulated sky.

Example: one amplitude. Let the model be $\boldsymbol{t} = A\,\boldsymbol{t}^{\text{fid}}$, the best-fit spectrum scaled by an unknown factor A. With Ψ the Hartlap-corrected inverse covariance, the χ^2 is a parabola in A,

$$\chi^2(A) \overset{(1)}{=} \left(\widehat{\boldsymbol{D}} - A\boldsymbol{t}^{\text{fid}}\right)^{\mathsf{T}} \Psi \left(\widehat{\boldsymbol{D}} - A\boldsymbol{t}^{\text{fid}}\right) \overset{(2)}{=} \widehat{\boldsymbol{D}}^{\mathsf{T}} \Psi \widehat{\boldsymbol{D}} - 2A\, \boldsymbol{t}^{\text{fid}\,\mathsf{T}} \Psi \widehat{\boldsymbol{D}} + A^2\, \boldsymbol{t}^{\text{fid}\,\mathsf{T}} \Psi \boldsymbol{t}^{\text{fid}},$$

$$\frac{\mathrm{d}\chi^2}{\mathrm{d}A} = 0 \overset{(3)}{=} \hat{A} = \frac{\boldsymbol{t}^{\text{fid}\,\mathsf{T}} \Psi \widehat{\boldsymbol{D}}}{\boldsymbol{t}^{\text{fid}\,\mathsf{T}} \Psi \boldsymbol{t}^{\text{fid}}}, \qquad \sigma_A \overset{(4)}{=} \left(\boldsymbol{t}^{\text{fid}\,\mathsf{T}} \Psi \boldsymbol{t}^{\text{fid}}\right)^{-1/2}. \tag{13.15}$$

(1) (13.14) with $\boldsymbol{t} = A\boldsymbol{t}^{\text{fid}}$.
(2) Multiply out; Ψ is symmetric, so the two cross terms are equal.
(3) Set the derivative $-2\boldsymbol{t}^{\text{fid}\,\mathsf{T}} \Psi \widehat{\boldsymbol{D}} + 2A\, \boldsymbol{t}^{\text{fid}\,\mathsf{T}} \Psi \boldsymbol{t}^{\text{fid}}$ to zero.
(4) $\hat{A}$ is linear in the Gaussian data, so its variance is $\boldsymbol{a}^{\mathsf{T}} \mathbf{C} \boldsymbol{a}$ with $\boldsymbol{a} = \Psi \boldsymbol{t}^{\text{fid}} / (\boldsymbol{t}^{\text{fid}\,\mathsf{T}} \Psi \boldsymbol{t}^{\text{fid}})$; with $\Psi\mathbf{C} \approx \mathbb{1}$ this is $1/\boldsymbol{t}^{\text{fid}\,\mathsf{T}} \Psi \boldsymbol{t}^{\text{fid}}$, the inverse of the curvature of $\chi^2/2$, as for every linear-Gaussian model (Eq. (8.80)).

On the SMICA bandpowers, $\hat{A} = 0.9992 \pm 0.0017$. The overall power of the CMB between $\ell = 30$ and 1019 agrees with the best fit to 0.1%, with an error of 0.2%. That size can be checked by counting. The number of independent modes behind the bandpowers is $f_{\text{sky}} w_2^2 / w_4 \sum_{\ell=30}^{1019} (2\ell + 1) = 0.708 \times (1020^2 - 30^2) \approx 7 \times 10^5$. Each mode is one squared Gaussian

number, whose relative scatter is $\sqrt{2}$ (Eq. (7.77)), so the average over N modes has relative error $\sqrt{2/N} = 0.17\%$, which is what we find. It is the same estimator that Section 8.13.4 used on simulated skies, now with one real sky.

13.5.3 Side by side

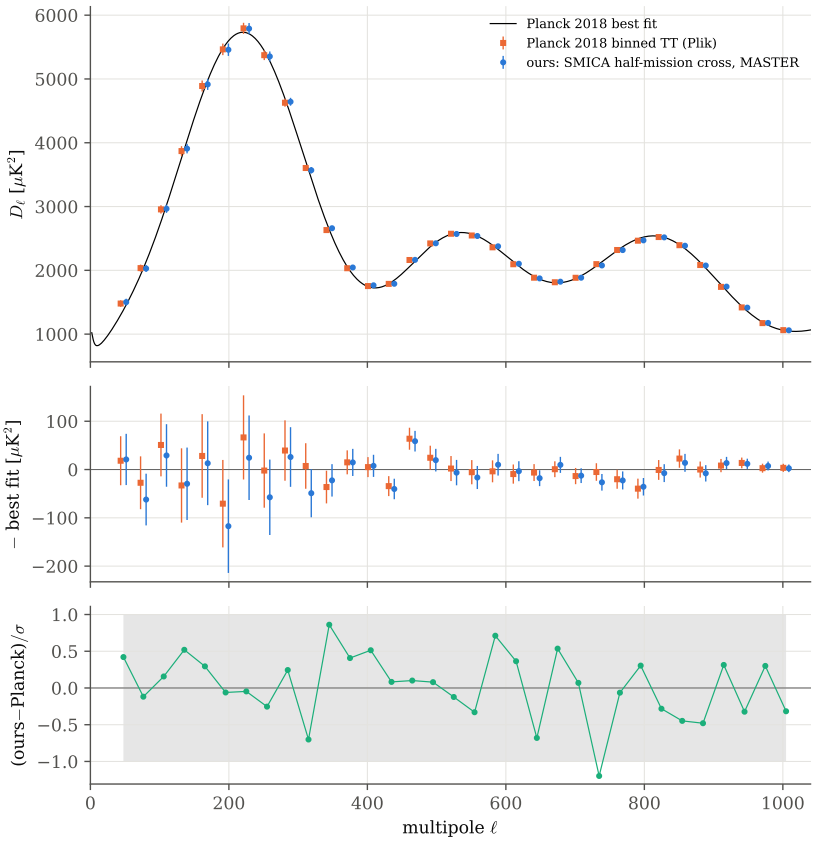


Figure 13.11: Top: our MASTER bandpowers of the SMICA half-mission cross-spectrum (blue) and *Planck*'s binned TT spectrum (orange), shifted by ±4 in ℓ to be seen side by side, with the best fit (black). Middle: both minus the best fit; each with its own error bars. Bottom: their difference in units of our error bar; the grey band is ±1.

Figure 13.11 puts the two spectra on one plot. Over the 33 bins they differ by 15 μK^2 in the root mean square, 0.43 of *Planck*'s error bar, and no bin differs by more than $|-1.2|$ of ours (at $\ell \approx 735$). The error bars themselves are almost equal: ours are 0.99 of *Planck*'s (median over the bins). An overall calibration difference would show up as a ratio different from one; the inverse-variance weighted mean of our bandpower divided by theirs is 0.9998.

13.5.4 Why they differ, and by how much they should

Two measurements of the *same* sky do not scatter around each other like two independent skies. Most of the error bar of each is cosmic variance, the luck of the particular sky, and the two share it. So the question "are 15 μK^2 differences reasonable?" needs the scatter of the *difference* of two estimators on one sky, and we can measure that with our simulations. Take the main window and the window with $|b| > 30°$, a part of the same sky, and compute both on each simulated sky. Their bandpowers are correlated with coefficient 0.81, and their difference scatters by 0.74 of the main error bar. The reason is a simple identity. Nested

regions share their sky, so write the smaller-sky estimate as the larger-sky estimate plus an extra piece, $A_{\text{small}} = A_{\text{large}} + E$, where E is what the less informative estimate adds, independent of A_{large} (the larger-sky estimate is the more precise one, and that is what makes the extra piece uncorrelated with it). The variance of a sum of independent pieces is the sum of the variances, $\sigma^2_{\text{small}} = \sigma^2_{\text{large}} + \text{Var}(E)$, and the difference $A_{\text{small}} - A_{\text{large}} = E$ has variance $\sigma^2_{\text{small}} - \sigma^2_{\text{large}}$. So it scatters by $\sqrt{(\sigma_{\text{small}}/\sigma_{\text{large}})^2 - 1} = \sqrt{1.26^2 - 1} \approx 0.77$ of the larger-sky error bar, close to the measured 0.74.

Planck's analysis uses a sky that overlaps ours far more than the $|b| > 30^\circ$ cut does (its masks keep 47–66% of the sky, almost all of it inside our 0.775), so the expected scatter of our difference is smaller than 0.74 of an error bar. We measure 0.43. The rest of the difference has identifiable causes, each small:

- *Different maps.* SMICA cleans foregrounds by weighting frequencies, Plik by modelling them. Where the foreground model and SMICA's weights disagree, the spectra differ; this is largest at high ℓ, where unresolved galaxies add power. Above $\ell = 700$ our bandpowers are on average $-3 \pm 4\,\mu\text{K}^2$ from *Planck*'s: no sign of excess point-source power in SMICA at these multipoles.
- *Different weights on the sky.* Different masks, and inside them different weights per frequency, sample the sky's fluctuations differently, which is the part estimated above.
- *Different best fits.* Our fiducial is the TT,TE,EE+lowE+lensing best fit, theirs in the file is their own; through our windows the two differ by up to 1.4%, which matters for the middle panel of Figure 13.11 but not for the difference of the two measurements.

What reproducing it taught us. A component-separated map, a common mask, our own coupling matrix and four hundred simulated skies at a sixteenth of the resolution give *Planck*'s temperature spectrum to a fraction of its error bar and its amplitude to 0.2%. What the comparison could not show by itself is how small the difference *should* be; for that we needed the joint scatter of two estimators on one sky, which only the simulations provide.

13.6 Is anything else in the map? The checks

The comparison with *Planck* told us that our spectrum agrees with theirs. It did not tell us that either spectrum is the spectrum of the CMB alone. Both could share a residual of the Galaxy, both are built from the same detectors, and a successful comparison between two analyses of one data set is weak evidence about what is in the data. Recall the three candidate explanations of the map listed in Section 13.1: a Gaussian isotropic CMB plus noise; residual foregrounds; something the processing did. The checks are how we put the second and third candidates on trial.

The idea of a null test. Choose a quantity that the first explanation predicts to be zero on average, and that the others would push away from zero. Spectra measured through two different windows are the standard example. If the map outside the mask is a statistically isotropic CMB, the spectrum through the northern half and the spectrum through the southern half have the same mean; their difference has mean zero. A residual of the Galaxy that is brighter in one hemisphere, or a stripe left by the scanning, would make it nonzero. The test then asks the question of every hypothesis test (Section 4.1.1): if the map were only CMB and noise, would a difference as large as the one we see, or larger, be something we would reasonably expect?

Why simulations, and not a formula. The answer needs the sampling distribution of the difference. The obvious guess, that the difference of two bandpowers has variance $\sigma_A^2 + \sigma_B^2$, is badly wrong whenever the two windows see the same sky. Take the binary common mask and the apodised window: almost the same pixels, slightly different weights. The simulations give a scatter of their difference of only 0.43 of a single bandpower's error, while $\sqrt{\sigma_A^2 + \sigma_B^2}$ would be about $\sqrt{2} = 1.4$ of it, three times too large, and a test built on it could never fail. The difference has the variance $\sigma_A^2 + \sigma_B^2 - 2\,\mathrm{Cov}(A, B)$, and the covariance of two estimators on one sky is exactly what our simulations contain, because each simulated sky is analysed through all five windows. So every test uses the same recipe:

1. the difference vector $\boldsymbol{\delta}$ of the data, over the 33 bins;
2. the same difference on each of the 400 simulated skies, giving its mean $\bar{\boldsymbol{\delta}}$ and covariance $\mathbf{C}_\delta$;
3. $\chi^2 = (\boldsymbol{\delta} - \bar{\boldsymbol{\delta}})^{\mathsf{T}} \Psi_\delta (\boldsymbol{\delta} - \bar{\boldsymbol{\delta}})$, with Ψ_δ the Hartlap-corrected inverse of $\mathbf{C}_\delta$;
4. the probability to exceed (PTE): the fraction of simulated skies whose χ^2, computed in the same way, is at least as large as the data's. For comparison, also the tail of χ^2_{33} above the data's value.

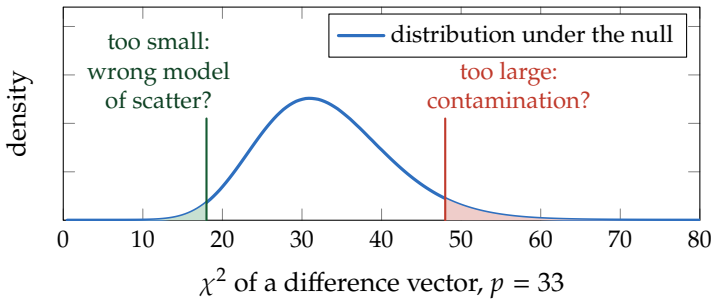


Figure 13.12: Reading a null test. Under the null hypothesis (CMB and our noise model, nothing else) the χ^2 of a difference vector of 33 bandpowers follows the blue curve, which the simulations sample. A value in the right tail (red) means the difference is larger than chance allows: something in the data is not in the model. A value in the left tail (green) means it is *smaller* than chance allows: the model of the scatter is wrong, because contamination can only add to a difference.

13.6.1 Four tests, and what each is for

difference	what would make it nonzero	χ^2 ($p = 33$)	PTE (sims)	PTE (χ^2_{33})
north − south	anisotropy, a one-sided residual	13.2	1.00	1.00
$\lvert b\rvert > 30°$ − main	Galactic residuals near the plane	27.5	0.59	0.74
binary − apodised	an error in the mode-coupling correction	14.7	1.00	1.00
half-mission null	a wrong noise model, sky in the difference	1312.3	0.00	0.00

Table 13.3: The checks. Each line is a difference that a Gaussian CMB plus our noise model predicts to vanish on average; the PTE is the fraction of simulated skies with a larger χ^2, and the tail probability of a χ^2 with 33 degrees of freedom.

The Galactic test passes. The window with $|b| > 30°$ removes everything within 30° of the plane, where dust and synchrotron residuals would be strongest. Its spectrum differs from the main one by $\chi^2 = 27.5$ for 33 bins, a PTE of 0.59 (Figure 13.13b). How small a residual

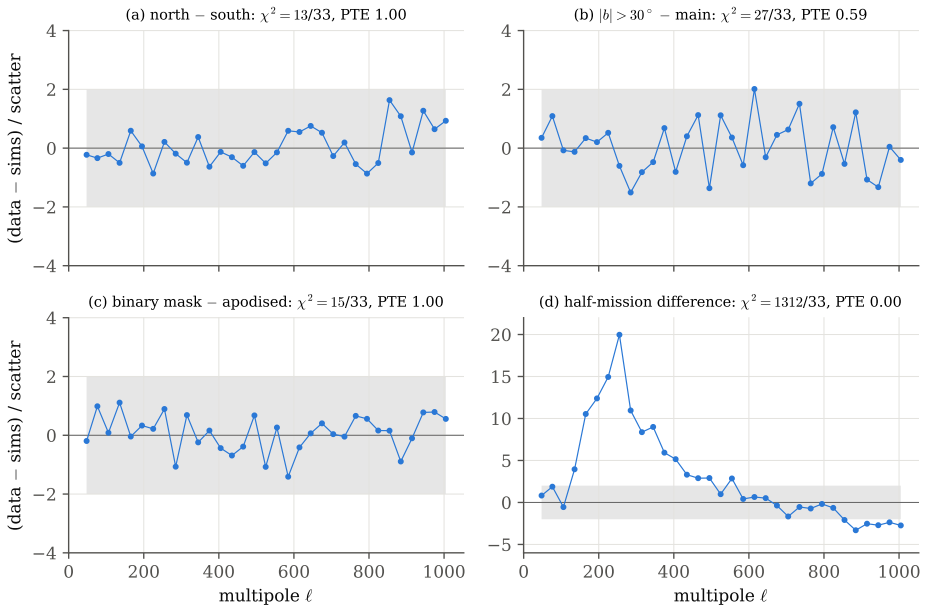


Figure 13.13: The four checks, bin by bin: the data's difference minus the average simulated difference, in units of the simulated scatter. The grey band is ±2. (a)–(c) compare spectra through different windows; (d) is the spectrum of the half-mission difference map against the simulated noise.

could it have seen? The scatter of this difference is 0.74 of a main error bar, so a residual that changed the main spectrum by a full error bar in several bins would have been detected; one much smaller than that could hide.

The half-mission null fails, and we can say why. Here the result is unambiguous: $\chi^2 = 1312.3$ for 33 bins, with single bins 20.0 standard deviations off (Figure 13.13d). The event: the spectrum of the half-difference map is up to 1.31 times the simulated one, at $\ell \approx 255$, and about 1% *below* it at high ℓ (on average 0.989 of it above $\ell = 600$). What could have produced it? We take the candidates one at a time.

1. *Our noise model has the wrong shape in ℓ.* It was built as a modulated isotropic field (13.10), with a spectrum averaged over 31 multipoles. If the real noise has a different spectrum in different parts of the sky, as SMICA's noise does, the product of one map v and one spectrum cannot reproduce it. This predicts a mismatch that varies slowly with ℓ, with either sign.
2. *Some sky survives in the difference.* If the two halves see the sky with slightly different gains, $1 + \epsilon/2$ and $1 - \epsilon/2$ (a calibration difference, or a slightly different effective beam), the difference map contains $\epsilon/2$ times the sky, and its spectrum contains $(\epsilon/2)^2$ times the sky's spectrum. This predicts an excess that follows the acoustic peaks, largest at the first one.
3. *The noise of the halves is correlated.* Correlated noise would *cancel* in the difference and lower its spectrum, not raise it. It cannot produce an excess.

How well does each explain the data? The excess peaks at $\ell \approx 255$, the first acoustic peak, which is what the second candidate predicts. Fitting the excess below $\ell = 600$ as $(\epsilon/2)^2$ times the fiducial bandpowers plus a constant gives $\epsilon = 1.2\%$, and explains 35% of the variance of the excess; the rest, and the deficit at high ℓ, is the first candidate. So both are at work: a percent-level difference between the halves' response to the sky, and a noise model whose shape is good to a few per cent at high ℓ and worse at low ℓ.

The test is this sensitive because the noise is measured with enormous precision: thousands of modes in each bin, and a simulated scatter of about 2% of the noise power. Does the failure matter for the cross-spectrum? Neither cause biases it. A gain difference changes the cross-spectrum by the factor $(1+\epsilon/2)(1-\epsilon/2) = 1-\epsilon^2/4$, a change of 0.0039%. A noise model wrong by 30% changes only the noise part of the error bars, and the noise is 0.15% of the signal at $\ell \approx 500$ and 1.9% at $\ell \approx 1000$. The failure is real, it is understood, and it is harmless for the questions we are asking; a likelihood that used the auto-spectra, where the noise must be subtracted, could not ignore it. (The identity (13.3) between the full-mission, the cross and the null spectra holds on the data to 11.41 $\mu\mathrm{K}^2$ at worst, not exactly, for the same reason: the SMICA full-mission map is not exactly the average of the two halves.)

Two tests are passed too well. The hemisphere test gives $\chi^2 = 13.2$ and the apodisation test $\chi^2 = 14.7$, for 33 bins each. Only 1 and 1 of the 400 simulated skies give smaller values (Table 13.3, where a PTE of 1.00 means "larger than in almost every simulation"). The data's differences are, bin by bin, 0.62 and 0.62 of the simulated scatter in the root mean square, against about one expected (and 0.89 for the Galactic test). What could produce two differences this small?

1. *Chance.* So few simulated skies fall this low that their count only bounds the chance: a χ^2 this low happens in at most about 0.01 of skies (95% upper limit). In the simulations the two tests are independent (the correlation of their χ^2 over the simulated skies is -0.05), so two of four tests this low has a probability below about 9×10^{-4}. Not plausible.
2. *A residual foreground.* No: contamination adds to a difference, it cannot remove one (Figure 13.12).
3. *The simulations overestimate the scatter of these differences.* The real map's spectra through different windows agree better than Gaussian skies passed through our model of the observation would. Our simulations lack two things the real map has been through: SMICA's component separation, and the inpainting that fills the holes (the band limit of the degrade spreads the filled values a few arcminutes into the kept pixels at the hole edges, exactly where the binary and apodised windows differ).

The third explanation is the likely one, and it limits what the simulations can certify: they describe the scatter of each window's spectrum well (the best-fit χ^2 of Section 13.5.2 is normal) but overestimate the scatter of the differences between windows. A low χ^2 cannot hide a contaminant, so the spectrum itself is not in question. What the full analysis would do is run the null tests on end-to-end simulations processed exactly as the data were, component separation and inpainting included, which our Gaussian skies are not.

Example: reading a PTE of 1.00 **correctly.** A PTE close to one is not "a perfect pass". The PTE is uniform between 0 and 1 when the null hypothesis and its simulated distribution are right (Section 4.4.2), so a value above 0.99 is as unusual as one below 0.01. A table of null tests whose PTEs cluster near one is a warning that the error model is too generous, just as a cluster near zero is a warning of contamination.

13.7 Two primordial numbers from the real sky

In Section 9.22 the Metropolis sampler of Chapter 9 was pointed at a mock *Planck*-like spectrum: $2 \le \ell \le 2500$, one 143 GHz-like channel, $f_{\mathrm{sky}} = 0.57$. With the amplitude and the tilt of the primordial spectrum free and everything else fixed, it returned errors of 0.0018 on $\ln(10^{10}A_s)$

and 0.0026 on n_s, with a correlation of -0.73; freeing the optical depth τ inflated the first error ninefold, to 0.0164, because the temperature spectrum above $\ell = 30$ measures only the combination $A_s e^{-2\tau}$ (9.132). Those were forecasts on a spectrum whose truth we chose. Now the same two parameters, the same emulator and the same sampler meet our bandpowers of the real sky.

The event, and the hypotheses. The observed event is our vector of 33 bandpowers $\widehat{\boldsymbol{D}}$. The hypotheses are the pairs $\boldsymbol{\theta} = (\ln 10^{10} A_s,\, n_s)$, with the baryon and dark-matter densities, the angular size of the sound horizon and τ held at the *Planck* best fit. For each pair the question is: had nature chosen these values, how plausible would the fluctuation be that turns the predicted bandpowers into the ones we measured? That is the likelihood of (13.14), and Bayes' theorem with a flat prior over $2.8 < \ln 10^{10} A_s < 3.3$, $0.85 < n_s < 1.10$ turns it into a posterior.

13.7.1 The model, the likelihood and the sampler

The model. The predicted bandpowers are $\boldsymbol{t}(\boldsymbol{\theta}) = \mathcal{F}\, \boldsymbol{C}(\boldsymbol{\theta})$: the theory spectrum passed through the bandpower windows (13.9). The spectrum $C_\ell(\boldsymbol{\theta})$ comes from the emulator of Section 9.20, a second-order Taylor expansion of $\ln C_\ell$ about the fiducial (9.123), which reproduces CAMB to far better than our error bars over this range (Section 9.20.3). A CAMB call takes seconds; the emulator takes microseconds, so a chain of tens of thousands of steps costs seconds.

The likelihood. Gaussian in the bandpowers, (13.14), with the simulated covariance and the Hartlap factor. The covariance is held fixed at the fiducial model, as *Planck* does; over the narrow range of parameters the data allow, the change of the covariance with the parameters is negligible [16].

The sampler. Random-walk Metropolis (Section 9.8), `code/ch09/lib_mcmc.py`, with a Gaussian proposal whose shape is the inverse Fisher matrix $\mathbf{F}^{-1}$ and whose scale is $2.38/\sqrt{d}$ for $d = 2$ parameters, the rule of Section 9.11. Four chains of 20 000 steps start from points scattered over three Fisher widths around the fiducial; the first 20% of each is discarded as burn-in (Section 9.12.1).

Example: the Fisher matrix as a first answer. For a model that is linear in the parameters near the peak, the posterior is Gaussian with covariance $\mathbf{F}^{-1}$, where $\mathbf{F} = J^{\mathsf{T}} \Psi J$ and $J_{bi} = \partial t_b / \partial \theta_i$ (Chapter 10). We compute J by central differences of the emulator. The Fisher errors are 0.0018 on $\ln 10^{10} A_s$ and 0.0039 on n_s. The chains give 0.0018 and 0.0039. Agreement to within a few per cent says the posterior is Gaussian to that precision: over a range of $\pm 3\sigma$ the bandpowers depend on the two parameters almost linearly.

Listing 13.5: The log-posterior: a flat prior and the Gaussian bandpower likelihood (`code/ch13/07_fit.py`)

```
def make_logpost(Fmat, data, prec):
    def logpost(th):
        if np.any(th < LO) or np.any(th > HI):
            return -np.inf
        r = data - Fmat @ emu.cl(th)[: Fmat.shape[1]]
        return -0.5 * r @ prec @ r
    return logpost
```

The function returned by `make_logpost` is all the sampler sees. Outside the prior box it returns $-\infty$, so every proposal there is rejected. Inside, it computes the residual between the bandpowers and the emulated theory through the windows, and returns $-\frac{1}{2}\chi^2$; the normalisation of the posterior never appears, because Metropolis only uses ratios (Section 9.8.1).

13.7.2 What the real sky says

The chains are healthy: acceptance 0.36, integrated autocorrelation time about 7 steps (Section 9.12.2), so the 64 000 kept samples are worth several thousand independent ones, and the split $\hat{R}$ of the four chains is 1.0009 (Section 9.13). The posterior means and standard deviations are

$$\ln(10^{10}A_s) = 3.0432 \pm 0.0018, \qquad n_s = 0.9686 \pm 0.0039, \qquad \rho = -0.28, \tag{13.16}$$

and at the maximum of the likelihood $\chi^2 = 32.4$ for $33 - 2 = 31$ degrees of freedom (a tail probability of 0.40): two parameters describe the 33 bandpowers as well as they should. With $\tau = 0.0544$ fixed, the quantity the temperature spectrum actually measures is $10^9 A_s e^{-2\tau} = 1.8811 \pm 0.0034$.

To separate what comes from *our* bandpowers from what comes from the multipole range, we run the same fit, same emulator, same sampler, on *Planck*'s published bandpowers, with their quoted errors as a diagonal covariance (the file gives no correlations): first over our range $30 \le \ell \le 1019$, then over $30 \le \ell \le 2489$, as far as our emulator goes.

data, multipoles, free parameters	$\ln(10^{10}A_s)$	n_s	ρ
ours (SMICA), 30–1019, 2 free	3.0432 ± 0.0018	0.9686 ± 0.0039	-0.28
Planck bandpowers, 30–1019, 2 free	3.0453 ± 0.0018	0.9661 ± 0.0040	-0.22
Planck bandpowers, 30–2489, 2 free	3.0451 ± 0.0017	0.9657 ± 0.0025	-0.73
mock spectrum of Section 9.22, 2–2500, 2 free	± 0.0018	± 0.0026	-0.73
Planck 2018 VI, TT+lowE, 6 free	3.040 ± 0.016	0.9626 ± 0.0057	

Table 13.4: The amplitude and tilt from our bandpowers, from *Planck*'s bandpowers through the same fit, from the forecast of Section 9.22, and as published with all six parameters free [25].

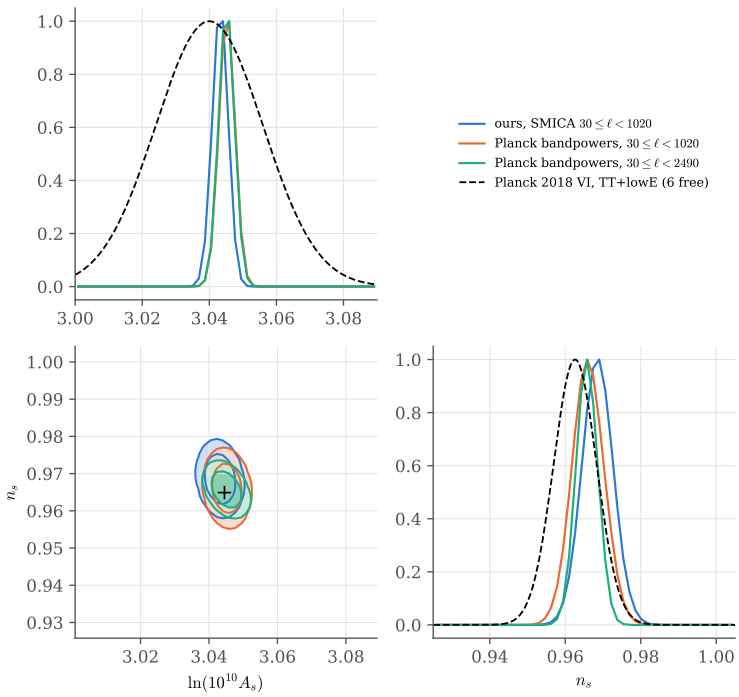


Figure 13.14: Posteriors of $\ln(10^{10}A_s)$ and n_s with the other four parameters fixed: from our SMICA bandpowers (blue), from *Planck*'s bandpowers over the same multipoles (orange) and over $30 \le \ell \le 2489$ (green). Contours enclose 68% and 95%. The black dashed curves are the published one-dimensional results with all six parameters free; the cross is the fiducial.

13.7.3 Reading the comparison with *Planck* 2018 VI

Their question and setup. *Planck* 2018 VI asked which ΛCDM parameters the CMB spectra prefer. Its "TT+lowE" result combines the high-ℓ temperature likelihood of (13.13), a low-ℓ temperature likelihood for $2 \le \ell < 30$, and the large-scale E-mode polarisation, which measures τ. All six parameters are free, together with nuisance parameters for the foregrounds and the calibration [25]. Its key equation is the same Gaussian likelihood we derived in Section 13.5.1, multiplied by the low-ℓ terms.

Side by side, and why the numbers differ. Figure 13.14 and Table 13.4 show three differences, each with a cause we can name.

1. *Our error on the amplitude is nine times smaller.* Above $\ell = 30$ the temperature spectrum depends on A_s and τ only through $A_s e^{-2\tau}$, so $\ln(10^{10}A_s) = \ln(10^{10}A_s e^{-2\tau}) + 2\tau$ and, the two terms being nearly independent, $\sigma^2(\ln 10^{10}A_s) \approx \sigma^2(\ln 10^{10}A_s e^{-2\tau}) + 4\sigma_\tau^2$ (9.132). We fixed τ, so the second term is zero and we quote the error of the first, 0.0018. *Planck* measures $\sigma_\tau = 0.008$ from polarisation, and $2 \times 0.008 = 0.016$ is its whole error on $\ln(10^{10}A_s)$. The well-measured combination agrees: $10^9 A_s e^{-2\tau} = 1.8811 \pm 0.0034$ for us, 1.884 ± 0.014 published.

2. *Our error on n_s is smaller than theirs although we use two fifths of the multipoles.* With six parameters free, a change of the tilt can be partly undone by a change of the baryon density, which also tilts the relative heights of the peaks; fixing the other parameters removes that escape route. Freeing them costs a factor of about two in Section 9.22 (0.0026 to 0.0053), the same factor that separates our full-range fit of *Planck*'s bandpowers (0.0025) from the published 0.0057.

3. *The multipole range matters for the tilt.* On *Planck*'s bandpowers, going from $\ell \le 1019$ to $\ell \le 2489$ shrinks the error on n_s from 0.0040 to 0.0025 and strengthens the anticorrelation with the amplitude from −0.22 to −0.73. A tilt is a lever, and the damping tail lengthens the arm; on a longer arm, raising the amplitude and lowering the tilt compensate each other over more of the range, hence the stronger correlation. The full-range numbers reproduce the forecast of Section 9.22 (0.0018, 0.0026, −0.73): that forecast described the real data correctly.

The central values agree within the errors. Our $n_s = 0.9686$ sits one standard deviation above the fiducial 0.9649, and 0.6σ above the fit to *Planck*'s bandpowers over the same range; the bin-by-bin differences of Section 13.5.3, a few tenths of an error bar each, are enough to move a two-parameter fit that much.

13.7.4 Do the low and the high multipoles agree?

A fit to a range of multipoles averages over it. If the low and the high multipoles wanted different parameters, the average would hide it, and *Planck* took this question seriously: in their temperature data some parameters differ by more than 2σ between $\ell < 800$ and $\ell > 800$, which they judge, after accounting for correlations, not to be especially unusual [25, p. 32]. We ask the same of our data, splitting at $\ell = 510$ so that both halves have similar constraining power.

The test statistic. Fit $(\ln 10^{10}A_s, n_s)$ separately to the bins below and above the split, and form the shift $\boldsymbol{\Delta} = \hat{\boldsymbol{\theta}}_{\text{low}} - \hat{\boldsymbol{\theta}}_{\text{high}}$. If both halves measure the same sky with the same model, $\boldsymbol{\Delta}$ has mean zero. Its covariance is not the sum of the two fits' covariances, since the two halves share

the mask and are weakly correlated, so we take it from the simulations: the same two fits on each simulated sky give 400 shifts and their covariance $\mathbf{C}_\Delta$. Then $\chi^2 = \boldsymbol{\Delta}^\mathsf{T}\mathbf{C}_\Delta^{-1}\boldsymbol{\Delta}$ has two degrees of freedom.

The result. The low multipoles prefer (3.051, 0.975) and the high ones (3.037, 0.986). The shift is $\Delta \ln 10^{10}A_s = +0.015 \pm 0.007$ and $\Delta n_s = -0.011 \pm 0.014$, and $\chi^2 = 4.49$. A fraction 0.11 of the simulated skies gives a larger one, the same as the χ^2_2 tail (0.11; Figure 13.15). The amplitude shift alone is a 2σ effect, but one of two correlated numbers at 2σ is what the two-dimensional test is for: it is a one-in-nine event, and the two halves agree.

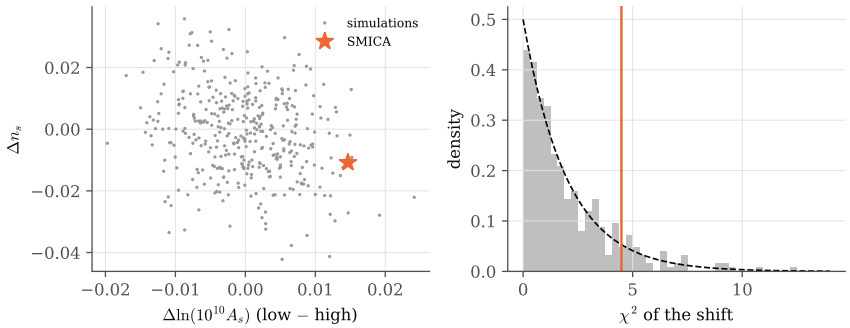


Figure 13.15: Low against high multipoles. Left: the shift between the parameters fitted to $\ell < 510$ and to $\ell > 510$ on each simulated sky (grey) and on the SMICA bandpowers (star). Right: the χ^2 of the shift over the simulations (histogram), the χ^2_2 density (dashed), and the data (vertical line).

Watch out: A fixed parameter is a prior of zero width

Holding four parameters at their best-fit values is the same as giving them delta-function priors. The errors in (13.16) are therefore *conditional* errors, smaller than the marginal ones of a six-parameter fit, and the central values inherit any error in the fixed values. Our ± 0.0018 on $\ln(10^{10}A_s)$ is not a better measurement than *Planck*'s ± 0.016; it is the answer to a narrower question.

13.8 The full analysis, and what the laptop version gives up

Everything above ran on four processor cores: the preparation of the maps in under a minute, each simulated sky in 1.7 s, the four hundred of them in about ten minutes, the fits in seconds. *Planck*'s analysis of the same sky used supercomputers for years. Each place where we took a shorter road costs something, and an honest result says what. Table 13.5 lists the reductions, and the paragraphs after it put a size on the ones that matter.

The multipole range. The largest single cost. Fitting *Planck*'s own bandpowers with our two-parameter model, the error on n_s is 0.0040 over our range and 0.0025 over theirs (Section 13.7.3): the damping tail is worth a factor 1.6 on the tilt. Going to $N_{\mathrm{side}} = 1024$, whose degraded maps already exist in `data/planck/`, would reach $\ell \approx 2000$ at eight times the cost per transform.

The simulations. *Planck*'s FFP10 simulations follow the photons through the instrument: time-ordered data of every detector, the map-maker, and the component separation. Ours start at the map and assume a Gaussian sky and a modulated Gaussian noise. For the error bars of the spectrum itself the difference is small, because below $\ell \approx 1000$ the error is cosmic variance,

step	full analysis (*Planck* 2018)	ours	what the reduction costs
maps	100, 143, 217 GHz half-missions, $N_{\rm side} = 2048$	SMICA half-missions degraded to $N_{\rm side} = 512$	multipoles above 1019; foreground uncertainty not propagated
multipoles	30–2508 (plus a separate 2–29 likelihood)	30–1019	error on n_s 0.0040 instead of 0.0025 on the same bandpowers
masks	per frequency, 47–66% of the sky	one common mask, 0.775	more exposure to residuals near the plane
foregrounds	modelled, amplitudes marginalised	cleaned by SMICA, then ignored	residuals at high ℓ unmodelled
beams	beam matrices per cross-spectrum, with leakage	one effective beam	a beam error would tilt n_s
covariance	analytic, validated on FFP10 end-to-end simulations	400 Gaussian skies, noise model from the data	errors uncertain by 3.5%; null and window-difference tests not validated
parameters	six, plus nuisance parameters, τ from polarisation	two, four fixed	conditional, not marginal, errors

Table 13.5: The full analysis and the laptop analysis, step by step.

which a Gaussian sky gets right; our errors agree with *Planck*'s to within a few per cent (0.99 of theirs, Section 13.5.3). For the checks it is not small: the half-mission null test fails because our noise model is too simple, and the window-difference tests come out too good, most likely because the real map has been through processing that our skies have not (Section 13.6).

The number of simulations. With $n = 400$ skies and $p = 33$ bins, the Hartlap factor is 0.915 and the error bars are known to 3.5%. Doubling n would reduce that to 2.5% at twice the cost; the gain matters for PTEs near the edges of the distribution, where $1/n$ sets the resolution, as in the two window-difference tests, which only 1 and 1 of the 400 simulated skies beat: too few to say how rare they are, only that they are rare.

How this is used in real research: Verification and research in one pipeline

Two different jobs were done by the same simulations. The bias test of Section 13.4.3 *verified* a known result: MASTER's mean is the theory through the bandpower windows, a statement derived in Section 8.11, and the simulations confirmed it to $1/\sqrt{400}$ of an error bar. Everything else was *research*: the covariance of masked, noisy, beamed bandpowers, the scatter of the difference between two windows on one sky, the distribution of a parameter shift between low and high multipoles. None of these has a formula we trust at the precision needed, and each was read off the simulations directly. Research codes package the same steps: NaMaster computes the coupling matrix, the binned inverse and the bandpower windows for spin-0 and spin-2 fields on any mask [169], and *Planck*'s likelihood code evaluates (13.13) with its analytic covariance. What no library supplies is the model of the observation, which here was ours to build and ours to test.

Every arrow of the pipeline, on real data. Each arrow of physical model → field → what matters → summary → its sampling distribution → inference was crossed on real data. The physical model was ΛCDM with two free numbers. The field was the SMICA map. What

mattered was the two-point statistics of a Gaussian field, so the summary was the binned power spectrum. Its sampling distribution came from four hundred skies observed like the real one. The inference was a posterior for A_s and n_s, and a set of tests that the data passed, failed, or passed suspiciously well. The same map returns in Section 13.9, where the summary changes from the spectrum to the shapes of the hot and cold regions, and with it the information it can see.

Recap

- A map is changed in resolution in harmonic space: multiply the $a_{\ell m}$ by the ratio of pixel windows and truncate at $3N_{\rm side} - 1$, after filling the masked holes so that the band limit cannot spread their residuals.
- The common mask has a few large holes and thousands of small ones; they are tapered differently (0.5° and 2°) because a wide taper around thousands of holes would cost most of the sky.
- The cross-spectrum of the half-mission maps has no noise bias; their half-difference is a null map of the noise.
- *Planck*'s bins weight C_ℓ by $\ell(\ell+1)$ and quote D_b at $\ell^{\rm eff}$; a bandpower is not D_ℓ at $\ell^{\rm eff}$, and theory must go through the same binning and the bandpower windows.
- On the real sky the naive correction is off by 10% in the troughs; MASTER with our coupling matrix agrees with *Planck*'s bandpowers to 1.5% at worst.
- Four hundred Gaussian skies observed like the real one give the covariance, an unbiasedness test, and the scatter of every difference used by the checks.
- The best fit describes our bandpowers ($\chi^2 = 33.5$ for 33), the amplitude agrees to 0.2%, and our spectrum differs from *Planck*'s by 0.43 of an error bar.
- Null tests: the Galactic test passes; the half-mission null fails because of a percent-level gain difference between the halves and an approximate noise model, neither of which affects the cross-spectrum; two window-difference tests are passed too well, a sign that our simulations overestimate the scatter of differences.
- With four parameters fixed, $\ln(10^{10}A_s) = 3.0432 \pm 0.0018$ and $n_s = 0.9686 \pm 0.0039$; the comparison with *Planck*'s six-parameter result is explained by τ, by the fixed parameters and by the multipole range.

13.9 The shape of the one sky we have

Twice already we have watched the microwave sky's shapes being tested, and both times the sky itself stayed off stage. In Section T3.5.1 we read how the *Planck* collaboration measured the area, the boundary length and the Euler characteristic of the hot and cold regions of their maps at twelve thresholds, stacked the 36 numbers into one vector, and compared it through a χ^2 with 999 end-to-end simulations; every temperature map at every resolution came out unremarkable [114]. In Section 12.7.4 we rebuilt the Gaussian half of that test: 300 simulated Gaussian skies at $N_{\rm side} = 256$ with a 40′ beam reproduced the curves of their Fig. 10, and the χ^2 of one Gaussian sky against the others came out at 1.01 per degree of freedom, as it should. What was missing both times is the step that makes it a measurement: putting the real map through the same code. The estimator, the thresholds and the normalisation stay exactly as they were. What changes is everything that comes with real data: the sky is cut by a mask with hundreds of holes, the map carries instrumental noise, the region behind the mask has been

filled in by the component-separation pipeline, and there is only one map.

The event. We point the pipeline at the *Planck* SMICA temperature map. At each of four resolutions it returns 36 numbers: the fraction of the observed sky above each of twelve thresholds, the length of the boundaries there, and the number of hot regions minus the number of holes in them. That vector is what happened. Call it $\boldsymbol{y}_{\text{obs}}$.

What could have produced it? Before computing anything we list the situations that could have made the sky look the way it does, one at a time, each with its reason.

1. *A Gaussian, isotropic sky with the best-fit spectrum, seen through our instrument and our processing.* This is what the standard model predicts: inflation produces Gaussian initial fluctuations, linear physics up to recombination keeps them Gaussian, and the instrument adds a beam, pixels and noise.
2. *A primordial non-Gaussianity.* Many inflationary models give a small three-point function (Section T3.2); it would bend the curves asymmetrically, as Matsubara's expansion showed [113].
3. *Something left over from our Galaxy or from point sources.* Dust and synchrotron residuals, or unresolved radio galaxies, are not Gaussian and not isotropic; they would add hot spots at high thresholds and change the boundaries near the mask.
4. *A mismatch between the data and their model*: a wrong noise level, a beam slightly different from the one assumed, a spectrum different from the best fit, or the in-filled region leaking into the observed pixels. None of these is about the early universe, and all of them can move the numbers.

Only the first candidate comes with a sampling distribution we know how to produce. We can draw as many Gaussian skies as we like, pass each through the same beam, pixels, noise and mask, and measure them with the same code. For the others we would have to say *which* non-Gaussianity, *which* residual, how large. This is the defining property of a null hypothesis (Section 4.1.1): it is the explanation whose behaviour we can simulate, "it happened by chance" with chance fully specified. The test can tell us that the first candidate is a poor explanation of $\boldsymbol{y}_{\text{obs}}$. It cannot, by itself, tell us which of the other three is right.

The question, in words. If the sky were a Gaussian, isotropic field with the best-fit spectrum, observed and processed exactly as the SMICA map was, would a vector of 36 numbers as far from the Gaussian average as ours, or even farther, be something we would reasonably expect?

13.9.1 What "Gaussian" contains

The null hypothesis sounds like one assumption. It is several, and it pays to separate them, because each can fail on its own and each failure would show up differently. Write the temperature fluctuation in direction $\hat{\boldsymbol{n}}$ as a sum of spherical harmonics, $T(\hat{\boldsymbol{n}}) = \sum_{\ell m} a_{\ell m} Y_{\ell m}(\hat{\boldsymbol{n}})$, as in Chapter 7. The inputs are these.

Physical inputs.

1. *Gaussianity*: the coefficients $a_{\ell m}$ are jointly Gaussian. Physically, the initial fluctuations were Gaussian and everything since was linear.

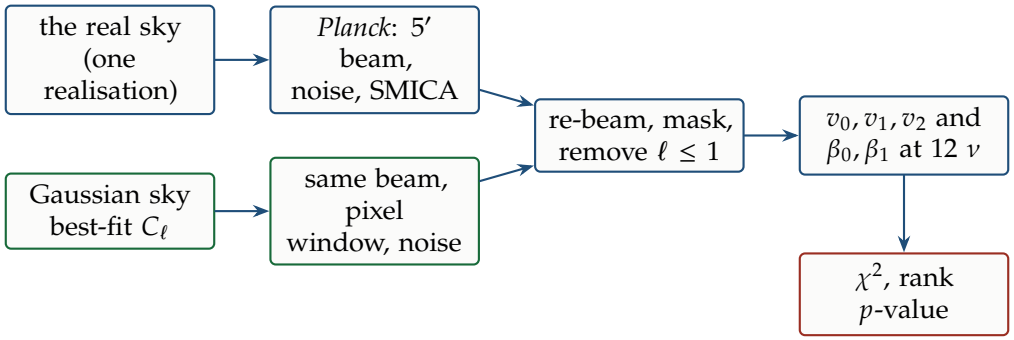


Figure 13.16: Two roads to the same statistics. The upper road is the data: one sky, observed once by *Planck* and cleaned by SMICA. The lower road (green) is the null hypothesis: Gaussian skies with the best-fit spectrum, given the same beam, pixels and noise. From the box "re-beam, mask" onwards the two roads run through identical code, so whatever that code does to the shapes, it does to both.

2. *Statistical isotropy*: the $a_{\ell m}$ are independent with variance C_ℓ that depends on ℓ only. Physically, no direction on the sky is special.
3. *The spectrum*: C_ℓ is the best-fit ΛCDM spectrum (Chapter T1).

Observational inputs (properties of the instrument and of the processing, not of the universe).

4. The map is the sky convolved with a Gaussian beam of 5′ FWHM and averaged over $N_{\rm side} = 2048$ pixels, plus noise.
5. The noise is Gaussian with a spectrum we can measure (Section 13.10.2).
6. Only the pixels outside the mask are used.

Nothing else is assumed. In particular, nobody assumes a formula for the distribution of the 36 numbers. Inputs 1–3 fix the distribution of the sky, inputs 4–6 fix what happens to it on the way to our code, and the distribution of the statistics follows by simulation, not by a derivation: the mask with its holes and the pixel counting (Section 12.2.2) make the exact distribution analytically out of reach, while simulating it is cheap.

Each input can break, and each break leaves a different fingerprint. A failure of input 1 bends the shapes of the curves at fixed spectrum, odd in ν for a skewed field. A failure of 2 makes one part of the sky differ from another; a whole-sky test like ours mixes the regions and is weak against it, which is why anisotropy is tested separately [114]. A failure of 3 changes σ_0 and σ_1, but the normalised functionals of Section 12.7.4 divide most of that out. Failures of 4–6 are the dangerous ones: they are mundane, they produce exactly the kinds of deviation a physical signal would, and they are the reason the simulations must copy the processing step by step.

13.9.2 From 36 numbers to one, and to a probability

Of the pipeline that runs through the book (a physical model, a field, a summary of it, the sampling distribution of that summary, and an inference), this part supplies the last two arrows for the real sky: the sampling distribution of the summaries comes from simulation, and the inference is the verdict on the Gaussian explanation. The test has three moving parts, each met earlier in the book. The statistics come from Chapter 12: for a threshold ν in units of the map's own standard deviation σ_0, the excursion set is the set of observed pixels with $u = (T - \bar{T})/\sigma_0 > \nu$, and we measure its area fraction v_0, its boundary length v_1 and its Euler

characteristic v_2, normalised as in (12.35), plus the two Betti numbers β_0 (pieces) and β_1 (holes) whose difference is the Euler characteristic. The comparison is the quadratic form

$$\chi^2(\boldsymbol{y}) = h\,(\boldsymbol{y}-\bar{\boldsymbol{y}})^{\mathsf{T}}\widehat{\mathbf{C}}^{-1}(\boldsymbol{y}-\bar{\boldsymbol{y}}), \tag{13.17}$$

where $\bar{\boldsymbol{y}}$ and $\widehat{\mathbf{C}}$ are the mean and the covariance of the statistics over n simulated skies and h is the Hartlap factor of (3.43); Section 13.12 derives why this is the right summary. And the probability is the rank of the data among the simulations,

$$\widehat{p} = \frac{\#\{\text{simulations } i:\ \chi^2(\boldsymbol{y}_i) \geq \chi^2(\boldsymbol{y}_{\text{obs}})\}}{n}, \tag{13.18}$$

the fraction of Gaussian skies that land at least as far from the Gaussian average as the real one, which is the p-value of Section 4.1.3 with the null distribution sampled instead of written down.

Example: how precisely n simulations measure a probability. Each simulation either lands beyond the data or does not, with probability p under the null hypothesis, independently of the others. So the count in (13.18) is binomial with n trials and success probability p, and

$$\operatorname{Var}\widehat{p} \overset{(1)}{=} \frac{1}{n^2}\operatorname{Var}\big(\#\{i:\chi_i^2 \geq \chi^2_{\text{obs}}\}\big) \overset{(2)}{=} \frac{1}{n^2}\,n\,p(1-p) = \frac{p(1-p)}{n}.$$

(1) $\widehat{p}$ is the count divided by the constant n, and a constant factor comes out of a variance squared.
(2) The variance of a binomial count is $np(1-p)$ (Chapter 2).

With $n = 1000$ simulations and $p = 0.9$ the Monte Carlo error is $\sqrt{0.09/1000} = 0.009$, so a quoted 90% means "between 89 and 91%". With $n = 400$ it is 0.015. And $1/n$ is the smallest probability that can be resolved at all: a sky beyond every one of 1000 simulations has $\widehat{p} = 0$, which means only "$p \lesssim 10^{-3}$". The precision is set by the number of simulations, not by the data; this is why *Planck* ran 999 (Figure 13.17).

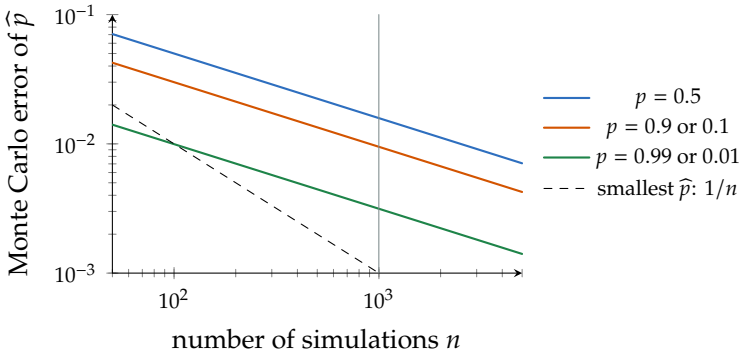


Figure 13.17: How well n simulated skies pin down a probability. The error $\sqrt{p(1-p)/n}$ falls as $1/\sqrt{n}$; the dashed line is the smallest nonzero probability that n simulations can express. The grey vertical line marks *Planck*'s 999 simulations.

Example: what a large p means. *Planck*'s Table 5 reports, for the SMICA temperature map at $N_{\text{side}} = 256$, a probability of 93.9% of a larger χ^2 [114]. Read with the question above: if the sky were Gaussian, 94 of every hundred Gaussian skies would land farther from the Gaussian average than the real one. Our sky is closer to the average than most. That is not evidence *for* Gaussianity in the sense of a probability that the null is true (Section 4.4.4); it says only

that these 36 numbers give no reason to doubt it. A probability close to 100% would itself be suspicious: a sky that matches the Gaussian average too well hints that the simulations are noisier than the data, for example a noise model that is too high. Both tails carry information, which is why *Planck* reports the fraction with a larger χ^2 and looks at both ends.

13.10 From the archive to a working map

The *Planck* Legacy Archive, mirrored at IRSA, holds the final (2018, "PR3") maps as HEALPix files at $N_{\rm side} = 2048$, about fifty million pixels of 1.7′ each. Three products are all a temperature-topology test needs, and the script `code/ch13/b01_prepare.py` turns them into maps we can measure. The SMICA CMB map (`COM_CMB_IQU-smica_2048_R3.00_full.fits`, 2.0 GB) is the sky after component separation: SMICA combines the nine frequency maps with weights, chosen multipole by multipole, that keep the black-body CMB and suppress everything with a different frequency dependence. It contains the temperature, the two Stokes polarization maps and, for each, an *inpainted* version in which the pixels inside the mask have been filled in by an inpainting algorithm with values that continue the CMB smoothly across the gap. The two half-mission maps (`..._hm1.fits` and `..._hm2.fits`, 0.6 GB each) are the same sky made from the first and the second half of the observing time; we use them only to measure the noise. The common temperature mask (`COM_Mask_CMB-common-Mask-Int_2048_R3.00.fits`, 0.2 GB) marks with zeros the Galactic plane and several hundred point sources where no component-separation method is trusted; it keeps 77.9% of the sky. All four files sit in `data/planck/` and are read with `healpy.read_map`, which also converts them from the "nested" pixel ordering of the archive to the "ring" ordering our code uses. The binned TT spectrum of the same release (`COM_PowerSpect_CMB-TT-binned_R3.01.txt`) is used only as a check.

Topology is not measured at $N_{\rm side} = 2048$. The pixel-counting estimators need a field that is smooth on the pixel scale (Section 12.2.2), and different smoothing scales look at different physics. We follow *Planck*'s Table 1 and work at four resolutions: $N_{\rm side} = 64, 128, 256, 512$ with Gaussian smoothing of FWHM 160′, 80′, 40′ and 20′, always about three pixels across [114]. At 160′ the map shows the large-angle sky, $\ell \lesssim 100$; at 20′ it reaches the third acoustic peak, $\ell \approx 800$ (Figure 13.18).

13.10.1 Changing the resolution of a map

The question. We hold a map that is the sky seen through a 5′ beam and stored in $N_{\rm side} = 2048$ pixels. We want the map that *would* have come out of an instrument with a 20′ beam and $N_{\rm side} = 512$ pixels. Can we get it without going back to the sky?

What the archive map is, in harmonic space. The beam and the pixel averaging are both (to a very good approximation) convolutions with a kernel that depends only on the angle, so each multiplies the harmonic coefficients of the sky by a number that depends only on ℓ: the beam transfer function b_ℓ of (8.10) and the pixel window p_ℓ of (8.11). Leaving out the noise for a moment, the coefficients of the archive map are $a^{\rm map}_{\ell m} = b^{5'}_\ell \, p^{2048}_\ell \, a^{\rm sky}_{\ell m}$, and the map we want has $a^{\rm new}_{\ell m} = b^{\rm new}_\ell \, p^{\rm new}_\ell \, a^{\rm sky}_{\ell m}$. Dividing one by the other removes the sky we do not know:

$$a^{\rm new}_{\ell m} \overset{(1)}{=} b^{\rm new}_\ell \, p^{\rm new}_\ell \, a^{\rm sky}_{\ell m} \overset{(2)}{=} b^{\rm new}_\ell \, p^{\rm new}_\ell \, \frac{a^{\rm map}_{\ell m}}{b^{5'}_\ell \, p^{2048}_\ell} = \frac{b^{\rm new}_\ell \, p^{\rm new}_\ell}{b^{5'}_\ell \, p^{2048}_\ell} \, a^{\rm map}_{\ell m} . \tag{13.19}$$

(1) The definition of the map we want: the sky through the new beam, averaged over the new pixels.

(2) Solve the archive relation for the sky, $a^{\rm sky}_{\ell m} = a^{\rm map}_{\ell m} / (b^{5'}_\ell \, p^{2048}_\ell)$, and substitute.

So we compute the $a_{\ell m}$ of the archive map once (up to $\ell = 1535$, the band limit $3N_{\rm side} - 1$ of the finest working map), multiply by the ratio of windows, and synthesise the new map directly at the new $N_{\rm side}$. This is the procedure of *Planck*'s Sect. 2 [114]. The step is safe because the new beam is much wider than the old one: the ratio in (13.19) is below one at every ℓ and falls steeply, so dividing by $b_\ell^{5'}$ never amplifies anything.

Example: the ratio at $\ell = 500$ for the 20′ map. With $\sigma_b = \mathrm{FWHM}/\sqrt{8\ln 2}$, the 20′ beam has $\sigma_b = 8.49' = 2.47 \times 10^{-3}$ rad and the 5′ beam $\sigma_b = 2.12' = 6.18 \times 10^{-4}$ rad. Then

$$\frac{b_{500}^{20'}}{b_{500}^{5'}} \overset{(1)}{=} \exp\Big[-\frac{500 \cdot 501}{2}\big(\sigma_{20'}^2 - \sigma_{5'}^2\big)\Big] \overset{(2)}{=} \exp\big[-125250 \times (6.10 - 0.38) \times 10^{-6}\big] = e^{-0.716} = 0.49.$$

(1) The ratio of two Gaussian beams (8.10) is a Gaussian beam whose σ_b^2 is the difference of the two: smoothing twice adds variances, and un-smoothing subtracts them.
(2) $\sigma_{20'}^2 = 6.10 \times 10^{-6}$ and $\sigma_{5'}^2 = 0.38 \times 10^{-6}$ in rad^2.

The pixel windows change this by a few per cent ($p_{500}^{512} = 0.956$, $p_{500}^{2048} = 0.997$). At $\ell = 500$ half of every coefficient survives; at $\ell = 1000$ the ratio is $e^{-2.9} = 0.06$, and the 20′ map holds almost nothing beyond.

Working definition 13.1 (changing the resolution of a sky map)

To bring a map with beam $b_\ell^{\rm old}$ and pixel window $p_\ell^{\rm old}$ to a new beam and pixelisation: compute its $a_{\ell m}$ [one harmonic transform at the old resolution, up to the new band limit $3N_{\rm side} - 1$], multiply by $b_\ell^{\rm new} p_\ell^{\rm new} / (b_\ell^{\rm old} p_\ell^{\rm old})$, and synthesise at the new $N_{\rm side}$. Valid when the new beam is wider than the old one, so that the ratio is below one at every ℓ.

The mask goes through the same procedure. The common mask is binary at $N_{\rm side} = 2048$. At $N_{\rm side} = 128$ with an 80′ beam, a working pixel near the Galactic plane is a weighted average of thousands of archive pixels, some trusted and some not. What is the right question to ask of it? Not "is the centre of this pixel inside the mask?", but "how much of the weight that went into this pixel came from trusted sky?" Smoothing the binary mask with the same beam answers exactly that: the smoothed mask at a working pixel is the beam-weighted fraction of trusted area around it. *Planck* keeps the pixels where this fraction is at least 0.9 [114], and so do we. The masks this produces keep 71.3%, 73.6%, 74.7% and 75.5% of the sky at the four resolutions, against 71.3, 73.6, 74.7 and 75.6% in their Table 1: the same masks, to within a tenth of a per cent of the sky. Wider beams lose more sky because more pixels near the edge collect weight from inside the mask.

Why the inpainted map, and what it does not give us. Smoothing mixes each pixel with its neighbours, so a pixel just outside the mask collects weight from pixels just inside it. With the raw map, an 80′ beam would drag the bright Galactic plane into every observed pixel near the edge of the mask. With the inpainted map it drags in the filled-in values instead, which continue the CMB smoothly and are tame. They are still not measurements. Statistics are computed only on the pixels the thresholded mask keeps; the in-filled values enter only through the smoothing near the edge, and the 0.9 threshold limits that to a tenth of the weight.

The monopole and the dipole. The map's average over the sky (its monopole) is set by an arbitrary zero level, and its dipole is dominated by our motion through the CMB rest frame; neither is part of the anisotropy we want to test. Both are removed by fitting $c_0 + \boldsymbol{c} \cdot \hat{\boldsymbol{n}}$ to

the observed pixels by least squares and subtracting the fit from the $\ell \le 1$ coefficients (the function `remove_monodipole` in `code/ch13/lib_skytopo.py`). Every simulated sky gets the same treatment, because a fit on a cut sky also removes a little of the $\ell \ge 2$ power that leaks into the fitted modes, and the simulations must lose the same little.

13.10.2 How much noise is in the map?

The simulations need a noise model, and the half-mission maps give it to us without any knowledge of the detectors. *The idea.* Both halves see the same sky, so the sky cancels in their difference and only the noise is left. Write the two half-mission maps as $h_1 = s + n_1$ and $h_2 = s + n_2$, with s the sky (after the same beam) and n_1, n_2 independent noise of equal variance σ_n^2 per pixel. The full-mission map averages the two halves. Its noise and the noise of the half-difference map are

$$\begin{aligned} \mathrm{Var}\Big(\frac{h_1+h_2}{2} - s\Big) &\overset{(1)}{=} \mathrm{Var}\Big(\frac{n_1+n_2}{2}\Big) \overset{(2)}{=} \frac{\sigma_n^2+\sigma_n^2}{4} = \frac{\sigma_n^2}{2}, \\ \mathrm{Var}\Big(\frac{h_1-h_2}{2}\Big) &\overset{(3)}{=} \mathrm{Var}\Big(\frac{n_1-n_2}{2}\Big) \overset{(4)}{=} \frac{\sigma_n^2+\sigma_n^2}{4} = \frac{\sigma_n^2}{2}. \end{aligned} \tag{13.20}$$

(1) The sky term cancels: $(s + n_1 + s + n_2)/2 - s = (n_1 + n_2)/2$.
(2) Independent terms add their variances, and a factor $\frac{1}{2}$ comes out squared.
(3) The sky cancels in the difference: $(s + n_1 - s - n_2)/2$.
(4) $\mathrm{Var}(-n_2) = \mathrm{Var}(n_2)$, so the minus sign changes nothing.

The half-difference map is therefore a noise map with exactly the statistics of the noise in the full map, and the same holds multipole by multipole. We measure its spectrum on the observed sky with the pseudo-C_ℓ of Section 8.9.3 divided by f_{sky} (good enough for a quantity that varies slowly with ℓ), and the simulations add Gaussian noise with that spectrum, re-beamed by (13.19) exactly as the data are. At multipoles 400–800 the noise corresponds to 32 μK arcmin, and after correcting for the 5′ beam it is 0.002, 0.019 and 0.144 of the CMB power at $\ell = 500$, 1000 and 1500 (Figure 13.18).

Example: does the noise matter for topology? The working beam has already removed the multipoles where the noise competes. The fraction of the working map's variance that is noise, $\sum(2\ell+1)N_\ell w_\ell^2$ over $\sum(2\ell+1)C_\ell w_\ell^2$ with w_ℓ the ratio of windows in (13.19), is 0.01%, 0.02%, 0.02% and 0.03% at 160′, 80′, 40′ and 20′. For the gradient variance σ_1^2, which weights each multipole by a further $\ell(\ell+1)$ and is the quantity the boundary length and the Euler characteristic depend on, it is at most 0.08%. So at these resolutions the temperature map is signal-dominated, and the noise model, crude as it is, cannot move the test. *Planck*'s noise simulations follow the scanning pattern, so the noise is deeper near the ecliptic poles than elsewhere; our noise is isotropic. That difference would matter at $N_{\mathrm{side}} = 1024$ and 2048, where the working beams reach $\ell \approx 2000$, and it does not matter here.

13.10.3 Does the map have the power of the simulations?

The simulations are drawn from the best-fit spectrum of Chapter T1. If the map's power differed from it substantially, the test would compare skies of different amplitude, and could fail for that reason alone. The check in Figure 13.18 is rough (the binary mask couples multipoles, Section 8.10, which the MASTER estimator of Section 8.11 would undo), but it is enough to see three things. Over $30 \le \ell < 800$, which carries almost all the variance of our working maps, the ratio of the map's power to the best fit is 1.005. Over $800 \le \ell < 1500$ it is 1.088, and 1.06 once

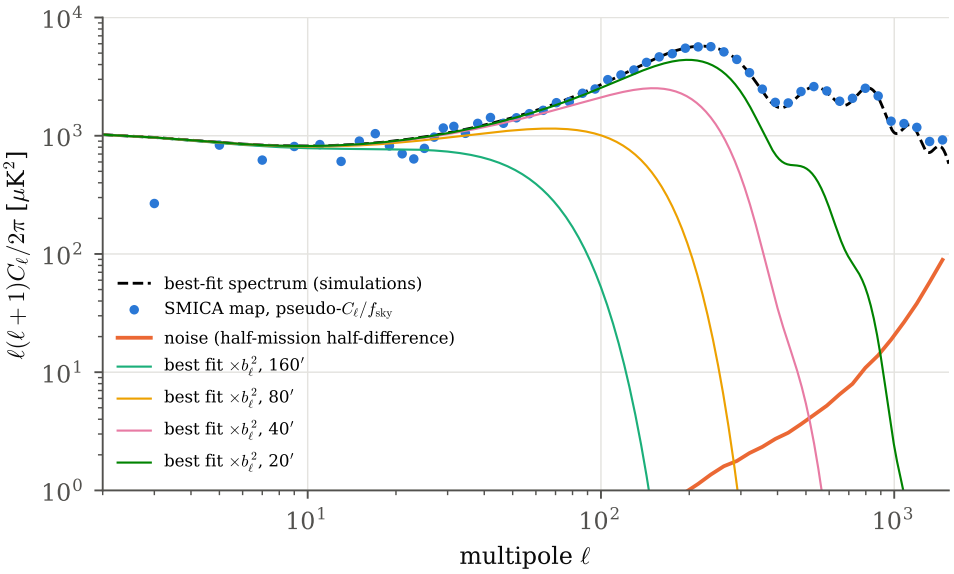


Figure 13.18: The power of the SMICA map against the spectrum used for the simulations. Blue points: the map's pseudo-spectrum on the observed sky divided by $f_{\rm sky}$, with the 5′ beam and the pixel window removed; dashed: the best-fit spectrum; orange: the noise from the half-difference map. The four thin curves are the best fit multiplied by the squared working beams: each working map sees only the multipoles under its curve, where the noise is far below the signal.

the noise power is subtracted; the remaining few per cent are the leakage of large-scale power through the edges of an unsoftened mask and the residual point sources that *Planck* also sees in the data and not in its simulations [114]. And over $2 \le \ell < 30$ the ratio is 0.71: the sky has less large-angle power than the best fit predicts. This is the well-known low-variance anomaly of the largest scales (Section T3.5.3), and it will show up again below as a smaller σ_0 of the real map at 160′.

Does a power mismatch spoil the test? Not much, by construction. The thresholds are in units of the map's *own* standard deviation, so a sky that is uniformly too faint or too bright gives the same excursion sets (Section 12.3.2); and the functionals are divided by the map's own $(\sigma_1/\sigma_0)^k$, which absorbs most of a change in the spectrum's shape (Section 12.7.4). What survives this normalisation is the shape of the curves, which is what Gaussianity predicts.

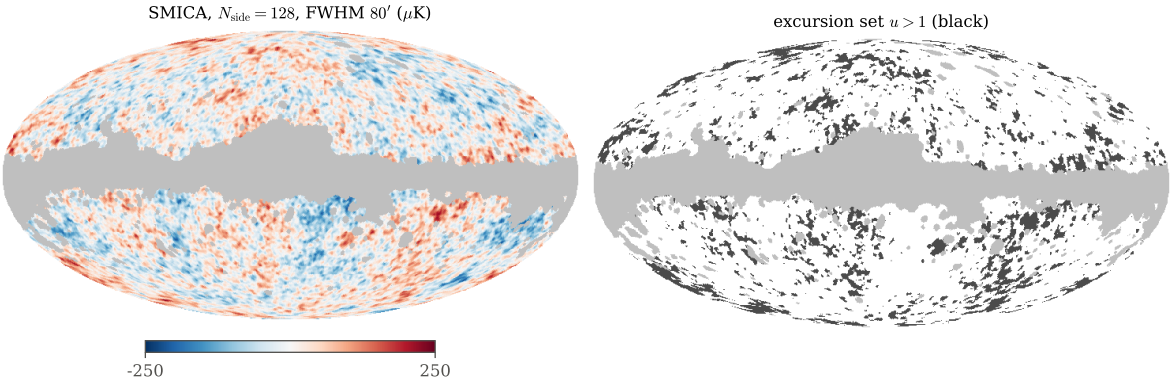


Figure 13.19: The working map at $N_{\rm side} = 128$ with an 80′ beam (left), outside the degraded common mask (grey), and its excursion set at one standard deviation (right): the observed pixels with $u = (T - \bar{T})/\sigma_0 > 1$. Every statistic below is a count or a measure of shapes like these, at twelve thresholds and four resolutions.

13.11 The area, boundary and Euler characteristic of the real sky

The working maps of Section 13.10 are ready: four resolutions, each with its mask of observed pixels. At each resolution we now ask the questions of Chapter 12 of the real sky. Standardise the map on its observed pixels, $u = (T - \bar{T})/\sigma_0$, with $\bar{T}$ and σ_0 the mean and the standard deviation over those pixels only. For a threshold ν, the excursion set X_ν is the set of observed pixels with $u > \nu$, the hot regions of Figure 13.19. How much of the observed sky does it cover? How long is its boundary? How many pieces does it have, minus how many holes? Each answer must come from code that a simulated sky passes through unchanged, and each needs one decision about pixels that the formulas of Section 12.5 never had to make.

13.11.1 Three estimators on a cut sphere

Area. The fraction of observed pixels with $u > \nu$. HEALPix pixels all have the same area, so counting them is measuring area, and $v_0(\nu)$ is this fraction.

Boundary length, without drawing a contour. Tracing the contours $u = \nu$ through a pixelised map and adding up their lengths is fiddly, especially where they run into the mask. There is a shortcut. *The question:* how is the length of a contour related to the area of a thin band around it? Take the band of points where u lies within $\Delta\nu/2$ of ν. Near a point of the contour, u changes fastest across the contour, at the rate $|\nabla u|$ per radian, so the band there is $\Delta\nu/|\nabla u|$ wide. A short piece of contour of length $\mathrm{d}s$ therefore carries a strip of band of area $\mathrm{d}A = \mathrm{d}s\,\Delta\nu/|\nabla u|$. Solving for $\mathrm{d}s$ and adding up the pieces,

$$L(\nu) \overset{(1)}{=} \int_{u=\nu} \mathrm{d}s \overset{(2)}{=} \lim_{\Delta\nu\to 0} \frac{1}{\Delta\nu} \int_{|u-\nu|<\Delta\nu/2} |\nabla u|\,\mathrm{d}A \overset{(3)}{=} \frac{A_{\mathrm{obs}}}{\Delta\nu} \left\langle |\nabla u|\, \mathbb{1}\left(|u-\nu| < \tfrac{\Delta\nu}{2}\right) \right\rangle_{\mathrm{obs}}. \tag{13.21}$$

(1) The length is the integral of the arc-length element along the contour.
(2) Each piece of contour, $\mathrm{d}s = |\nabla u|\,\mathrm{d}A/\Delta\nu$, comes with the piece of band beside it, so the integral over the contour becomes an integral over the band (this is the coarea formula, here in its simplest form).
(3) On equal-area pixels an integral over the observed sky is A_{obs} times the average over observed pixels; $\mathbb{1}(\cdot)$ is 1 inside the band and 0 outside.

The gradient comes exactly from the harmonic coefficients (`healpy.alm2map_der1` returns $\partial_\theta u$ and $\partial_\phi u/\sin\theta$), so no finite differences are needed; we use $\Delta\nu = 0.1$. *Planck*'s V_1 is a quarter of the length per unit area (their eq. 14), and we follow that convention: $v_1 = L/(4A_{\mathrm{obs}})$.

Example: the band estimator on a straight contour. Take $u = k\,x$ on a flat patch of side a, a field that rises steadily in one direction. The contour $u = \nu$ is the straight line $x = \nu/k$, of length a. The band $|u - \nu| < \Delta\nu/2$ is the strip $|x - \nu/k| < \Delta\nu/(2k)$, of area $a\,\Delta\nu/k$, and in it $|\nabla u| = k$. Then (13.21) gives $k \cdot a\,\Delta\nu/k/\Delta\nu = a$, the right answer for any slope: a steep field has a thin band but a large gradient, and the two compensate exactly.

The Euler characteristic, at every threshold at once. Section 12.2.3 computed χ of a union of closed pixels as $V - E + F$: vertices minus edges plus faces of the union. On the cut sphere we keep that rule and add one: an unobserved pixel never belongs to the set. *The question:* at which threshold does a given vertex, edge or face join the union? A face (a pixel) joins when ν drops below its value. An edge or a vertex belongs to the union as soon as *one* of the observed pixels

around it does, so it joins when ν drops below the largest observed value around it. Every element thus has an entry level, and

$$\chi(\nu) = \#\{\text{vertices entered above } \nu\} - \#\{\text{edges entered above } \nu\} + \#\{\text{faces entered above } \nu\}. \tag{13.22}$$

Sorting the three lists of entry levels once turns every $\chi(\nu)$ into three binary searches, which is how `entry_levels` and `chi_from_levels` in `code/ch13/lib_skytopo.py` work. On full-sky maps the result agrees exactly, threshold by threshold, with the pixel-by-pixel counter of Chapter 12.

Example: four pixels by hand. A 2×2 block of pixels has values 3 (top left), 1 (top right), 0 (bottom left) and 2 (bottom right): 9 vertices, 12 edges, 4 faces. Lower the threshold to $\nu = 1.5$. The faces above it are the top-left and bottom-right pixels, which touch only at the central vertex. Entry levels: the central vertex sees all four values, so it joins at $\max(3, 1, 0, 2) = 3$; every other corner of the top-left pixel has that pixel among its neighbours and joins at 3, every other corner of the bottom-right pixel joins at 2 or 3; an edge of either pixel joins at the larger of its two neighbouring values, at least 2; and every remaining vertex and edge touches only the pixels with values 1 and 0, so it joins below 1.5. Counting the elements that joined above 1.5: $V = 4 + 4 - 1 = 7$ (the central vertex is shared), $E = 4 + 4 = 8$, $F = 2$, so $\chi = 7 - 8 + 2 = 1$. The two diagonal pixels form one piece, because closed pixels that share a corner touch. This is the convention of Section 12.2.2 (hot pixels connect through corners, cold ones only through edges), and (13.22) implements it automatically.

Normalisation. As in Section 12.7.4, the functionals are divided by the power of σ_1/σ_0 that carries their dependence on the spectrum, with σ_1^2 the mean of $|\nabla T|^2$ over the observed pixels: v_1 by σ_1/σ_0, and the Euler characteristic per steradian of observed sky, $v_2 = \chi/A_{\text{obs}}$, by $(\sigma_1/\sigma_0)^2$. On the cut sky v_2 is not the Gaussian curve of (12.35) alone: the mask's own topology and its boundary add the terms $\mathcal{L}_0\Psi(\nu)$ and $\mathcal{L}_1\rho_1(\nu)$ of Section 12.8.1. The simulations are cut by the same mask and carry the same terms, so nothing has to be modelled.

13.11.2 What the real sky gives

Example: one threshold, read off the sky. At $N_{\text{side}} = 256$ and $\nu = 0.27$ (one of *Planck*'s twelve thresholds), the excursion set of the SMICA map has 1854 separate hot regions and 894 holes in them, so $\chi = 960$. The simulated skies, cut and processed identically, give $\chi = 830$ on average with a spread of 103.6 from sky to sky. The real sky lies $(960 - 830)/103.6 \approx 1.3$ standard deviations above the mean, a deviation that about one Gaussian sky in five exceeds in one direction or the other. A single number like this is easy to read, and also easy to over-read: there are 36 such numbers per resolution, and they are not independent.

Figure 13.20 shows the three normalised curves at $N_{\text{side}} = 256$ for the data and for the simulations; Figure 13.21 shows, at all four resolutions, each number's deviation from the simulated mean in units of the simulated standard deviation, $\Delta v/\sigma$. Three things stand out. First, the real sky's curves lie inside the band of Gaussian skies everywhere, at every resolution; the largest single deviation is 2.2, 1.8, 1.9 and 2.2 standard deviations at $N_{\text{side}} = 64$, 128, 256 and 512. Second, the deviations mostly change smoothly with ν: neighbouring thresholds tend to move together, because the excursion sets at ν and at $\nu + 0.55$ share most of their pixels. That smoothness is the correlation we will have to take into account. Third, no curve stays at zero: each wanders up and down by about one standard deviation, at every resolution. That is what a single realisation does. At $N_{\text{side}} = 64$ the observed sky holds only a few hundred

independent patches, so a slight surplus of hot area at some thresholds and a deficit at others is expected, not a sign of anything.

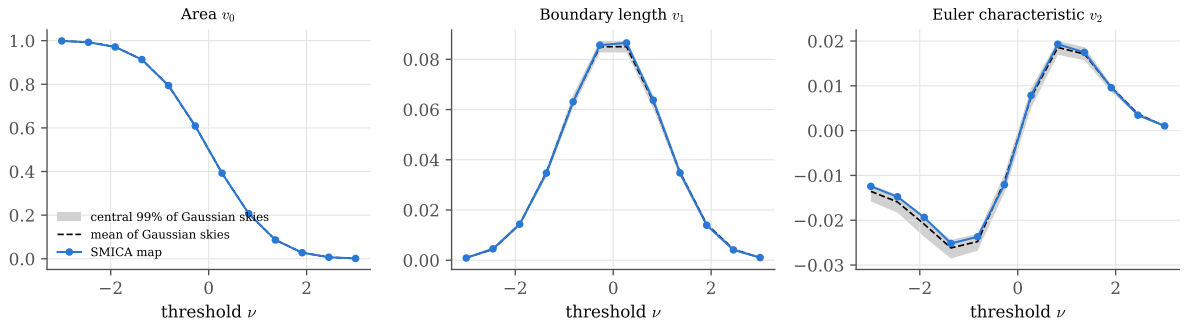


Figure 13.20: The three normalised Minkowski functionals of the SMICA map at $N_{\mathrm{side}} = 256$ (40′), against 500 Gaussian skies with the best-fit spectrum, the same beam, pixels, noise and mask. The grey band encloses the central 99% of the simulated skies and the dashed curve gives their mean. The mask breaks the symmetry of the genus curve about zero because its boundary and its holes add pieces to every excursion set.

The amplitude that was divided out. The normalisation removed σ_0 and σ_1/σ_0; it is worth looking at them directly, because they are where the spectrum lives. The real map's σ_0 is 46.7, 58.7, 75.6 and 90.9 μK at the four resolutions; the Gaussian skies give 52.7 ± 3.4, 63.4 ± 2.7, 79.5 ± 2.3 and 94.1 ± 1.9 μK. Only 1.6% of the skies have a σ_0 as small as the data at 160′: the deficit of large-angle power found in Section 13.10.3, now as a variance. The ratio σ_1/σ_0 is 114.1 per radian in the data at $N_{\mathrm{side}} = 256$, against 108.7 ± 3 in the skies. Had we compared the unnormalised functionals, or set the thresholds in units of the simulations' average σ_0, the low variance would have shifted every curve and been mistaken for a change of shape. The normalisation keeps the two questions apart: "is the amplitude typical?" is a question about the power spectrum, and "is the shape Gaussian?" is the question asked here.

Watch out: looking for the largest deviation

With 36 numbers per resolution and four resolutions, the largest of 144 deviations of a Gaussian sky is typically well beyond two standard deviations, even though each single number exceeds two standard deviations only 5% of the time. A deviation found by scanning a figure must be judged against the largest deviation of the simulations scanned in the same way, not against the normal distribution: this is the look-elsewhere effect of Chapter 4. The χ^2 of Section 13.12 avoids the problem by fixing in advance one number per resolution.

13.12 Thirty-six numbers, one verdict

Figure 13.21 shows 36 deviations per resolution, all modest, all moving together with their neighbours. How surprised should we be by the whole pattern? Adding up the squared deviations would count the same fluctuation many times, once for every threshold it touches; picking the largest one invites the look-elsewhere trap. We need one number per resolution that measures how far the vector $\boldsymbol{y}_{\mathrm{obs}}$ lies from the Gaussian average, counting each independent fluctuation once, and then the distribution of that number under the null hypothesis of Section 13.9.

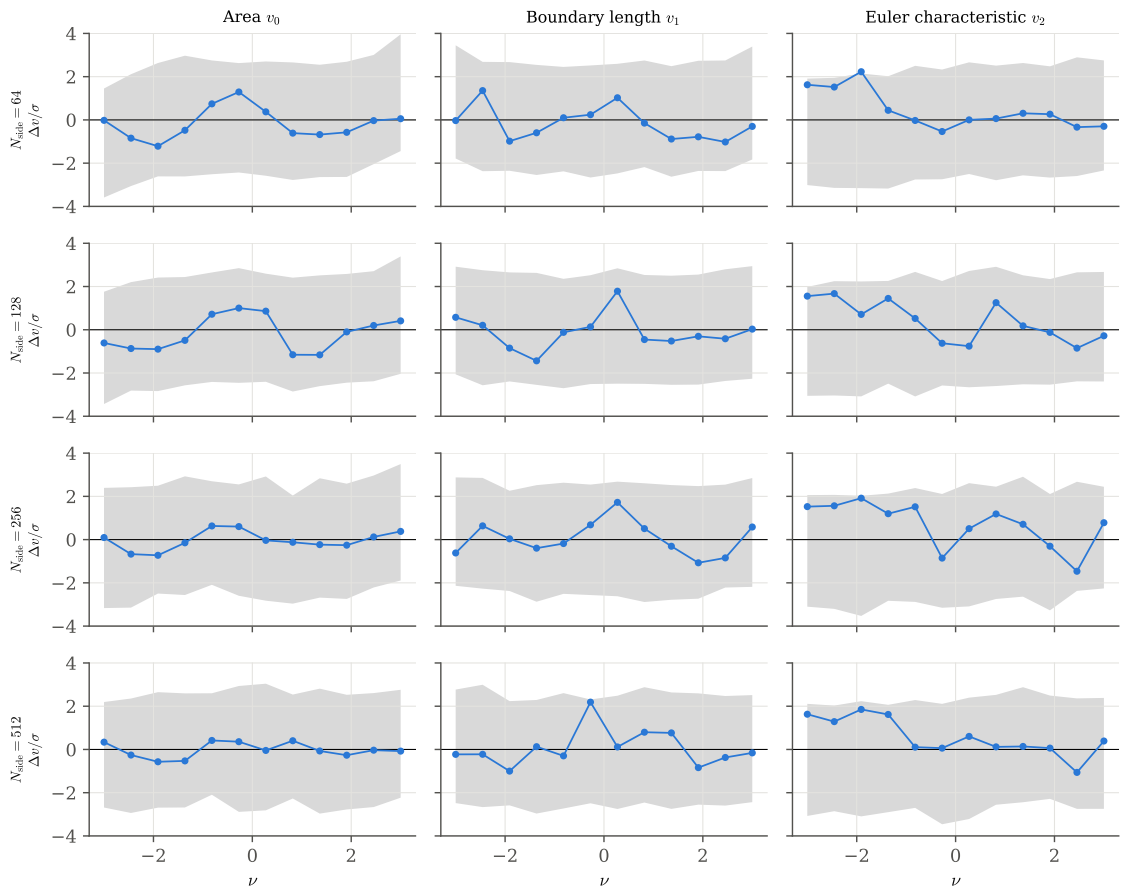


Figure 13.21: Deviation of each of the 36 numbers of the SMICA map from the mean of the Gaussian skies, in units of their standard deviation, at the four resolutions (rows). Grey: the central 99% of the skies, each measured the same way. Neighbouring thresholds move together, which is why the 36 deviations cannot be judged one at a time.

13.12.1 The distance that the null hypothesis itself defines

The question. Among all the vectors $\boldsymbol{y}$ that Gaussian skies can produce, which ones count as "farther from the average than ours"? The natural answer is: those that the null hypothesis makes *less probable* than ours. To use it we need the density of $\boldsymbol{y}$ under the null.

The density. Each component of $\boldsymbol{y}$ is an average over the observed sky: a fraction of pixels, a length per area, a count per area. The observed sky contains many patches a correlation length apart, and their contributions are nearly independent, so by the central limit theorem (Chapter 2) the vector is close to a multivariate Gaussian with some mean $\boldsymbol{\mu}$ and covariance $\mathbf{C}$:

$$p(\boldsymbol{y} \mid H_0) \approx \frac{1}{\sqrt{(2\pi)^p \det \mathbf{C}}} \exp\!\left[-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^{\mathsf{T}}\mathbf{C}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right], \tag{13.23}$$

with $p = 36$ components. The density depends on $\boldsymbol{y}$ only through the quadratic form in the exponent, and falls as it grows. So "less probable than ours" is the same as "a larger value of $Q(\boldsymbol{y}) = (\boldsymbol{y}-\boldsymbol{\mu})^{\mathsf{T}}\mathbf{C}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})$". The surfaces of constant Q are ellipsoids aligned with the correlations: a deviation along a direction in which the skies naturally fluctuate a lot costs little, a deviation in a direction they rarely take costs much. And if $\boldsymbol{\mu}$ and $\mathbf{C}$ were known exactly, Q would follow the χ^2 distribution with p degrees of freedom (Section 2.13.3): mean 36, standard deviation $\sqrt{72} \approx 8.5$, the scale of a χ^2 with 36 components that Section T3.5.1 already met.

Example: why correlated thresholds must not be added as if independent. Take two thresholds whose deviations, in units of their standard deviations, have correlation ρ, and suppose both came out at +1. With $\mathbf{C} = \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)$,

$$Q \overset{(1)}{=} \frac{1}{1-\rho^2}\,(1,\ 1)\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} \overset{(2)}{=} \frac{2-2\rho}{1-\rho^2} \overset{(3)}{=} \frac{2}{1+\rho}.$$

(1) The inverse of a 2×2 matrix: swap the diagonal, negate the off-diagonal, divide by the determinant $1 - \rho^2$.
(2) Multiply out: $1 - \rho - \rho + 1$.
(3) $1 - \rho^2 = (1 - \rho)(1 + \rho)$, and $1 - \rho$ cancels.

For independent thresholds ($\rho = 0$) the two deviations count twice, $Q = 2$. For $\rho = 0.9$, the strongest correlation between neighbouring area thresholds in Figure 13.22 (most neighbouring pairs lie between 0.5 and 0.9), $Q = 1.05$: two deviations that the skies almost always make together are worth hardly more than one. Had the second deviation come out at -1 instead, $Q = 2/(1 - \rho) = 20$: a pattern that cuts across the correlation is very improbable, even though each number on its own is ordinary. The quadratic form sees both effects.

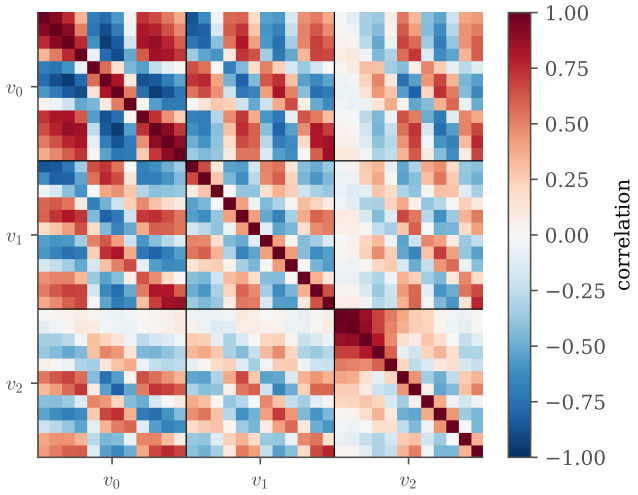


Figure 13.22: Correlation matrix of the 36 numbers at $N_{\text{side}} = 256$, estimated from the Gaussian skies. Blocks: area, length and Euler characteristic, each at twelve thresholds from -3 to 3. Each sky is measured in units of its own mean and rms, so its areas cannot all grow together: the area fractions at the tail thresholds ($|\nu| > 1$) move together, and against those in the core ($|\nu| < 1$). Neighbouring area thresholds on the same side of $|\nu| = 1$ are correlated at 0.5 to 0.9; the two pairs that straddle it hardly at all. Length and Euler characteristic are less tightly tied, but many entries far from the diagonal still reach ± 0.5 or more.

13.12.2 When the mean and covariance come from simulations

We do not know $\boldsymbol{\mu}$ and $\mathbf{C}$; we estimate them from n simulated skies, $\bar{\boldsymbol{y}}$ and $\widehat{\mathbf{C}}$. Two corrections follow, both from Chapter 3.

The inverse of a noisy covariance is biased. $\widehat{\mathbf{C}}$ is unbiased, but its inverse is not: random errors make some directions look less variable than they are, and $\widehat{\mathbf{C}}^{-1}$ inflates them. For Gaussian vectors, $\mathbb{E}\,\widehat{\mathbf{C}}^{-1} = \frac{n-1}{n-p-2}\,\mathbf{C}^{-1}$ (3.43), so the unbiased estimate of the inverse is $h\,\widehat{\mathbf{C}}^{-1}$ with the Hartlap factor $h = (n - p - 2)/(n - 1)$ [6]. With 500 skies at $N_{\text{side}} = 256$ it is 0.926; with 300 skies at $N_{\text{side}} = 512$ it is 0.876. That is the factor h in (13.17).

A simulation must not judge itself. To get the null distribution of χ^2 we compute it for every simulated sky. If sky i were included in the mean and covariance it is compared with, it would pull both towards itself and look too typical. How much too typical? Sum the uncorrected quadratic forms of all n skies, with $\boldsymbol{d}_i = \boldsymbol{y}_i - \bar{\boldsymbol{y}}$ and $\widehat{\mathbf{C}} = \frac{1}{n-1}\sum_j \boldsymbol{d}_j \boldsymbol{d}_j^{\mathsf{T}}$:

$$\sum_{i=1}^{n} \boldsymbol{d}_i^{\mathsf{T}} \widehat{\mathbf{C}}^{-1} \boldsymbol{d}_i \overset{(1)}{=} \sum_{i=1}^{n} \operatorname{tr}\!\left(\widehat{\mathbf{C}}^{-1} \boldsymbol{d}_i \boldsymbol{d}_i^{\mathsf{T}}\right) \overset{(2)}{=} \operatorname{tr}\!\left(\widehat{\mathbf{C}}^{-1} \sum_{i=1}^{n} \boldsymbol{d}_i \boldsymbol{d}_i^{\mathsf{T}}\right) \overset{(3)}{=} \operatorname{tr}\!\left(\widehat{\mathbf{C}}^{-1} (n-1)\widehat{\mathbf{C}}\right) \overset{(4)}{=} (n-1)\,p. \tag{13.24}$$

(1) A number equals its own trace, and $\operatorname{tr}(\mathbf{AB}) = \operatorname{tr}(\mathbf{BA})$ moves $\boldsymbol{d}_i$ to the end.
(2) The trace is linear, so the sum goes inside.
(3) The definition of the sample covariance.
(4) $\operatorname{tr} \mathbb{1}_p = p$.

So the average in-sample value is $p(n-1)/n$, *exactly*, whatever the skies did: slightly below p, with no scatter in the average at all. A new, independent sky compared with the same $\bar{\boldsymbol{y}}$ and $\widehat{\mathbf{C}}$ has instead an average of about $p\,(n-1)/(n-p-2)$, above p. The data are a new sky in this sense, so each simulation is treated the same way: its χ^2 uses the mean and covariance of the *other* $n-1$ skies (with Hartlap factor $(n-p-3)/(n-2)$). This "leave-one-out" distribution is the null distribution the data are ranked in. Its mean, 1 per degree of freedom at $N_{\text{side}} = 256$, confirms that the corrections work.

The Gaussian approximation is not used for the probability. Equation (13.23) told us which statistic to use. It is not trusted for the tail: at $|\nu| = 3$ the counts are small and skewed, and $\bar{\boldsymbol{y}}$ and $\widehat{\mathbf{C}}$ carry their own noise. The probability is the rank (13.18) of the data's χ^2 among the simulations' χ^2, which needs no formula. The tail of the χ^2_{36} distribution is reported alongside, as a check on how far the approximation can be trusted.

Key idea: the simulation test of a sky map

Stack the statistics into a vector. Use simulations of the null hypothesis, processed exactly like the data, to estimate its mean and covariance; form the Hartlap-corrected χ^2 of the data and, leaving each one out in turn, of every simulation. The p-value is the fraction of simulations whose χ^2 is at least as large as the data's. The Gaussian density chooses the statistic; the simulations, not a formula, give its distribution.

13.12.3 The verdict at four resolutions

The simulated skies are made and measured once, by `code/ch13/b02_sims.py`, and cached. The test itself is a short function in `code/ch13/b03_mf_test.py`:

Listing 13.6: The χ^2 of the data and the leave-one-out χ^2 of every simulation (`code/ch13/b03_mf_test.py`)

```
    n, p = Y.shape
    C = np.cov(Y, rowvar=False)
    d = y - Y.mean(0)
    c_data = (n - p - 2) / (n - 1) * d @ np.linalg.solve(C, d)
    c_sims = np.zeros(n)
    for i in range(n):
        rest = np.delete(Y, i, axis=0)
        Ci = np.cov(rest, rowvar=False)
        di = Y[i] - rest.mean(0)
        c_sims[i] = (n - 1 - p - 2) / (n - 2) * di @ np.linalg.solve(Ci, di)
    return c_data, c_sims
```

The first four lines are (13.17) for the data, with the covariance of all n skies; the loop repeats it for each simulation against the other $n-1$. The rank (13.18) is then one line, `np.mean(c_sims >= c_data)`. Table 13.6 collects the results.

	$N_{\text{side}} = 64$ 160′	128 80′	256 40′	512 20′
simulated skies n	1000	1000	500	300
Hartlap factor h	0.963	0.963	0.926	0.876
χ^2/N_{dof} of the data	0.65	0.91	0.96	1.18
$\widehat{p}$ (rank among skies), %	91.7 ± 0.9	60.3 ± 1.5	50 ± 2.2	23 ± 2.4
tail of χ^2_{36}, %	94.9	63.1	53.5	21.5
area alone, %	92.7	33.6	32.4	58.7
length alone, %	64.2	79.8	59.4	6.3
Euler characteristic alone, %	78.2	79.6	30	61.3
Planck 2018, SMICA, %	90	85.4	93.9	80.7

Table 13.6: The Minkowski-functional test of the SMICA temperature map. $\widehat{p}$ is the fraction of Gaussian skies with a χ^2 at least as large as the data's, with its binomial Monte Carlo error; the three rows below use the twelve numbers of one functional at a time. The last row is the same probability from *Planck*'s Table 5 [114], with 999 end-to-end simulations.

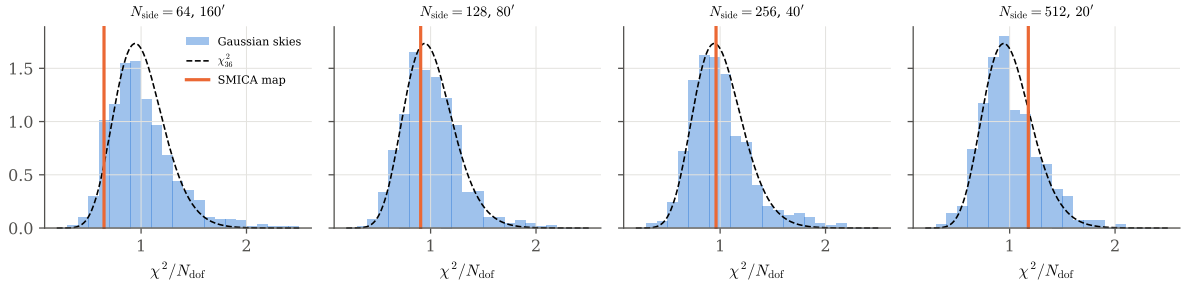


Figure 13.23: The χ^2 per degree of freedom of each Gaussian sky against the others (histogram), the χ^2_{36} density (dashed), and the SMICA map (vertical line), at the four resolutions. The data sit inside the bulk of the Gaussian skies at every resolution.

Reading the verdict. Go back to the four candidate explanations of Section 13.9. The data's χ^2 lands in the bulk of the null distribution at every resolution (Figure 13.23): between 23% and 91.7% of Gaussian skies are farther from the Gaussian average than the real one. So the first explanation, a Gaussian sky seen through our instrument, accounts for the shapes of the real sky without strain, and the other three are not needed to explain these 36 numbers. That is a statement about sensitivity as much as about the sky: a primordial non-Gaussianity, a foreground residual or a processing error smaller than the scatter of these statistics would pass unseen, and Section T3.5.2 shows how much tighter the bispectrum constrains the primordial kind.

One small number among twelve. The single-functional rows matter because a defect in one estimator (the boundary length near the mask, say) would show in one row and not in the others. Eleven of the twelve are unremarkable; the boundary length at $N_{\text{side}} = 512$ gives 6.3%. Is that a sign of something? Ask how often the smallest of twelve probabilities is that small when nothing is wrong. If the twelve were independent, each below 0.063 with probability 0.063, the chance that none of them is would be $(1 - 0.063)^{12} = 0.46$; in all the other perfectly Gaussian skies at least one row is this small. The rows are in fact correlated (the same sky

enters all of them), which lowers that chance somewhat but not to anything rare. This is the look-elsewhere effect again: the smallest probability was found by looking at twelve numbers, and it has to be judged as the smallest of twelve. The combined test, fixed in advance, is the one that carries the verdict.

13.13 Pieces, holes and their lifetimes on the real sky

The Euler characteristic of Section 13.11 is a difference, pieces minus holes, $\chi = \beta_0 - \beta_1$ (Section 12.2.1). A sky with ten more hot regions and ten more holes than a Gaussian one has the same χ. Could the real sky hide such a compensation? The Betti numbers separate the two counts, and the persistence diagram of Section 12.9 goes further: it follows each hot region from the threshold at which it appears to the threshold at which it merges with an older one. Both were developed for cosmic-web catalogues [5] and for primordial non-Gaussianity in galaxy surveys [116]; here they are applied to the real microwave sky, through the same simulation test.

The question. Does the real sky have, threshold by threshold, the number of separate hot regions and the number of holes in them that Gaussian skies have? And is the whole history of its hot regions, births and mergers, as close to that of a Gaussian sky as Gaussian skies are to one another?

13.13.1 Betti numbers on a cut sphere

Pieces. $\beta_0(\nu)$, the number of connected hot regions, is computed by the union–find algorithm of Section 12.11.1, run on the HEALPix neighbour graph: the observed pixels are added in order of decreasing u; each new pixel is joined to the observed neighbours already present (all neighbours, up to eight, that share an edge or a corner, the convention of Section 13.11.1); when it joins two regions, the younger dies (the elder rule). One pass gives the birth and death level of every hot region, and $\beta_0(\nu)$ is the number of regions born above ν and not yet merged at ν. In `code/ch13/lib_skytopo.py` the loop is compiled with `numba`, and it handles the 3 million pixels of an $N_{\mathrm{side}} = 512$ map in 1.01 s on one core of the cluster, most of it spent following the eight neighbours of each pixel.

Holes, from pieces and the Euler characteristic. Counting holes directly on a sphere with a mask is awkward: is a cold region that touches the mask a hole or not? There is no need to decide. For a surface, $\chi = \beta_0 - \beta_1 + \beta_2$, where β_2 counts closed two-dimensional shells. A subset of the sphere contains a closed shell only if it is the whole sphere, and with a mask our excursion sets never are, so $\beta_2 = 0$ and

$$\beta_1(\nu) = \beta_0(\nu) - \chi(\nu), \tag{13.25}$$

with χ from (13.22). A hole is then whatever χ says it is: a cold region enclosed by hot pixels, and also a hot region wrapped around one of the mask's holes. Both are features of the observed excursion set, and the simulations see the same mask. As a check of the code, on unmasked skies Alexander duality gives β_1 another way, as the number of cold regions (four-connected) minus one (Section 12.11.2, here on the sphere instead of the plane). The script `code/ch13/b05_checks.py` makes this comparison on 20 full-sky Gaussian maps at $N_{\mathrm{side}} = 64$ and 128, at 61 thresholds each: of the 1220 cases in which both the hot and the cold set are non-empty, 0 disagree.

Example: the split of χ at one threshold. At $N_{\mathrm{side}} = 256$ and $\nu = 0.27$ the real sky has $\beta_0 = 1854$ hot regions and $\beta_1 = 894$ holes, so that $\chi = 1854 - 894 = 960$ (Section 13.11.2). Near

$\nu = 0$ hot and cold areas are about equal and both counts are large; the Euler characteristic, their small difference, has a relative scatter much larger than either count. The Betti numbers are therefore not redundant with χ: they measure the two large numbers that χ subtracts.

The test. The statistics vector now holds β_0 and β_1 per steradian of observed sky at the twelve thresholds, 24 numbers per resolution. A threshold at which no simulated sky varied (if no sky ever had a hole in its hot set at $\nu = 3$, for example) would carry no information and would have to be dropped, since its row of the covariance would be zero; the script checks this, and at our four resolutions all 24 numbers vary. The test of Section 13.12 then runs unchanged. The SMICA map gives

	$N_{\text{side}} = 64$	128	256	512
numbers used	24	24	24	24
χ^2/N_{dof}	0.7	0.98	1.22	1.09
$\widehat{p}$, %	73.5 ± 1.4	38.9 ± 1.5	22 ± 1.9	33.3 ± 2.7

with the curves in Figure 13.24. As with the functionals, the real sky's pieces and holes are where Gaussian skies put them.

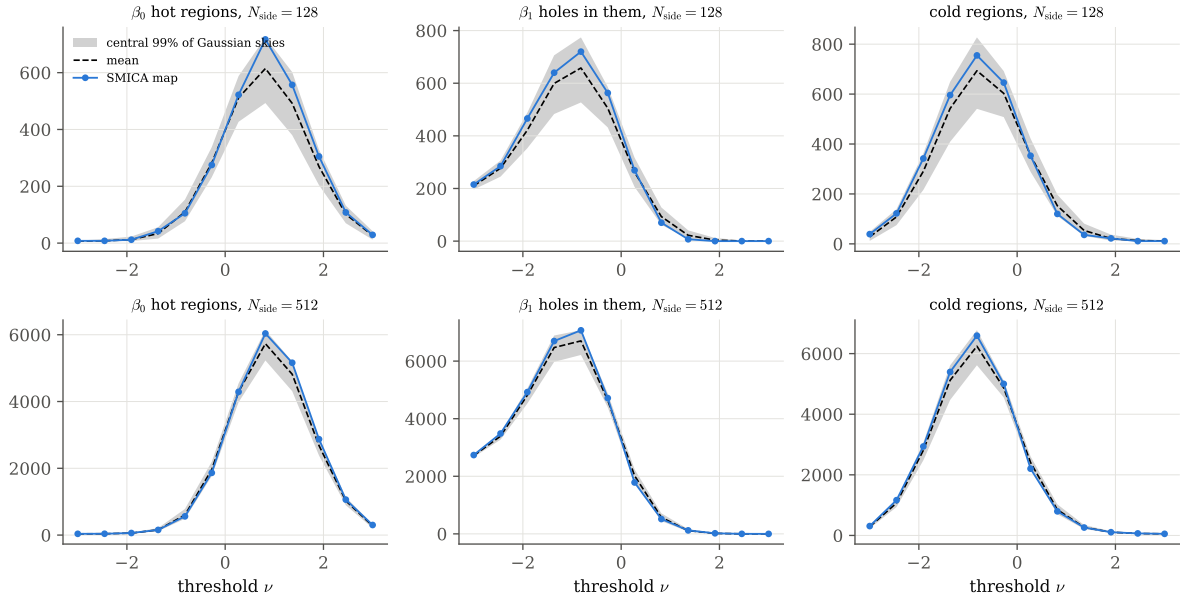


Figure 13.24: Betti curves of the SMICA map (blue) at $N_{\text{side}} = 128$ (80′, top) and 512 (20′, bottom): the number of hot regions β_0, the number of holes in them $\beta_1 = \beta_0 - \chi$, and the number of cold regions, against the central 99% (grey) and the mean (dashed) of the Gaussian skies. The counts are totals over the observed sky.

13.13.2 The whole history: persistence diagrams

The same union–find pass that counts β_0 records, for every hot region, a pair (birth level, death level). Plotted as points, these pairs form the H_0 persistence diagram of the map (Section 12.9): a point far from the diagonal is a hot spot that stood out over a wide range of thresholds, a point near the diagonal a blip. Two features of the cut sky are worth naming. A region cut off from the rest of the observed sky by the mask can never merge with it, so the diagram has one never-dying point for each separate island of observed sky; these carry information about

the mask, not the sky, and are left out. And the levels are in units of the map's own σ_0, so the diagram, like the curves, is insensitive to the overall amplitude.

Comparing a diagram with a population of diagrams. A diagram is a set of points, not a vector, so the χ^2 of Section 13.12 does not apply directly; nor is there a simple "mean diagram" of the Gaussian skies to compare with. What we do have is a distance. The 1-Wasserstein distance W_1 (Section 12.10.2) matches the points of two diagrams one to one, sending unmatched points to the diagonal, and adds up the costs of the best matching; each cost is the larger of the shifts in birth and in death, and the cost of a point sent to the diagonal is half its lifetime.

Example: W_1 by hand. Let $D = \{(2, 0.5)\}$ and $D' = \{(2.2, 0.4), (0.3, 0.2)\}$. Matching $(2, 0.5)$ with $(2.2, 0.4)$ and sending $(0.3, 0.2)$ to the diagonal costs

$$\max(|2 - 2.2|, |0.5 - 0.4|) + \frac{0.3 - 0.2}{2} = 0.2 + 0.05 = 0.25,$$

while sending everything to the diagonal costs $\frac{1.5}{2} + \frac{1.8}{2} + \frac{0.1}{2} = 1.7$, and matching $(2, 0.5)$ with $(0.3, 0.2)$ is worse still. So $W_1(D, D') = 0.25$: the long-lived spot moved by 0.2, and one short blip was created. Unlike the bottleneck distance, which keeps only the worst pair, W_1 adds up every change, so a sky with many slightly displaced spots is far from a Gaussian one even when no single spot moved much.

Computing the best matching. Finding the cheapest matching looks like a search over all pairings, but it is a standard problem in disguise. *The trick:* give every point a private copy of itself on the diagonal. If D has n points and D' has m, list as "rows" the n points of D followed by the m diagonal copies of the points of D', and as "columns" the m points of D' followed by the n diagonal copies of the points of D. A row–column pair is then one of four things: a point of D matched to a point of D' (cost: the larger shift), a point of D sent to its own diagonal copy (cost: half its lifetime), a point of D' sent to its diagonal copy (the same), or two diagonal copies paired with each other (cost 0, nothing happens). Every matching of the two diagrams is a one-to-one assignment of the $n + m$ rows to the $n + m$ columns, and the cheapest assignment of a square cost matrix is what `scipy.optimize.linear_sum_assignment` computes exactly (the Hungarian algorithm). The function `w1` in `code/ch13/b04_betti_persistence.py` builds this matrix in five lines, and on the example above it returns 0.25. Research code uses `gudhi`'s `hera.wasserstein_distance` (order 1, ground metric L^∞), which solves the same problem with a faster approximate auction algorithm.

The test statistic. For each map we compute its mean W_1 distance to a fixed set of reference skies (10 Gaussian skies at $N_{\text{side}} = 64$, 5 at 128). Before matching, points that live less than $0.1\,\sigma_0$ (at 64) or $0.5\,\sigma_0$ (at 128) are dropped. They are the many tiny wiggles of the field, each costing almost nothing, and they make the cost matrix large (the matching time grows as the cube of the number of points) without adding information about the long-lived spots. A Gaussian sky is at a typical distance from the references; a non-Gaussian one, whose spots are born and merge at the wrong levels, would be farther. The null distribution comes from 400 and 200 further Gaussian skies, not among the references, and the p-value is again the rank of the data. At $N_{\text{side}} = 128$ the SMICA diagram keeps 1016 points after the cut (a Gaussian sky about 956). Its mean distance to the references is 99.3, against 99.6 ± 11.9 for Gaussian skies, so 35.5% of Gaussian skies are farther; at $N_{\text{side}} = 64$ the numbers are 47.2 against 45.9 ± 4, and 26.2% (Figure 13.25).

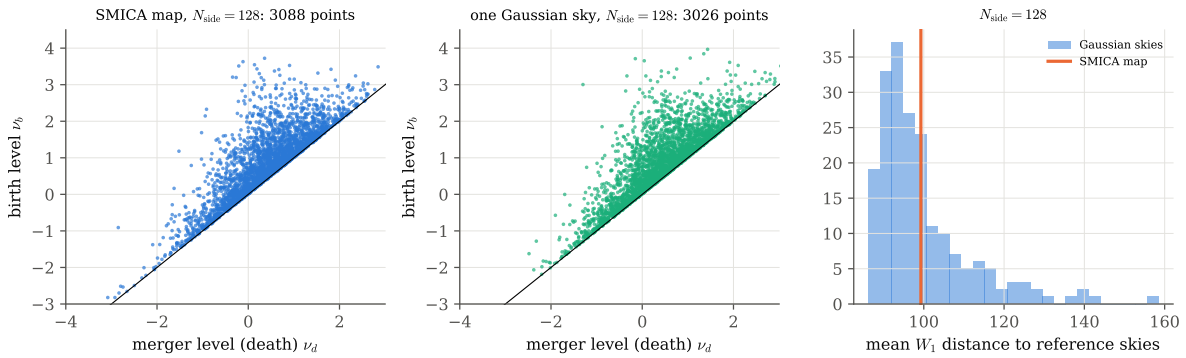


Figure 13.25: Left and middle: H_0 persistence diagrams of the hot regions of the SMICA map and of one Gaussian sky at $N_{side} = 128$, birth level against merger level in units of σ_0; points far above the diagonal are hot spots that stood out over a wide range of thresholds. All finite points are shown; the distances use only those living longer than $0.5\,\sigma_0$. Right: the mean W_1 distance of each Gaussian sky's diagram to the reference skies (histogram), and the same for the SMICA map (vertical line).

Why the noise does not matter here either. The stability theorem (Section 12.10.3) bounds the change of a diagram by the largest change of the field: adding noise of amplitude ϵ (in units of σ_0) moves every point by at most ϵ in the bottleneck distance. At these resolutions the noise is a few hundredths of a per cent of the variance (Section 13.10.2), an amplitude of order $10^{-2}\sigma_0$, so the noise moves the long-lived spots by far less than the sky-to-sky differences of Figure 13.25. Only the very short-lived blips near the diagonal are affected, and W_1 charges them only half their tiny lifetimes.

13.14 Side by side with *Planck* 2018

Their question. By 2018 the *Planck* temperature maps had reached the limit set by cosmic variance over most of the sky. The simplest models of inflation predict a Gaussian, statistically isotropic sky; non-Gaussianity of a few parts in 10^4 would point to more complicated physics in the early universe, and larger departures to foreground residuals or systematic errors in the maps. The collaboration therefore ran a battery of model-independent tests, one of which is the Minkowski-functional test we have just repeated on the same map [114]. The morphological approach on the sky goes back to the analysis of the four-year COBE maps by Schmalzing and Górski, who derived the estimators on the curved, pixelised and masked sphere [110] that Section 12.7.2 reproduced.

Their setup, in plain terms. Four temperature maps, one from each component-separation method; at each of eight resolutions ($N_{side} = 16$ to 2048), the map is degraded and smoothed as in (13.19), a common mask degraded and thresholded at 0.9 removes the Galaxy and the point sources, and the three normalised functionals are measured at twelve thresholds between $-3\sigma_0$ and $3\sigma_0$, with σ_0 the map's own standard deviation. The null hypothesis is represented by 999 FFP10 simulations: lensed Gaussian CMB skies from a fiducial ΛCDM spectrum (with $\tau = 0.060$, slightly above the final best fit), observed with the real beams, scanning and noise, and passed through each component-separation pipeline with the weights derived from the data.

	Planck 2018 VII	ours
maps	Commander, NILC, SEVEM, SMICA	SMICA
resolutions	$N_{\text{side}} = 16, \ldots, 2048$	$N_{\text{side}} = 64, 128, 256, 512$
simulations	999 FFP10, end to end	1000, 1000, 500, 300 Gaussian skies
noise	simulated time streams, map-making	isotropic, half-difference spectrum
statistics	v_0, v_1, v_2 (and needlet v_k)	v_0, v_1, v_2; β_0, β_1; H_0 diagrams
Euler characteristic	$(N_{\text{hot}} - N_{\text{cold}})/2\pi A$	$V - E + F$ per steradian

Their equation, and ours. Their eqs. (20)–(22) define $\chi^2 = \sum_{ij} [\mathsf{C}^{-1}]_{ij} (y_i - \bar{y}_i^{\text{sim}})(y_j - \bar{y}_j^{\text{sim}})$ with $\mathsf{C}_{ij} = \langle (y_i^{\text{sim}} - \bar{y}_i^{\text{sim}})(y_j^{\text{sim}} - \bar{y}_j^{\text{sim}}) \rangle$ over the simulations. Symbol map: their $\boldsymbol{y}$ is our $\boldsymbol{y}$ (the 36 normalised numbers), $\bar{\boldsymbol{y}}^{\text{sim}}$ is our $\bar{\boldsymbol{y}}$, their C our $\widehat{\mathbf{C}}$. This is (13.17) without the Hartlap factor, and it is the quadratic form that Section 13.12.1 derived from the Gaussian density of $\boldsymbol{y}$. Their normalised functionals, eqs. (16)–(19), are (12.35), derived in Section 12.7.4. Their probability, Table 5, is the fraction of simulations with a larger χ^2: our (13.18).

Why can they leave out the Hartlap factor, which Section 13.12.2 insisted on? Because of how the probability is computed. *The question:* does multiplying every χ^2 by the same positive number change the rank of the data among the simulations? It does not: if $\chi_i^2 \geq \chi_{\text{obs}}^2$, then $h\chi_i^2 \geq h\chi_{\text{obs}}^2$ for any $h > 0$, so the count in (13.18) is unchanged. The factor matters only when the χ^2 is compared with the χ_{36}^2 formula, as in the "tail of χ_{36}^2" row of Table 13.6. (In our leave-one-out version the simulations use $n - 1$ skies and the data n, so the two factors differ by one part in n; irrelevant here, but this is why we keep them.) What does matter, with either convention, is that the data and the simulations are compared with the same mean and covariance in the same way.

Their choice of σ_0. The thresholds could be set in units of the population standard deviation (the average over simulations) or of each map's own. *Planck* chose the map's own sample standard deviation, noting that it makes the test more sensitive (their Sect. 5.3 and footnote 9). Section 13.11.2 showed the reason on the real sky: at 160′ the data's σ_0 is below most of the simulations', and with population units that deficit of large-scale power would have shifted every curve and masqueraded as a change of shape.

13.14.1 What agrees, and what differs

N_{side} (FWHM)	64 (160′)	128 (80′)	256 (40′)	512 (20′)
f_{sky}, ours, %	71.3	73.6	74.7	75.5
f_{sky}, *Planck* Table 1, %	71.3	73.6	74.7	75.6
p, ours, %	91.7 ± 0.9	60.3 ± 1.5	50 ± 2.2	23 ± 2.4
p, *Planck*, SMICA, %	90.0	85.4	93.9	80.7
p, *Planck*, range over four methods, %	63.2–99.0	79.7–97.2	81.5–95.4	78.0–87.2

Table 13.7: The masks and the probabilities of a larger χ^2 for the Minkowski-functional test of the temperature map, from our pipeline and from *Planck* 2018 VII, Tables 1 and 5 [114].

What agrees. The masks agree with *Planck*'s: the sky fractions are identical at three resolutions and differ by a tenth of a per cent of the sky at the fourth (Table 13.7), so the degrading and thresholding of Section 13.10.1 is theirs. The conclusion agrees: at every resolution the real sky's χ^2 sits well inside the distribution of Gaussian skies (Figure 13.26), and no probability is

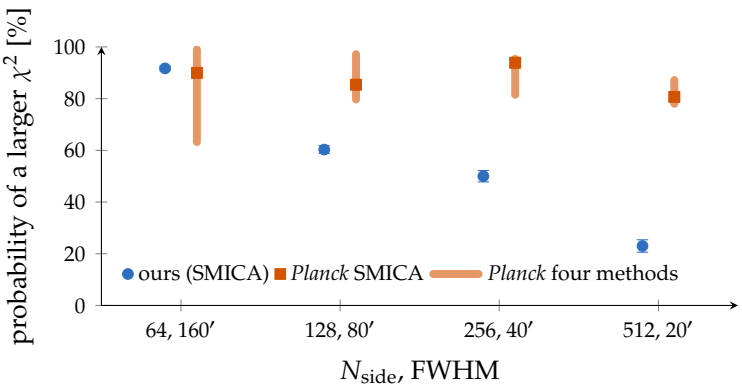


Figure 13.26: The probability of a larger χ^2 at four resolutions: our pipeline on the SMICA map (blue, with its Monte Carlo error), *Planck*'s SMICA result (orange square) and the spread of *Planck*'s four component-separation maps (orange bar). Every value lies in the bulk of the null distribution; no resolution is anomalous in either analysis.

small enough to suggest a departure from Gaussianity. The Betti curves and the persistence diagrams of Section 13.13, which *Planck* did not use, agree with the same conclusion.

What differs, and why. The individual probabilities are not the same numbers, and they should not be expected to be. Three causes, in decreasing order of size.

1. *A probability is a property of a statistic, not only of the sky.* If the null hypothesis is true, the p-value of a valid test is uniformly distributed between 0 and 1 over realisations (Section 4.4.2). Two different statistics computed on the same sky are two correlated draws of such a uniform number; unless the statistics are almost identical, their p-values can differ by tens of per cent. Our estimators measure the same three functionals in a different way (the Euler characteristic by $V - E + F$ instead of hot minus cold spots, the length by the band formula (13.21) instead of contour lengths), and those differences reweight the same fluctuations of the sky. *Planck*'s own Table 5 gives the size of this effect for a milder change: the same test on four cleanings of the same sky spans 81.5 to 95.4% at $N_{\text{side}} = 256$ and 63.2 to 99.0% at 64.
2. *Different simulations.* FFP10 skies are lensed, carry the real noise pattern and go through component separation and inpainting; ours are Gaussian skies with the lensed spectrum and isotropic noise, not inpainted. Lensing makes the sky very slightly non-Gaussian, and the inpainting changes the pixels near the mask after smoothing; both shift the null distribution a little, and differently from the data.
3. *Monte Carlo error.* With 300 to 1000 skies our probabilities carry errors of one to two and a half per cent (Section 13.9.2); *Planck*'s, with 999, about one per cent. This is the smallest of the three.

None of these can turn an unremarkable sky into an anomalous one. A real non-Gaussian signal would push the probability towards zero in every version of the test at once; differences in the processing move it around within the bulk.

Lesson. Reproducing a published p-value means reproducing the published statistic and the published null distribution, not the published conclusion; with our own estimators and our own simulations we reproduce the masks to a tenth of a per cent and the conclusion fully, and the individual probabilities only within the spread that the choice of statistic itself creates. A

test of a single sky is a statement about one draw, and its p-value moves with every honest change of method: what is robust is whether it lands in the tail.

How this is used in real research: from a capstone test to a research result

Planck's analysis differs from ours in what it can rule out, not in its logic. Its simulations follow the data through the real instrument and the real component separation, so a non-Gaussianity created by the pipeline is in the null distribution and cannot be mistaken for the sky; ours would attribute such an effect to the sky. It tests four cleanings of the same data, whose agreement is a test of foreground residuals. It repeats the analysis on needlet maps, each restricted to a band of multipoles, which localises any deviation in scale: there the largest scales ($\ell \lesssim 32$) give probabilities of a few per cent (their Table 6), an echo of the large-angle anomalies (Section T3.5.3). It analyses E-mode polarization the same way. And when a specific non-Gaussian model is in mind, the functionals are fitted rather than only tested: the first-order expansion of [113] (Section T3.2.1) predicts their response to $f_{\rm NL}$, so the measured curves can be turned into an estimate of it; for a known three-point template the bispectrum, which weights every triangle of multipoles by the template, is far more sensitive (Section T3.5.2). Persistence-based statistics are being developed for the same purpose in galaxy surveys [116]. For polarization, the singularities of the Q, U field (Section 12.14) have predicted densities and correlations [1, 2] that Section 12.15.1 reproduced; applying them to *Planck*'s polarization needs the end-to-end noise simulations, because at the small scales where singularities are dense the polarization maps are noise-dominated.

Laptop scale, and what it costs. Each reduction, with what the full analysis does instead:

- *Resolutions. Planck* goes to $N_{\rm side} = 2048$ (5′); we stop at 512 (20′, $\ell \lesssim 800$). The cost: no test of the small scales where noise inhomogeneity and residual point sources would matter.
- *Simulations. Planck* uses 999 end-to-end FFP10 skies per method; we use 1000 to 300 Gaussian skies with isotropic noise from the half-difference spectrum. The cost: Monte Carlo errors of one to two and a half per cent on each probability, Hartlap factors down to 0.876, and a null distribution that does not contain lensing non-Gaussianity, the scanning pattern of the noise, component separation or inpainting.
- *One cleaning. Planck* tests four component-separation maps; we test SMICA only. The cost: no internal check on foreground residuals.
- *Persistence.* The W_1 test runs at $N_{\rm side} = 64$ and 128 against 10 and 5 reference skies; a research version would use all resolutions and a summary designed for sensitivity (persistence images or landscapes with a covariance). The cost: a coarse test whose power against specific non-Gaussian models has not been measured.

Recap

- The real-sky test asks: if the sky were a Gaussian, isotropic field with the best-fit spectrum, observed and processed like the SMICA map, would its topological statistics be as far from the Gaussian average as they are, or farther? The null hypothesis is the one explanation whose sampling distribution we can simulate.
- A map is moved to a working resolution by multiplying its $a_{\ell m}$ by $b_\ell^{\rm new} p_\ell^{\rm new} / (b_\ell^{\rm old} p_\ell^{\rm old})$; the mask by smoothing it with the same beam and keeping pixels with at least 90% trusted weight. The noise spectrum comes from the half-mission half-difference map, whose noise has exactly the full map's statistics.

- On a cut sphere the Euler characteristic is $V - E + F$ with entry levels, the boundary length a band integral of $|\nabla u|$, and $\beta_1 = \beta_0 - \chi$ because an excursion set of a masked sky has $\beta_2 = 0$.
- The χ^2 is the quadratic form of the (approximately Gaussian) statistics; with a simulated covariance it needs the Hartlap factor and leave-one-out simulations, and the probability is the rank of the data among the simulations, with Monte Carlo error $\sqrt{p(1-p)/n}$.
- The SMICA map is consistent with Gaussian skies at $N_{\rm side} = 64$–512 in its Minkowski functionals, its Betti curves and its H_0 persistence diagrams, as *Planck* 2018 found; the individual probabilities differ from *Planck*'s within the spread that different statistics and simulations produce.

13.15 A real galaxy catalogue: CMASS in the southern sky

Everything we know about measuring galaxy clustering we learned on catalogues we made ourselves. In Chapter 11 a lognormal field was sampled by a Poisson process, the points were counted in pairs, and the acoustic bump near 100 h^{-1}Mpc came out of mock surveys whose truth we had chosen. A real survey hands us a list of 208,426 galaxies, each with two angles on the sky and a redshift, and nothing else. Nobody tells us the true correlation function, the true distance to the galaxies, or how many of them the telescope missed. The machinery of Chapter 11 has to work on that list, and when it does, the bump tells us how far away the galaxies are.

Where we have met this measurement before. In Section 11.4.4 we rebuilt the first detection of the acoustic peak [42] on mock surveys: periodic boxes with the volume and galaxy density of the 2005 sample, a covariance matrix from more than a thousand of them, and a fit of the template of Eq. (11.30) with the dilation α. A survey of that size typically detected the peak at about 2σ, and the 2005 paper's 3.4σ sat in the upper tail. In Section 11.4.6 we compared the published DESI distances D_V/r_d with the Planck prediction. Here the estimator and the fit are the same, but three things change. The data are real, so the galaxies sit in a survey with an irregular edge, holes, and a density that falls with distance. There are no mocks, so the error bars must come from the data themselves. And we do not compare published numbers; we produce one, and then set it beside the numbers of the BOSS collaboration, which measured the same galaxies.

The question. Is the acoustic peak visible in the correlation function of real galaxies, how strongly, and what distance $D_V(z)/r_d$ does its position give? Answering it takes five steps: turn the catalogue into positions in space; correct the catalogue for the galaxies the survey could not observe properly; count pairs; estimate the uncertainty from the data alone; and fit the peak.

13.15.1 What the survey observed

The Baryon Oscillation Spectroscopic Survey (BOSS) used the 2.5-metre Sloan telescope in New Mexico between 2009 and 2014. Galaxies were first chosen from images, then a spectrograph measured a redshift for each through an optical fibre plugged into a metal plate at the galaxy's position, about a thousand fibres per plate [276]. The **CMASS** sample ("constant mass") was selected by cuts in colour and magnitude designed to pick galaxies of roughly the same stellar

mass at every distance, massive red galaxies at $0.43 < z < 0.7$. Massive galaxies are rare and bright, so they can be seen far away, and they live in dense regions, so they cluster strongly; both help to see a faint feature at a large scale.

The survey covered two regions of sky on either side of the Milky Way's disc, the northern and the southern Galactic caps. We use only the southern cap: its random catalogue is 1.2 GB, while the northern one is 3.2 GB, more than a laptop analysis should download. Table 13.8 sets our sample beside the full sample.

	CMASS North	CMASS South (ours)	CMASS total
galaxies used, $0.43 < z < 0.7$	568,776	208,426	777,202
effective area [deg^2]	6,851	2,525	9,376
close pairs without a fibre	34,151	11,163	45,314
redshift failures	10,188	5,157	15,345

Table 13.8: The DR12 CMASS sample [Reid2016 Table 2]. After the redshift cut our copy of the southern catalogue holds exactly the 208,426 galaxies of the middle column. The southern cap is 27% of the area.

13.15.2 Two files: the galaxies and the randoms

The public large-scale-structure catalogue for one sample and one cap is two files [276]. The *galaxy file* lists 230,831 objects with right ascension RA, declination Dec (the two angles on the sky), redshift z, four weights that we meet in Section 13.16, and the expected number density $\bar{n}(z)$ at each galaxy's redshift, in a column called `NZ`. Cutting to $0.43 < z < 0.7$ leaves 208,426 galaxies.

The *random file* lists 11,636,252 points, about fifty per galaxy, that mimic everything the survey did except clustering. It is the random catalogue of Section 11.2.1, built for a real survey, and its construction tells us exactly what "everything the survey did" means [Reid2016 §5.2].

- *The angles.* Points are thrown uniformly on the sphere and kept only inside the survey footprint, a union of polygons on the sky. In each small patch of sky (a "sector") a point is kept with a probability equal to the fraction of targets in that sector that received a good redshift, so a patch where the survey was 80% complete gets 80% of the randoms.
- *The holes.* Points are removed wherever galaxies could not have been seen: around bright stars, in fields with bad images, and at the centre of each plate, where the bolt that holds it leaves no room for a fibre.
- *The redshifts.* Each random point is given the redshift of a galaxy chosen at random from the galaxy file, with probability proportional to the galaxy's weight w_{tot}.

The last item needs a reason. The true redshift distribution of CMASS galaxies is not known from theory; it is whatever the colour cuts let through. So it is estimated from the galaxies themselves, and the randoms copy it. The price is that any real fluctuation in the number of galaxies with redshift, a large structure at one distance, is copied into the randoms too, and removed from the measured clustering along the line of sight. Ross et al. compared this "shuffled" choice with smooth fits to $\bar{n}(z)$ on mock catalogues: it gave the smallest systematic error of the methods they tried, and a negligible one for the spherically averaged correlation function that we measure [277]. We keep a random subsample of 10 randoms per galaxy, 2,084,260 points, chosen with `rng.choice`; a subsample of a uniform sample is still uniform.

13.15.3 From two angles and a redshift to a position

A galaxy's redshift is not a distance. To place it in space we need the distance–redshift relation, and that needs a cosmology. We take the comoving distance from the cosmology notes, for a flat universe:

$$D_C(z) = \int_0^z \frac{c\,\mathrm{d}z'}{H(z')} = \frac{c}{H_0}\int_0^z \frac{\mathrm{d}z'}{E(z')}, \qquad E(z) = \sqrt{\Omega_m(1+z)^3 + 1 - \Omega_m}, \tag{13.26}$$

with radiation neglected in E, which is accurate to better than 0.1% below $z = 1$ (Section T1.2 lists the full set of distances). The value of D_C we actually use comes from CAMB's background module, which keeps radiation and neutrinos; the formula is how to think about it.

Why surveys measure in h^{-1}Mpc. Write $H_0 = 100\,h\,\mathrm{km\,s^{-1}Mpc^{-1}}$, so $c/H_0 = 2997.9\,h^{-1}\mathrm{Mpc}$. The integral in (13.26) depends only on Ω_m. So in units of h^{-1}Mpc the distances of all galaxies depend on one parameter, Ω_m, and h never enters. A survey of redshifts cannot measure h by itself, because every distance in it scales with $1/h$ together; quoting lengths in h^{-1}Mpc removes that unknown common factor from every number. (It is the same invisible calibration that Section 11.20 met: an overall distance scale that the data cannot see.)

The position. A galaxy at angles $(\mathrm{RA}, \mathrm{Dec}) = (\phi, \delta)$ lies along the unit vector $\hat{\boldsymbol{n}}$, and its comoving position is

$$\boldsymbol{x} = D_C(z)\,\hat{\boldsymbol{n}}(\phi,\delta), \qquad \hat{\boldsymbol{n}} = (\cos\delta\cos\phi,\ \cos\delta\sin\phi,\ \sin\delta). \tag{13.27}$$

Because the observed redshift includes the galaxy's own motion along the line of sight, these are redshift-space positions (Section 11.5.1), and separations between them are written s rather than r. The cosmology we chose is the Planck 2018 model used throughout the book, with $\Omega_m = 0.315$. If it is wrong, every separation is stretched by a few per cent, in the way Section 11.4.1 worked out: the acoustic peak then appears at the wrong place, by the factor α that the fit will measure. A wrong fiducial cosmology is not an error in the analysis; it is what the analysis measures.

Example: the distance to $z = 0.565$ by hand. The weighted mean redshift of our galaxies turns out to be $z_{\mathrm{eff}} = 0.565$ (Section 13.16). Simpson's rule with four intervals of width $\Delta z = 0.14125$, with $\Omega_m = 0.3152$, needs $1/E$ at five points: 1, 0.93116, 0.86076, 0.79194, 0.72683. Then

$$D_C(0.565) \overset{(1)}{=} 2997.9\,h^{-1}\mathrm{Mpc} \times \frac{\Delta z}{3}\Big[1 + 4(0.93116) + 2(0.86076) + 4(0.79194) + 0.72683\Big]$$
$$\overset{(2)}{=} 2997.9 \times 0.04708 \times 10.3410\ h^{-1}\mathrm{Mpc} = 1459.6\,h^{-1}\mathrm{Mpc}.$$

(1) Simpson's rule, $\int_a^b f \approx \frac{\Delta z}{3}[f_0 + 4f_1 + 2f_2 + 4f_3 + f_4]$, applied to $1/E(z)$, times c/H_0.
(2) $\Delta z/3 = 0.047083$, and the bracket adds to 10.3410.

CAMB, with radiation and neutrinos, gives $1460\,h^{-1}$Mpc. The sample spans $D_C = 1153$ to $1742\,h^{-1}$Mpc, a shell $590\,h^{-1}$Mpc thick, about six acoustic scales.

Example: the separation of a pair. Two galaxies 3° apart on the sky, at $z = 0.55$ and $z = 0.56$, have $D_C = 1426.7$ and $1448.6\,h^{-1}$Mpc. Their separation follows from the law of cosines for the triangle formed with the observer,

$$s = \sqrt{D_1^2 + D_2^2 - 2D_1D_2\cos\theta} = 78.4\,h^{-1}\mathrm{Mpc}.$$

The transverse part is about $D\,\theta = 75\,h^{-1}$Mpc and the radial part $D_2 - D_1 = 22\,h^{-1}$Mpc. A change of 0.01 in redshift is a radial step of $22\,h^{-1}$Mpc, which is why redshift errors of $\sim 10^{-4}$ do not matter for the acoustic scale, while galaxy velocities of a few hundred km/s, $\Delta z \sim 10^{-3}$, do (Section 11.5).

Example: how dense is the sample? The effective area of $2525\,\mathrm{deg}^2$ is a solid angle $\Omega = 2525\,(\pi/180)^2 = 0.769$ sr. The comoving volume of the shell between D_1 and D_2 seen through that solid angle is $\Omega(D_2^3 - D_1^3)/3 = 0.96\,h^{-3}\mathrm{Gpc}^3$, and the mean density is $\bar{n} = 208{,}426/V = 2.17 \times 10^{-4}\,h^3\mathrm{Mpc}^{-3}$. The density peaks at 4.23×10^{-4} near $z = 0.5$. In the language of Section 11.3.2, with $P \approx 10^4\,h^{-3}\mathrm{Mpc}^3$ at the acoustic scale, $\bar{n}P$ is between 1 and 4: the sample is dense enough that sample variance, not shot noise, limits the measurement. The effective volume, $V_{\mathrm{eff}} = \int \mathrm{d}V\,[\bar{n}P_0/(1+\bar{n}P_0)]^2$ (Eq. (11.24)), is $0.40\,h^{-3}\mathrm{Gpc}^3$, against 0.13–0.55 for the 2005 sample of Section 11.4.4.

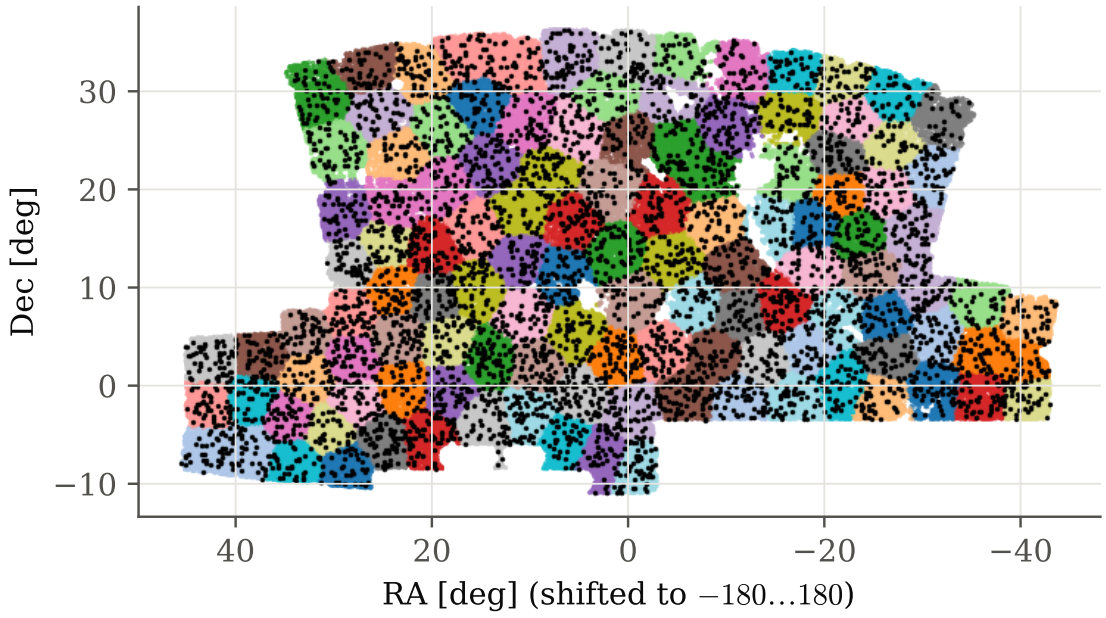


Figure 13.27: The southern CMASS footprint. Coloured dots: a subsample of the randoms, coloured by the jackknife region they belong to (Section 13.18); black dots: a subsample of the galaxies. The sky is seen from inside, so RA increases to the left. The ragged edge and the small holes are the mask the randoms copy.

The code. Figures 13.27 and 13.28 show the result of `code/ch13/21_cmass_catalogue.py`. The lines that matter are these.

Listing 13.7: From angles and redshift to a position (`code/ch13/21_cmass_catalogue.py`)

```
def comoving_distance_table():
    import camb
    p = camb.set_params(H0=FIDUCIAL["H0"], ombh2=FIDUCIAL["ombh2"], omch2=FIDUCIAL["omch2"],
                        mnu=FIDUCIAL["mnu"], ns=FIDUCIAL["ns"], As=FIDUCIAL["As"])
    bg = camb.get_background(p)
    zt = np.linspace(0.0, 1.0, 2001)
    h = FIDUCIAL["H0"] / 100
    dc = bg.comoving_radial_distance(zt) * h                  # Mpc -> Mpc/h
    hz = bg.hubble_parameter(zt)                              # km/s/Mpc
    rd = bg.get_derived_params()["rdrag"]                     # Mpc
    return zt, dc, hz, rd, h
```

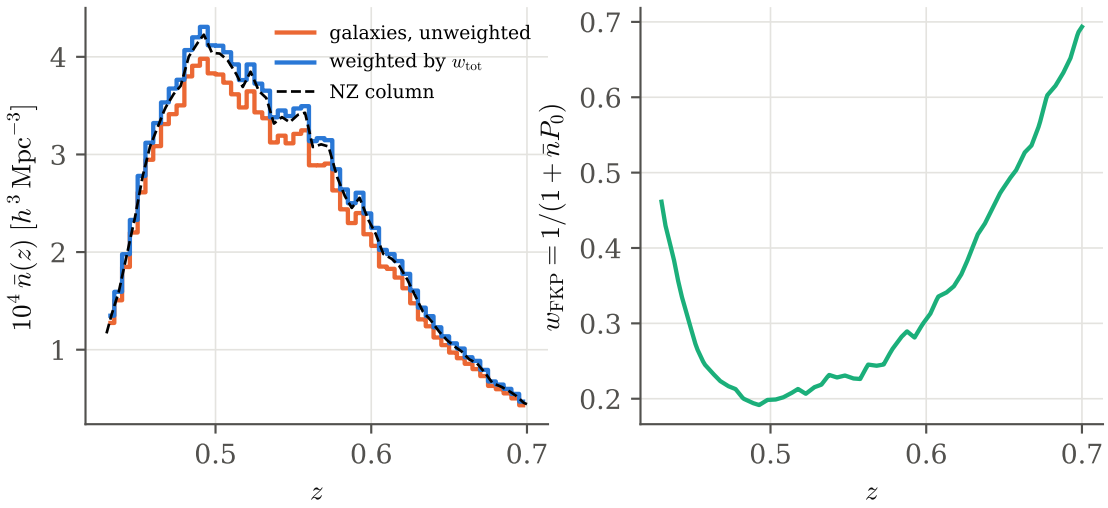


Figure 13.28: Left: the number density of CMASS South galaxies against redshift, counted in shells of the fiducial cosmology, without weights (orange) and with the weight $w_{\rm tot}$ of Section 13.16 (blue), and the catalogue's `NZ` column (dashed), which matches the weighted histogram. Right: the FKP weight each galaxy receives; it is lowest where the sample is densest.

```
67
68
69 zt, dct, hzt, RD_MPC, H = comoving_distance_table()
70 dc_of_z = lambda z: np.interp(z, zt, dct)
71
72
73 def to_xyz(ra, dec, z):
74     """Comoving Cartesian position: distance D_C(z) along the unit vector of (RA, Dec)."""
75     r = dc_of_z(z)
76     a, d = np.radians(ra), np.radians(dec)
77     return np.column_stack([r * np.cos(d) * np.cos(a), r * np.cos(d) * np.sin(a), r *
       np.sin(d)])
```

`comoving_distance_table` evaluates CAMB's $D_C(z)$ once on a fine grid and multiplies by h to work in h^{-1}Mpc; every galaxy's distance is then a linear interpolation in that table, and `to_xyz` is (13.27). The script also reads the weights (Section 13.16), keeps 10 randoms per galaxy, and assigns each object to one of 120 patches of sky for the jackknife (Section 13.18).

13.16 What the survey could not see: the weights

The random catalogue of Section 13.15.2 encodes the survey completeness across the sky and the mean density as a function of redshift. Additional corrections are needed for losses that depend on the galaxies themselves or on the observing conditions. A galaxy that sits right next to another one may not get a fibre; a faint galaxy may not give a usable spectrum; a patch of sky crowded with stars may yield fewer galaxy targets. None of these has anything to do with cosmology, and all of them imprint a pattern on the galaxy counts that would be read as clustering. The catalogue fixes them by attaching a **weight** to each galaxy: a number that says how many galaxies this one stands for.

The principle behind every weight. The model of Section 11.1 is that the galaxies are a Poisson sample of the density $\bar{n}(\boldsymbol{x})[1 + \delta_g(\boldsymbol{x})]$, where $\bar{n}(\boldsymbol{x})$ is the density the survey would

record if galaxies were not clustered. Suppose an effect that has nothing to do with cosmology multiplies the probability of recording a galaxy at $\boldsymbol{x}$ by a factor $f(\boldsymbol{x})$, so the recorded galaxies are a Poisson sample of $f\bar{n}(1+\delta_g)$. If each recorded galaxy is counted with weight $w = 1/f$, the expected weighted count in a small volume $\mathrm{d}V$ is

$$\mathbb{E}\Big[\sum_{i\in \mathrm{d}V} w_i\Big] \overset{(1)}{=} \frac{1}{f(\boldsymbol{x})}\,\mathbb{E}\big[N(\mathrm{d}V)\big] \overset{(2)}{=} \frac{1}{f(\boldsymbol{x})}\, f(\boldsymbol{x})\,\bar{n}(\boldsymbol{x})\big[1+\delta_g(\boldsymbol{x})\big]\mathrm{d}V = \bar{n}(\boldsymbol{x})\big[1+\delta_g(\boldsymbol{x})\big]\mathrm{d}V. \quad (13.28)$$

(1) All galaxies in the small volume share the same f, so the weight comes out of the sum, which is then the number $N(\mathrm{d}V)$ of recorded galaxies.
(2) The mean of a Poisson count is its intensity times the volume (Eq. (11.2)).

So the weighted galaxies are, on average, the galaxies a perfect survey would have recorded. The inputs are physical ones we should state: the loss factor f must not depend on the cosmological field δ_g itself, and it must be known, or measurable from something other than the galaxy clustering. Each weight below is a way of estimating $1/f$ for one kind of loss, and each one bends one of these inputs a little.

13.16.1 Close pairs: the galaxy that did not get a fibre

Two fibres cannot be plugged closer than 62″ on one plate [Reid2016 §6.1]. When two targets are closer than that and only one plate covers them, one of them gets no spectrum. At $z = 0.55$, 62″ is a transverse distance $D_C\theta = 1427 \times (62/206265) = 0.43\,h^{-1}$Mpc. The lost galaxy is, more often than not, a physical companion of the one that was observed: among close pairs where both members did get redshifts (because two plates overlapped), 52% belong to a component concentrated within about $5\,h^{-1}$Mpc of each other along the line of sight, and the rest are chance projections [Reid2016 §6.2, Fig. 10]. So the loss depends on the clustering itself, and the random catalogue, which knows nothing about clustering, cannot reproduce it.

The catalogue's fix is to give the lost galaxy's weight to its nearest neighbour with a good redshift: the column `WEIGHT_CP` is $w_{\mathrm{cp}} = 1$ plus the number of collided targets a galaxy stands for (Figure 13.29). This preserves the number of galaxies in every region much larger than $0.4\,h^{-1}$Mpc: one galaxy counted twice in place of two galaxies a few h^{-1}Mpc apart moves some pairs by a few h^{-1}Mpc, which hardly matters in bins $8\,h^{-1}$Mpc wide at the acoustic scale. Because the galaxy that got the fibre was chosen at random within the colliding group, it is statistically equivalent to the one that did not, and the correction is unbiased at transverse separations larger than the collision scale; below it, pairs are missing, which we do not fit [Reid2016 §6.2]. In our sample 4.3% of the galaxies carry $w_{\mathrm{cp}} > 1$, and they stand for 10164 collided targets in all.

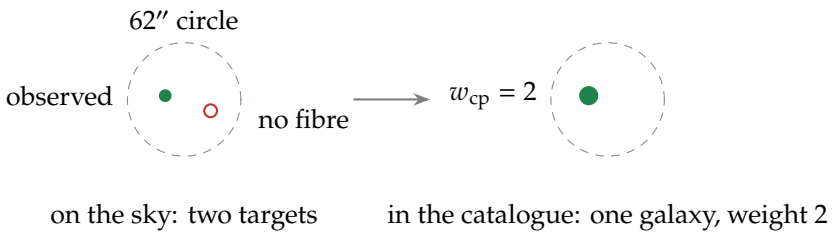


Figure 13.29: Nearest-neighbour upweighting. Left: two targets closer than 62″; only one received a fibre. Right: the observed galaxy carries the missing one's weight. Counts in any region much larger than the circle are unchanged on average.

13.16.2 Redshift failures, stars and seeing

Redshift failures. Some galaxies got a fibre but no reliable redshift, more often on fibres near the edge of a plate and for galaxies that look fainter through the fibre [276, 277]. Again the missing galaxy's weight goes to its nearest neighbour with a good redshift, in the column `WEIGHT_NOZ`; a small refinement in DR12 moves weight towards fainter neighbours, so the values are not exactly integers [Reid2016 §6.3]. In our sample 2.0% of galaxies carry $w_{\rm noz} > 1$, standing for 4396 failures.

Stars, and the reasoning that finds them. *The event.* Ross et al. found that the number of CMASS targets per square degree falls in parts of the sky where many stars are seen [277]. *What could have produced it?* Two explanations are possible. Either the regions near the Galactic plane, where stars are many, happen to hold fewer galaxies, or stars interfere with finding galaxies. *How likely is the observation under each?* The galaxies are billions of light years away and the stars are in our own Galaxy. Under the first explanation the galaxy density would know nothing about the star density, and a steady trend with it across the whole survey would be a coincidence. Under the second the trend is expected: a star close to a galaxy on the sky hides it or spoils the measurement of its brightness and colours, on which the selection cuts act. The trend is also stronger for galaxies of lower surface brightness, the ones most easily spoiled, and only the second explanation predicts that [277]. So the density is modelled as the true density times $A + B\,n_s$, a linear function of the local star density n_s, with A and B fitted in bins of galaxy brightness [Reid2016 eq. (45)], and the weight is the inverse, $w_{\rm star} = (A + Bn_s)^{-1}$: (13.28) with $f = A + Bn_s$. A second weight $w_{\rm see}$ does the same for the blurring of the images by the atmosphere, the seeing, and the catalogue stores their product, $w_{\rm systot} = w_{\rm star}\, w_{\rm see}$ [Reid2016 eq. (49)]. In our sample 95% of the galaxies have $0.918 < w_{\rm systot} < 1.263$, with mean 1.011. The correction is small per galaxy, but it varies smoothly over tens of degrees, and on large scales that is exactly the kind of pattern that masquerades as clustering: without it the stellar trend imprints an error larger than the statistical one at $s > 120\,h^{-1}$Mpc [277].

13.16.3 Putting the weights together

How many galaxies does one galaxy stand for? A galaxy with a good redshift stands for itself, plus the $w_{\rm cp} - 1$ collided targets given to it, plus the $w_{\rm noz} - 1$ failures given to it. All of these stand-ins are at the same place on the sky, so the same star and seeing correction applies to each. Therefore

$$w_{\rm tot} \overset{(1)}{=} w_{\rm systot}\left[1 + (w_{\rm cp} - 1) + (w_{\rm noz} - 1)\right] \overset{(2)}{=} w_{\rm systot}\left(w_{\rm cp} + w_{\rm noz} - 1\right) \tag{13.29}$$

(1) The count of galaxies this one represents, each corrected by the angular factor $1/f = w_{\rm systot}$ of (13.28).
(2) Collect the ones.

which is the total weight of the DR12 catalogues [Reid2016 eq. (50)], [Ross2012 eq. (25)]. Writing the sum minus one, rather than the product $w_{\rm cp}w_{\rm noz}$, says that a collided target and a failed redshift are two different missing galaxies, not one galaxy missing twice.

Example: the books balance. Summed over our 208,426 galaxies, $\sum(w_{\rm cp} + w_{\rm noz} - 1) = 222986$: the galaxies with redshifts plus the 10164 collided targets and the 4396 failures that they stand for. With the angular weights the total is $\sum w_{\rm tot} = 225469$, about 1% more, the mean of $w_{\rm systot}$. Figure 13.28 shows the redshift distribution before and after weighting; the weighted histogram is the one the randoms copy, and it matches the catalogue's `NZ` column.

The FKP weight is a choice, not a correction. The weights so far undo losses. The last one decides how much each part of the survey should count. In Section 11.3.2 we derived that the variance of the measured clustering at wavenumber k is smallest when each region is weighted by $w = 1/(1+\bar{n}P)$, where P is the power at that k (Eq. (11.23)). A sparse region is noisy, because its fluctuations are mostly shot noise, so it gets weight near 1; a dense region is limited by sample variance, so piling up more of its galaxies helps little, and each of them is down-weighted. The catalogue uses $P_0 = 10^4\,h^{-3}\mathrm{Mpc}^3$, the observed power near $k \approx 0.15\,h\,\mathrm{Mpc}^{-1}$ [Reid2016 eq. (53)], so $w_{\mathrm{FKP}} = 1/(1+\bar{n}(z)\,10^4)$. Recomputing it from the `NZ` column reproduces the stored values to $6.5e-08$. Across our sample it runs from 0.19 at the density peak to 0.69 at the far edge (Figure 13.28, right).

Why the randoms carry the FKP weight but not w_{tot}. The randoms already follow $\bar{n}(\boldsymbol{x})$, the density of a complete survey: their angular density follows the completeness and their redshifts follow the weighted galaxies. The galaxies needed w_{tot} to be brought to that same density. The FKP weight is different: it multiplies the field we analyse, galaxies minus randoms (Eq. (11.17)), and both terms must be multiplied by the same function of position, or the weighted field would not average to zero where there is no clustering. So the final weights are [Reid2016 §6.7]

$$w_i^{\mathrm{gal}} = w_{\mathrm{tot},i}\,w_{\mathrm{FKP},i}, \qquad w_j^{\mathrm{ran}} = w_{\mathrm{FKP},j}. \tag{13.30}$$

Example: one galaxy's weight. A galaxy at $z = 0.50$, where $\bar{n} = 4.2\times10^{-4}\,h^3\mathrm{Mpc}^{-3}$, that received the weight of one collided neighbour ($w_{\mathrm{cp}} = 2$, $w_{\mathrm{noz}} = 1$) in a slightly star-crowded field ($w_{\mathrm{systot}} = 1.05$) has $w_{\mathrm{tot}} = 1.05\times(2+1-1) = 2.10$ and $w_{\mathrm{FKP}} = 1/(1+4.2) = 0.19$, so it enters the pair counts with weight 0.40. An isolated galaxy at $z = 0.68$, where $\bar{n} = 0.7\times10^{-4}$, has $w_{\mathrm{tot}} \approx 1$ and $w_{\mathrm{FKP}} = 0.59$: it counts for half again as much as the first, although the first stands for two galaxies.

13.16.4 Weighted pair counts, and why they still estimate ξ

With weights, the normalised counts of Eq. (11.11) become weighted sums, each divided by the total weight of all possible pairs of its kind:

$$DD(s) = \frac{\sum_{i\neq j} w_i w_j\,[s_{ij}\in\mathrm{bin}]}{\left(\sum_i w_i\right)^2 - \sum_i w_i^2}, \quad DR(s) = \frac{\sum_{i,j} w_i w'_j\,[s_{ij}\in\mathrm{bin}]}{\sum_i w_i \sum_j w'_j}, \tag{13.31}$$

and RR like DD with the random weights w'. Here $[\,\cdot\,]$ is 1 if the pair's separation falls in the bin and 0 otherwise, and the sums run over ordered pairs. The denominator of DD is $\sum_{i\neq j} w_i w_j$, the weight of all ordered pairs with $i \neq j$: the square of the sum contains every product $w_i w_j$ once for each order, including $i = j$, and subtracting $\sum w_i^2$ removes those.

Why the ratio still measures $1+\xi$. Take the galaxies as a Poisson sample of intensity $\lambda(\boldsymbol{x}) = f\bar{n}(1+\delta_g)$, each carrying the weight $w(\boldsymbol{x})$ of (13.30), so that $w\lambda = w_{\mathrm{FKP}}\bar{n}(1+\delta_g)$ on average by (13.28). Write $u(\boldsymbol{x}) \equiv w_{\mathrm{FKP}}(\boldsymbol{x})\bar{n}(\boldsymbol{x})$ for this weighted mean density. For distinct points of a Cox process the density of pairs at $\boldsymbol{x}_1, \boldsymbol{x}_2$ is $\bar{n}_1\bar{n}_2[1+\xi(s)]$ (Eq. (11.6)), so

$$\begin{aligned}\mathbb{E}\Big[\sum_{i\neq j} w_i w_j [s_{ij}\in\mathrm{bin}]\Big] &\overset{(1)}{=} \int\!\!\int u(\boldsymbol{x}_1)\,u(\boldsymbol{x}_2)\,\big[1+\xi(s_{12})\big][s_{12}\in\mathrm{bin}]\,\mathrm{d}V_1\mathrm{d}V_2 \\ &\overset{(2)}{=} \big[1+\bar{\xi}_{\mathrm{bin}}\big]\int\!\!\int u(\boldsymbol{x}_1)\,u(\boldsymbol{x}_2)\,[s_{12}\in\mathrm{bin}]\,\mathrm{d}V_1\mathrm{d}V_2.\end{aligned} \tag{13.32}$$

(1) The pair density times the two weights, integrated over all positions; the weight times the recorded density is u, by (13.28).
(2) In a bin narrow enough that ξ hardly changes across it, $1+\xi$ comes out as its value $1+\bar{\xi}_{\rm bin}$ in the bin.

The random catalogue has intensity proportional to $\bar{n}$ and weights $w_{\rm FKP}$, so its weighted pair count is the same integral without the factor $1+\bar{\xi}$, multiplied by a constant that the normalisations remove. The two survey integrals cancel in the ratio, and $DD/RR \to 1+\bar{\xi}_{\rm bin}$; the Landy–Szalay combination of Section 11.2.1 inherits the same property, with its reduced sensitivity to the edges. The derivation also shows what can go wrong: if the galaxies' weighted density $w\lambda$ and the randoms' weighted density differ in shape, through a systematic not captured by a weight, the cancellation fails and the difference is read as clustering.

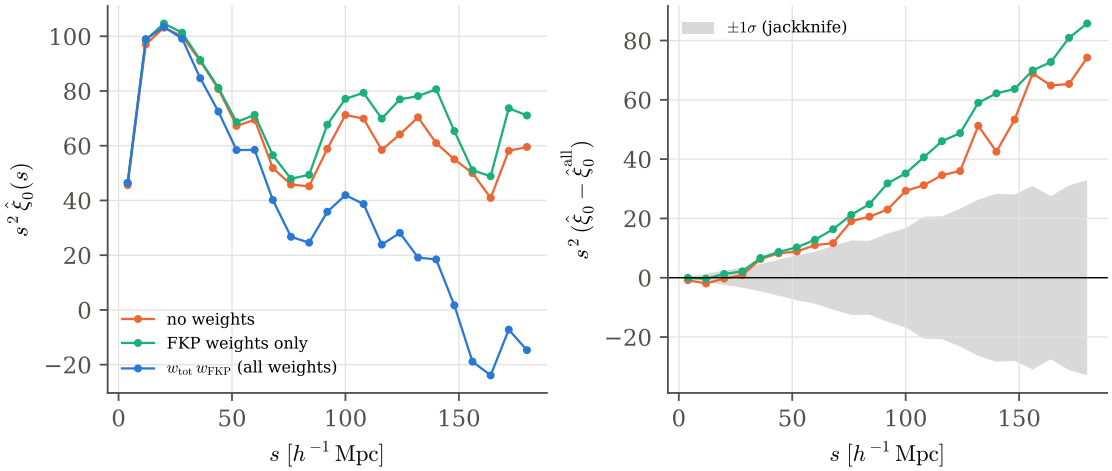


Figure 13.30: What the weights do to the measured monopole. Left: $s^2\hat{\xi}_0$ with no weights, with FKP weights only, and with all weights. Right: the difference from the fully weighted measurement, against the jackknife error band of Section 13.18. Without $w_{\rm tot}$ the clustering beyond 35 h^{-1}Mpc is too high, by more than one error bar over most of that range and by about two at the largest separations.

Figure 13.30 shows the measurement of Section 13.17 with three choices of weights. Below 30 h^{-1}Mpc the three agree to a few per cent. In the first bin, which holds ξ at 4 h^{-1}Mpc, the unweighted value is 1.8% lower, and with the small jackknife error there that is a 2σ difference; the FKP-only measurement reproduces the fully weighted value to a tenth of an error bar. So at small scales it is the redistribution of weight by the FKP factor, not the corrections for the lost galaxies, that matters, and the lost companions of Section 13.16.1, a few per cent of the galaxies, move ξ at 4 h^{-1}Mpc by much less than its error. The large scales are where the weights act. Without $w_{\rm tot}$ the monopole lies above the weighted one at every separation beyond 35 h^{-1}Mpc. The shift is nearly the same in ξ at all of these separations, about 0.003, which is 1–1.5 jackknife errors at 40–70 h^{-1}Mpc and about two errors at 150–180 h^{-1}Mpc, where the error bars are smaller (the right panel multiplies both by s^2). A nearly constant offset is not a feature at one scale but a change in the mean density that the pairs are measured against. That is the signature Ross et al. traced to the density of foreground stars [277]: where stars are dense, fewer galaxies are targeted, and since the stellar density varies smoothly over tens of degrees, the missing galaxies imprint a pattern on very large scales that pair counts read as extra clustering. The weight $w_{\rm star}$ removes the pattern, as (13.28) says it should. Beyond 35 h^{-1}Mpc the FKP weight alone moves the monopole by at most two-thirds of an error bar, since it only reweights galaxies by redshift. In all three versions the bump near 100 h^{-1}Mpc is present; what differs is the smooth shape it sits on, which the broad-band terms of the fit in

Section 13.19 are designed to absorb. Without the corrective weights the broad band would be wrong by more than the statistical error, and any analysis of the full shape of ξ, rather than only of the peak position, would be biased.

13.17 Counting pairs: the correlation function of CMASS South

We now have 208,426 weighted galaxies and, for the counts, 1,042,130 randoms (5 per galaxy) at known positions. The Landy–Szalay estimator of Section 11.2.1,

$$\hat{\xi}_0(s) = \frac{DD(s) - 2DR(s) + RR(s)}{RR(s)}, \tag{13.33}$$

with the weighted, normalised counts of (13.31), asks for three numbers in each separation bin. The subscript 0 says that pairs are counted in every direction, so the result is the angle average, the monopole, of the redshift-space correlation function. The difficulty is only the number of pairs: the randoms alone form 10^{12} ordered pairs, and a double loop that computes each distance would take days.

13.17.1 How a tree counts pairs without visiting them

A k-d tree sorts the points into nested boxes: the whole set is split in half along its longest side, each half again, and so on, until each box holds a few dozen points. For each box the tree stores its corners and the total weight of its points. To count the pairs between two boxes in our bins, `cKDTree.count_neighbors` looks first at the smallest and the largest distance that any point in one box can have from any point in the other, which follow from the corners alone.

- If both lie inside the same bin, every pair between the two boxes is in that bin, and the product of the two total weights is added at once.
- If both lie beyond the last bin edge, no pair counts, and the boxes are skipped.
- Otherwise the boxes are split and their children are compared.

Example. Two boxes of side 10 h^{-1}Mpc whose centres are 100 h^{-1}Mpc apart contain pairs from $100 - 17$ to $100 + 17$ h^{-1}Mpc (the diagonal of a cube of side 10 is 17), which spans five of our 8 h^{-1}Mpc bins; they must be opened. Two boxes of side 2 h^{-1}Mpc at the same distance span at most 3.5 h^{-1}Mpc, and most such pairs of boxes fall inside one bin. The work is therefore spent on box pairs that straddle a bin edge, a small fraction of all pairs, and the counts are exact. (The script of Section 11.2.4 used the same routine in two dimensions.) The cost still grows with the number of pairs inside the largest separation, which is why we stop at 184 h^{-1}Mpc and use 5 randoms per galaxy rather than the fifty of the BOSS analyses [278]. The random catalogue's own counting noise shrinks as the number of randoms grows; at five per galaxy it is a small addition to the counting noise of the galaxies, which is itself smaller than the sample variance at the acoustic scale for a sample with $\bar{n}P$ between 1 and 4 (Section 13.15.3).

13.17.2 Counts for the jackknife at no extra cost

Section 13.18 will need the correlation function recomputed 120 times, each time with one patch of sky left out. Recounting all pairs each time would multiply the cost by 120. It is not necessary, because a pair count is a sum over pairs, and a sum can be split by where the pairs live.

Divide the footprint into regions $k = 1, \dots, 120$ (the colours of Figure 13.27). For two sets of points A and B write $C(A, B)$ for the weighted count of ordered pairs with the first member in A and the second in B, in a given bin, and let D be all galaxies, D_k those in region k, and R, R_k the same for the randoms. Leaving region k out removes every pair with at least one member in D_k. Those are the pairs whose first member is in D_k, plus those whose second member is in D_k, minus the pairs counted twice, which have both members in D_k:

$$DD_{-k} \overset{(1)}{=} C(D,D) - \big[C(D_k,D) + C(D,D_k) - C(D_k,D_k)\big]$$
$$\overset{(2)}{=} C(D,D) - 2\,C(D_k,D) + C(D_k,D_k), \tag{13.34}$$
$$DR_{-k} \overset{(3)}{=} C(D,R) - C(D_k,R) - C(D,R_k) + C(D_k,R_k), \tag{13.35}$$

and RR_{-k} like DD_{-k} with R in place of D.

(1) The number of elements in a union of two sets is the sum of their numbers minus the number in their intersection (inclusion–exclusion).
(2) A distance does not care about the order of the pair, so $C(D, D_k) = C(D_k, D)$.
(3) For galaxy–random pairs the two members are of different kinds: remove pairs whose galaxy is in region k and pairs whose random is in region k, and add back the pairs where both are.

So for each region we compute five counts, $C(D_k, D)$, $C(D_k, D_k)$, $C(D_k, R)$, $C(D, R_k)$, $C(D_k, R_k)$, and two more for the randoms, $C(R_k, R)$ and $C(R_k, R_k)$. Summing $C(D_k, D)$ over k gives the full count $C(D, D)$, so the full measurement is a by-product, and the total work is about that of one full count. The normalisations of (13.31) are recomputed for each leave-one-out sample from the sums of the weights outside region k.

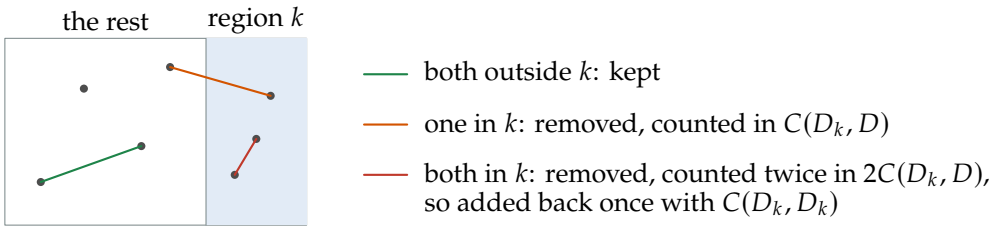


Figure 13.31: Leaving one region out of a pair count. A pair with one member in region k appears once, in $C(D_k, D)$ or in $C(D, D_k)$ according to its order. A pair with both members in k appears in both of those, so subtracting $2C(D_k, D)$ removes it twice, and $C(D_k, D_k)$ restores one copy.

13.17.3 The code, and what it measured

Listing 13.8: Region-by-region pair counts and the Landy–Szalay estimator (`code/ch13/22_cmass_xi.py`)

```
def region_counts(args):
    """Counts that involve region k: C(k, all) and C(k, k) for DD, DR (both ways) and RR."""
    k, wname = args
    wgal, wran = WEIGHTS[wname]
    ig, ir = np.flatnonzero(rg == k), np.flatnonzero(rr == k)
    tgk, trk = cKDTree(xg[ig]), cKDTree(xr[ir])
    out = {
        "dd_ka": pairs(tgk, TG, wgal[ig], wgal), "dd_kk": pairs(tgk, tgk, wgal[ig],
    wgal[ig]),
        "dr_ka": pairs(tgk, TR, wgal[ig], wran), "dr_ak": pairs(TG, trk, wgal, wran[ir]),
        "dr_kk": pairs(tgk, trk, wgal[ig], wran[ir]),
```

```
    }
    if wname != "fkp":                          # "fkp" has the same random weights as "full"
        out["rr_ka"] = pairs(trk, TR, wran[ir], wran)
        out["rr_kk"] = pairs(trk, trk, wran[ir], wran[ir])
    out["wg"], out["wg2"] = wgal[ig].sum(), (wgal[ig] ** 2).sum()
    out["wr"], out["wr2"] = wran[ir].sum(), (wran[ir] ** 2).sum()
    return out


def landy_szalay(DD, DR, RR, Wg, Sg, Wr, Sr):
    """xi_LS = (DD - 2 DR + RR)/RR with each count divided by its total number of weighted
    pairs."""
    dd = DD / (Wg ** 2 - Sg)                    # ordered pairs i != j: (sum w)^2 - sum w^2
    dr = DR / (Wg * Wr)
    rr_ = RR / (Wr ** 2 - Sr)
    return (dd - 2 * dr + rr_) / rr_
```

`pairs` calls the tree with the bin edges $0, 8, \ldots, 184\,h^{-1}$Mpc and the weights of both sets, and drops the first entry, which holds the pairs at zero distance: a point paired with itself. `region_counts` builds small trees for the galaxies and randoms of one region and counts them against the full trees, the five plus two counts listed above. `landy_szalay` is (13.33) with the normalisations of (13.31). The regions are independent of each other, so the script farms them out to 8 processes; the fully weighted measurement took 338 s.

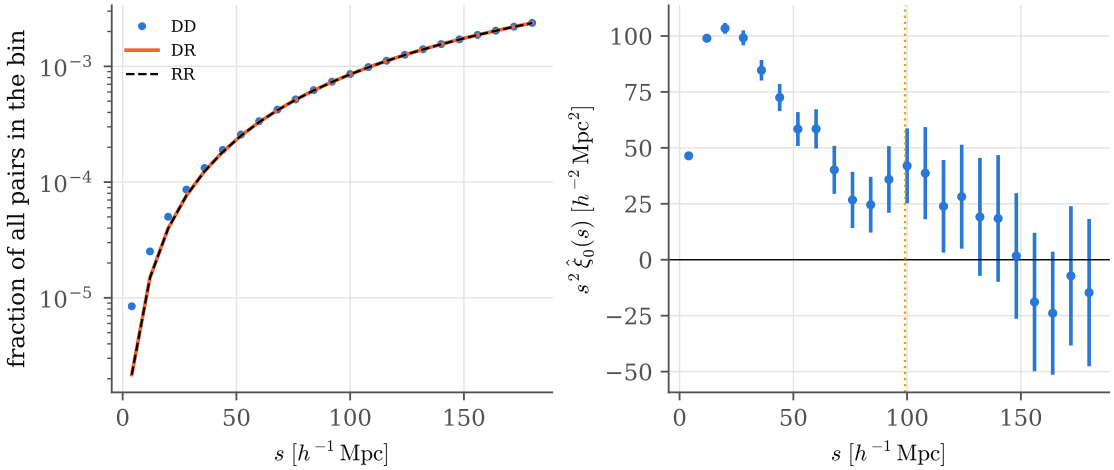


Figure 13.32: Left: the three weighted, normalised pair counts. Beyond $50\,h^{-1}$Mpc they differ by a few per cent at most, too little to see on a logarithmic axis; the correlation function lives in these small differences. Right: the Landy–Szalay monopole times s^2, with jackknife error bars (Section 13.18). The dotted line marks the fiducial sound horizon, $r_d h = 99.09\,h^{-1}$Mpc.

Figure 13.32 is the measurement. The three counts on the left follow the geometry of the survey: they rise as s^2, since a shell of radius s holds a volume growing as s^2, and grow more slowly once the shell starts to leave the survey. Only at small separations does DD visibly exceed RR; beyond $50\,h^{-1}$Mpc they agree to a few per cent, and the correlation function is that small excess. On the right, multiplying by s^2 makes the slow decay of ξ_0 visible on a linear axis, and a modest bump rises above the declining trend near $100\,h^{-1}$Mpc, close to the fiducial ruler $r_d h$. At $s = 100\,h^{-1}$Mpc we measure $\hat{\xi}_0 = (4.20 \pm 1.67) \times 10^{-3}$: a one-part-in-two-hundred excess of pairs, read off from 1.092×10^9 ordered galaxy pairs closer than $184\,h^{-1}$Mpc. Whether the bump is the acoustic peak or a fluctuation is a question for an error bar and a fit.

13.18 Error bars from the data alone: the jackknife

In the mock surveys of Section 11.4.4 the covariance of $\hat{\xi}$ came from fifteen hundred simulated boxes. The BOSS analyses did the same with about a thousand simulated copies of the real survey, each with its footprint, its density and its redshift distribution [224, 278]. Producing those copies is a research project in itself. We have one sky. The question an error bar answers is: if we could observe another patch of the Universe with the same shape and the same galaxies, how different would $\hat{\xi}_0$ be? We cannot observe another patch. But the survey is made of many smaller patches, and how much the answer moves when one of them is left out tells us how much it depends on which galaxies we happened to observe. That is the idea of the **jackknife**, a cousin of the bootstrap of Section 3.15 and of the leave-one-group-out error of Section 11.12.3.

13.18.1 The simplest case first: the mean of N numbers

Take N independent measurements $x_1, \dots, x_N$ with a common variance σ^2, and the estimate $\bar{x}$. Its variance is σ^2/N, estimated by s^2/N with $s^2 = \sum_k (x_k - \bar{x})^2/(N-1)$. The jackknife reaches the same number without knowing that the estimate is a mean. Leave out measurement k and recompute: $\bar{x}_{-k} = (N\bar{x} - x_k)/(N-1)$. The average of these N leave-one-out values is $\bar{x}$ itself, because each x_j is left out exactly once. How far does each leave-one-out value sit from it?

$$\bar{x}_{-k} - \bar{x} \overset{(1)}{=} \frac{N\bar{x} - x_k - (N-1)\bar{x}}{N-1} = \frac{\bar{x} - x_k}{N-1},$$

$$\frac{N-1}{N}\sum_{k=1}^{N}\big(\bar{x}_{-k} - \bar{x}\big)^2 \overset{(2)}{=} \frac{N-1}{N}\,\frac{1}{(N-1)^2}\sum_{k=1}^{N}(x_k - \bar{x})^2 \overset{(3)}{=} \frac{1}{N}\,\frac{\sum_k (x_k - \bar{x})^2}{N-1} = \frac{s^2}{N}. \tag{13.36}$$

(1) Put both terms over the common denominator $N-1$.
(2) Square the first line and sum over k.
(3) Cancel one factor of $N-1$.

So the jackknife variance, $\frac{N-1}{N}\sum_k (\bar{x}_{-k} - \bar{x})^2$, is exactly the usual estimate of the variance of a mean. The prefactor is the surprising part. For N independent repetitions we would divide by $N-1$; here we multiply by $N-1$. The leave-one-out values are far closer to each other than independent repetitions would be, because any two of them share $N-2$ of their $N-1$ measurements: each deviation is the deviation of one measurement shrunk by $1/(N-1)$, and the prefactor undoes that shrinkage.

Example: three numbers. For $x = (1, 2, 6)$, $\bar{x} = 3$ and $s^2 = (4 + 1 + 9)/2 = 7$, so $s^2/N = 2.33$. Leaving one out gives 4, 3.5, 1.5, which average to 3 with deviations 1, 0.5, −1.5; the jackknife gives $\frac{2}{3}(1 + 0.25 + 2.25) = 2.33$. The same number, from a recipe that only needed to recompute the estimate.

Working definition 13.2 (jackknife covariance)

Split the data into N subsets [here, 120 patches of sky of similar area]. Recompute the estimate with subset k removed, giving the vector $\hat{\xi}_{-k}$ [the binned correlation function without patch k], for each k, and let $\bar{\xi}$ be their average. Then

$$\hat{\mathbf{C}}_{\mathrm{JK}} = \frac{N-1}{N}\sum_{k=1}^{N}\big(\hat{\xi}_{-k} - \bar{\xi}\big)\big(\hat{\xi}_{-k} - \bar{\xi}\big)^{\mathsf{T}}. \tag{13.37}$$

Operationally: Section 13.17.2 gives every $\hat{\xi}_{-k}$ from region-by-region pair counts at the cost of one measurement.

The off-diagonal elements come for free: if leaving out a patch raises $\hat{\xi}$ in one bin, it usually raises it in the neighbouring bins too, because the same pairs and the same structures contribute to both, and (13.37) records that co-movement as a covariance.

13.18.2 What the jackknife gets wrong

For a mean, (13.36) is exact. For any other statistic of independent data, Efron and Stein proved that the jackknife variance is biased upward in expectation: on average it is too large, never too small [279]. The smallest example is the square of a mean, $\hat{\theta} = \bar{x}^2$, for N Gaussian measurements of mean μ and variance σ^2. Its true variance is that of the square of a Gaussian variable of mean μ and variance σ^2/N, namely $4\mu^2\sigma^2/N + 2\sigma^4/N^2$. With $d_k = \bar{x}_{-k} - \bar{x} = (\bar{x} - x_k)/(N-1)$ from (13.36),

$$\begin{aligned} \hat{\theta}_{-k} - \hat{\theta} &\overset{(1)}{=} (\bar{x} + d_k)^2 - \bar{x}^2 = d_k\,(2\bar{x} + d_k), \\ \frac{N-1}{N}\sum_k \big(\hat{\theta}_{-k} - \hat{\theta}\big)^2 &\overset{(2)}{=} \frac{N-1}{N}\Big[4\bar{x}^2\sum_k d_k^2 + 4\bar{x}\sum_k d_k^3 + \sum_k d_k^4\Big] \overset{(3)}{=} \frac{4\bar{x}^2 s^2}{N} + O(N^{-3}), \\ \mathbb{E}\Big[\frac{4\bar{x}^2 s^2}{N}\Big] &\overset{(4)}{=} \frac{4}{N}\Big(\mu^2 + \frac{\sigma^2}{N}\Big)\sigma^2 = \frac{4\mu^2\sigma^2}{N} + \frac{4\sigma^4}{N^2}. \end{aligned} \tag{13.38}$$

(1) Square out the bracket.
(2) Square the first line and sum. The sum is centred on $\hat{\theta}$ rather than on the average of the leave-one-out values, which differs from it by $O(N^{-2})$ and changes the result only at order N^{-3}.
(3) $\sum_k d_k^2 = s^2/(N-1)$ by (13.36). For a Gaussian sample $\bar{x}$ is independent of the deviations $x_k - \bar{x}$, so $\sum_k d_k^3$, odd in them, averages to zero, and $\sum_k d_k^4 \sim 3\sigma^4/N^3$.
(4) $\mathbb{E}[\bar{x}^2] = \mu^2 + \sigma^2/N$, because $\bar{x}$ has mean μ and variance σ^2/N, and $\mathbb{E}[s^2] = \sigma^2$.

The first term of the jackknife variance is the standard error-propagation result $(2\mu)^2\sigma^2/N$, and its next term, $4\sigma^4/N^2$, is twice the $2\sigma^4/N^2$ of the true variance. The jackknife is too large, here by a second-order amount that fades as N grows, in line with the theorem. A correlation function from patches of sky bends the inputs of that result in three ways.

- *The patches are not independent.* Neighbouring patches share large structures, and a fluctuation larger than a patch moves several of them together; the leave-one-out values then move together too, and that part of the variance is under-represented. Fluctuations on scales larger than the whole survey are invisible to any internal method.
- *Pairs straddle the boundaries.* Removing a patch removes the pairs between it and its neighbours as well. Our patches are about 21 deg^2, some 120 h^{-1}Mpc across on the sky at the sample's distance, so at the acoustic scale most of the pairs that a patch removes cross its boundary.
- *The patches are unequal.* Ours hold between 1018 and 2466 galaxies, while the derivation assumed exchangeable pieces.

Norberg et al. compared jackknife errors with the scatter of many simulated surveys [280]. The jackknife returned the right variance on large scales and too large a variance on small scales, and its errors were noisy enough that an interval quoted as 2σ from it could correspond to anything from 1σ to 3σ. We should therefore read our error bars as good to some tens of per cent, which is the price of not having mocks.

Inverting a noisy matrix. The fit of Section 13.19 needs $\hat{\mathbf{C}}^{-1}$, and the inverse of a noisy covariance estimate is biased high: noise makes some directions look better measured than

they are. For a covariance from N independent Gaussian samples and n_b bins, multiplying the inverse by $(N - n_b - 2)/(N - 1)$ removes the bias (Section 8.13.2). With $N = 120$ jackknife samples and the $n_b = 19$ bins of the fit this factor is 0.832. Jackknife samples are not independent draws, so the factor is only a guide; it is the reason for using many regions.

13.18.3 A check against the published mocks

There is one test we can make. Cuesta et al. published the CMASS monopole of the whole sample, north and south, with error bars from 956 simulated surveys [278]. How should our errors compare with theirs? The variance of a clustering measurement comes from the finite number of independent Fourier modes in the volume, and that number is proportional to the volume (the mode count of Section 11.4.5). At the same redshifts and density the error therefore scales as $V^{-1/2}$, so a sample with 3.71 times our area should have errors $\sqrt{9376/2525} = 1.93$ times smaller.

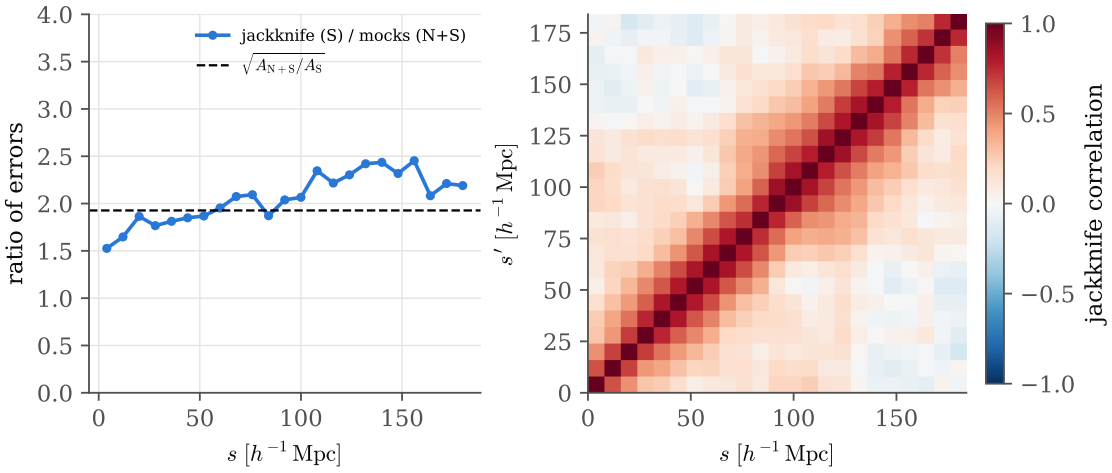


Figure 13.33: Left: our jackknife error on $\hat{\xi}_0$ (south only) divided by the mock error of the published north-plus-south measurement, bin by bin; the dashed line is the ratio $\sqrt{A_{N+S}/A_S}$ expected from the areas alone. Right: the correlation matrix of our jackknife covariance; neighbouring bins are strongly correlated, as they share pairs and structures.

Figure 13.33 shows the comparison. Over the fitted range $30 < s < 180\,h^{-1}$Mpc the median ratio is 2.09, against 1.93 expected from the areas: our jackknife errors are 1.09 times what scaling the mock errors would give. That is within the tens of per cent expected of a jackknife, and since the two error bars come from entirely different routes, one from the data alone and one from a thousand simulations, it is a real test of both. The right panel shows that neighbouring bins are strongly correlated. A fit that used only the diagonal errors would count the same information several times, which is why the fit uses the full matrix.

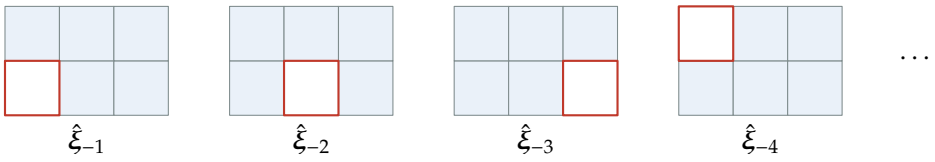


Figure 13.34: The jackknife in pictures: the footprint (six patches drawn here; 120 in the analysis) is measured again and again, each time with one patch removed (red). The spread of the leave-one-out results, multiplied by $(N - 1)/N$, is the covariance.

13.19 The acoustic peak of CMASS South, and the BOSS distance

The ratio $D_V(z)/r_d$ has appeared twice before. In Section 11.4.1 we saw why a galaxy survey measures it: the acoustic shell has a comoving radius r_d fixed in the early Universe, and the survey sees that shell at a size set by the distance to the galaxies, so only the ratio of the two is observed. In Section 11.4.6 we set the published DESI values of D_V/r_d beside the prediction of the Planck 2018 model and found them within one or two standard deviations of it, with errors of one to two per cent. Those were numbers someone else had measured. Now the same ratio comes out of our own pipeline, at the redshift $z_{\text{eff}} = 0.565$ of the CMASS galaxies, from the correlation function $\hat{\xi}_0(s)$ of Section 13.17 and the jackknife covariance of Section 13.18. What is new is that nothing has been done for us: the coordinates, the weights, the pair counts and the error bars are all ours, and the comparison is with collaborations that measured the same galaxies.

13.19.1 Their question

When BOSS released its ninth data set in 2012, the acoustic peak had been seen in galaxy surveys for seven years, but never with enough galaxies to measure its position to better than a few per cent. Anderson et al. asked whether the first 264,283 CMASS galaxies could detect the peak beyond reasonable doubt and turn it into a distance at $z = 0.57$ precise to two per cent [262]. They found the peak at about 5σ and $D_V(0.57)/r_s = 13.67 \pm 0.22$, a 1.7% distance [Anderson2012 abstract, Table 2]. Four years later the survey was complete. Cuesta et al. repeated the measurement on all 777,202 CMASS galaxies of the final release [278], and Alam et al. combined it with the other BOSS analyses into the collaboration's final distances at three redshifts [224].

Our question is narrower. With 27% of the final sample (the southern cap, Table 13.8) and with error bars from the data alone, do we see the same peak, how strongly, and do we get a distance consistent with theirs?

13.19.2 Their setup, in plain terms

Every BAO analysis makes the same four choices, and the numbers in Table 13.9 are those choices for the three analyses we compare.

A fiducial cosmology. To turn $(\text{RA}, \text{Dec}, z)$ into positions one needs a distance–redshift relation (Section 13.15.3), so each analysis picks a cosmology and works in its coordinates. Anderson et al. used a model close to WMAP's ($\Omega_m = 0.274$, $h = 0.7$); Cuesta et al. used the cosmology of their simulated surveys ($\Omega_m = 0.29$, $h = 0.7$); we use the Planck 2018 model of Section 13.15.3. A wrong choice does not bias the answer; it only moves the peak, which is exactly what α measures.

A template. The model fitted to the data is (11.30): a damped acoustic template ξ_t, stretched by α, with an amplitude B^2 and three smooth terms $a_0 + a_1/s + a_2/s^2$ that absorb what is not modelled. The template is computed in some cosmology, and its bump sits at that cosmology's sound horizon, which we call the **template ruler** r_d^{tmpl}.

A covariance. Anderson et al. and Cuesta et al. took it from 600 and 956 simulated copies of the survey; we take it from 120 jackknife regions.

A fit range and binning. All use the monopole in bins of $8\,h^{-1}$Mpc; Cuesta et al. and we fit $30 < s < 180\,h^{-1}$Mpc.

	Anderson et al. 2012	Cuesta et al. 2016	ours
sample	DR9 CMASS	DR12 CMASS, N+S	DR12 CMASS, S
galaxies	264,283	777,202	208,426
effective area [deg^2]	3,275	9,376	2,525
fiducial Ω_m, h	0.274, 0.70	0.29, 0.70	Planck 2018
template ruler [Mpc]	153.19 (fitting formula)	147.10	147.10
$D_V^{\rm fid}(z)$ [Mpc]	2026.49 at 0.57	2009.55 at 0.57	2048 at 0.565
covariance	600 mocks	956 mocks	120 jackknife regions
damping $\Sigma_{\rm nl}$ [h^{-1}Mpc]	8	from mocks	8

Table 13.9: The three analyses of the CMASS monopole before reconstruction [Anderson2012 §§3–5], [Cuesta2016 §§2–4, Tables 2, 5]. Our template ruler happens to agree with Cuesta et al.'s to the quoted digits, although the two cosmologies differ: theirs has a lower h and a higher $\Omega_m h^2$.

13.19.3 Their key equation: from a stretch to a distance

All three papers report D_V/r_d through one equation, $D_V/r_d = \alpha\, D_V^{\rm fid}/r_d^{\rm tmpl}$. Section 11.4.1 derived it in the special case where the coordinates and the template use the same cosmology. Comparing three analyses with three cosmologies, and refitting one paper's data with our template, needs the general case, with the coordinate cosmology ("fid") and the template cosmology ("tmpl") kept apart.

The event. The fit finds the template's bump best matched when it is stretched by α. *What does that say about the Universe?* Follow the true acoustic shell, of radius r_d, into the data, and the template's bump into the model, and ask where each one lands.

$$
\begin{aligned}
r_{\rm app} &\overset{(1)}{=} r_d\,\frac{D_V^{\rm fid}}{D_V}, \\
\xi_t(\alpha s)\text{ peaks at }\alpha s = r_d^{\rm tmpl} &\overset{(2)}{=} \quad s = \frac{r_d^{\rm tmpl}}{\alpha}, \\
\frac{r_d^{\rm tmpl}}{\alpha} = r_d\,\frac{D_V^{\rm fid}}{D_V} &\overset{(3)}{=} \quad \frac{D_V}{r_d} = \alpha\,\frac{D_V^{\rm fid}}{r_d^{\rm tmpl}}.
\end{aligned}
\tag{13.39}
$$

(1) A true separation r is placed at $r\,D^{\rm fid}/D$ in the fiducial coordinates (Eq. (11.26)); for the angle-averaged monopole the relevant distance is the volume average D_V of (11.27). Here $r_{\rm app}$ is where the true shell appears in the data.

(2) The template $\xi_t(s)$ has its bump at $s = r_d^{\rm tmpl}$; the stretched model $\xi_t(\alpha s)$ has it where its argument equals $r_d^{\rm tmpl}$.

(3) The fit puts the model's bump where the data's bump is; solve for D_V/r_d.

In the special case of Section 11.4.1, $r_d^{\rm tmpl} = r_d^{\rm fid}$ and (13.39) is $D_V/r_d = \alpha\,(D_V/r_d)^{\rm fid}$, which is (11.27).

Where h goes. Positions are measured in h^{-1}Mpc. A distance in these units, $D^{\rm fid}[h^{-1}{\rm Mpc}] = (c/100\,{\rm km\,s^{-1}})\int_0^z {\rm d}z'/E(z')$, depends only on $\Omega_m^{\rm fid}$ and not on h; the template, computed from a power spectrum in $h\,{\rm Mpc}^{-1}$, has its bump at $r_d^{\rm tmpl}[{\rm Mpc}]\times h^{\rm tmpl}$. So (13.39) holds with both $D_V^{\rm fid}$ and $r_d^{\rm tmpl}$ in h^{-1}Mpc, each converted with its own cosmology's h, and the result is a pure number.

Example: our own conversion. In our Planck coordinates $D_V^{\rm fid}(0.565) = 1379.5\,h^{-1}$Mpc, and our template's ruler is $r_d^{\rm tmpl}h = 99.09\,h^{-1}$Mpc, so $D_V^{\rm fid}/r_d^{\rm tmpl} = 13.92$. This is the Planck prediction itself, because the fiducial model is the Planck model: a fit that returned $\alpha = 1$ would say that the galaxies are where Planck says they should be.

Example: Cuesta et al.'s number. Their fit gave $\alpha = 1.0153 \pm 0.0134$ before reconstruction [Cuesta2016 Table 10], in coordinates with $D_V^{\rm fid}(0.57) = 2009.55$ Mpc and with a template whose ruler is 147.10 Mpc; both use $h = 0.7$, which cancels. Then

$$\frac{D_V}{r_d} = 1.0153 \times \frac{2009.55}{147.10} = 1.0153 \times 13.661 = 13.87, \qquad \sigma = 0.0134 \times 13.661 = 0.18,$$

the value quoted in their §4.4 (where systematic errors raise the error to 0.19).

Example: a ruler with a different name. Anderson et al. quoted $D_V/r_s = 13.44 \pm 0.22$ from the correlation function [Anderson2012 Table 2], where r_s is the sound horizon from a fitting formula of Eisenstein and Hu [109], 153.19 Mpc in their fiducial model. CAMB's integral (11.25) for the same model gives $r_d = 149.28$ Mpc, the value BOSS adopted for that cosmology in later papers [Cuesta2016 §3]. Both are names for the same physical ruler; they differ because the formula is a fit to the integral, and their ratio is almost independent of cosmology near the fiducial model. By (13.39) the fitted α is the same whatever the ruler is called, so

$$\frac{D_V}{r_d} = \frac{D_V}{r_s} \times \frac{r_s^{\rm fid}}{r_d^{\rm fid}} = 13.44 \times \frac{153.19}{149.28} = 13.44 \times 1.0262 = 13.79,$$

with error 0.23. A 2.6% difference in convention is larger than the 1.6% error of the measurement itself, which is why a paper's ruler must be identified before its number is compared with anything.

13.19.4 What we compute

The script `code/ch13/23_cmass_bao.py` fits the model (11.30) to two data sets with the same function: our monopole with the jackknife covariance, and the published monopole of all CMASS galaxies with its mock covariance.

Listing 13.9: The fit: $\chi^2(\alpha)$ with the linear coefficients profiled out (`code/ch13/23_cmass_bao.py`)

```
def fitter(s, edges, d, C, nmock):
    """chi^2(alpha) for both templates with the linear coefficients profiled out (GLS)."""
    m = (s > SMIN) & (s < SMAX + 1)
    s_, d_, nb = s[m], d[m], int(m.sum())
    Cm = C[np.ix_(m, m)]
    hart = (nmock - nb - 2) / (nmock - 1)
    Ci = hart * np.linalg.inv(Cm)
    e = np.concatenate([edges[:-1][m], [edges[1:][m][-1]]])
    res = {}
    for name, xf in [("w", xi_w_fine), ("nw", xi_nw_fine)]:
        chi2, coefs = [], []
        for a in ALPHAS:
            X = np.column_stack([binned(xf, e, a), np.ones(nb), 1 / s_, 1 / s_ ** 2])
            b = np.linalg.solve(X.T @ Ci @ X, X.T @ Ci @ d_)
            r = d_ - X @ b
            chi2.append(r @ Ci @ r)
            coefs.append(b)
        chi2 = np.array(chi2)
```

```
        i = int(np.argmin(chi2))
        p = np.exp(-0.5 * (chi2 - chi2.min()))
        p /= p.sum()
        mean = np.sum(p * ALPHAS)
        sd = np.sqrt(np.sum(p * ALPHAS ** 2) - mean ** 2)
        res[name] = dict(chi2=chi2, i=i, alpha=ALPHAS[i], mean=mean, sd=sd, coef=coefs[i],
    chimin=chi2[i])
    res["dchi2"] = res["nw"]["chimin"] - res["w"]["chimin"]
    res["nb"], res["hart"], res["mask"], res["edges"] = nb, hart, m, e
    return res
```

The mask keeps the bins of the fit range. The inverse covariance is multiplied by the Hartlap factor of Section 13.18.2, here with the number of jackknife regions standing in for the number of simulations. For each α on a grid from 0.8 to 1.2, the design matrix holds the stretched template, bin-averaged by `binned` with weight s^2 because the number of pairs in a thin shell grows as s^2, and the three smooth terms; the next line is the generalised least-squares solution for (B^2, a_0, a_1, a_2) (Section 3.10.5). The resulting $\chi^2(\alpha)$ is turned into a posterior with a flat prior on the grid, $p(\alpha) \propto e^{-\chi^2(\alpha)/2}$ (Section 11.4.2), whose mean and standard deviation are reported. The same loop with the no-wiggle template gives the $\Delta\chi^2$ of (11.31). The template is the one of Section 11.4.2: the Planck linear spectrum from CAMB, its wiggles damped with $\Sigma_{\rm nl} = 8\,h^{-1}$Mpc as appropriate before reconstruction.

13.19.5 Side by side

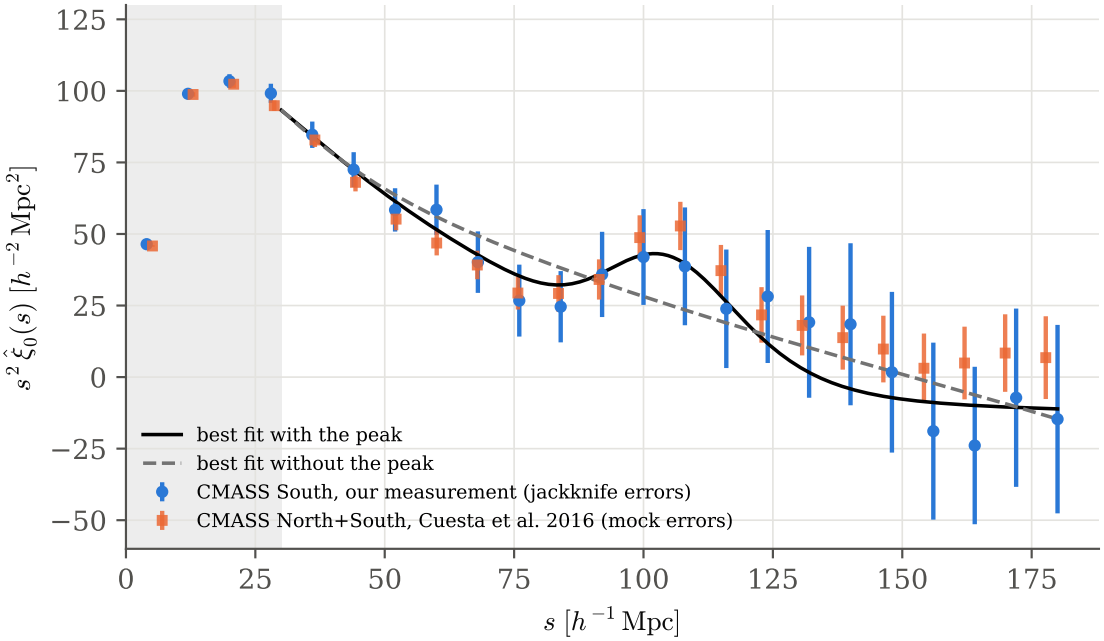


Figure 13.35: Our monopole of the southern cap (blue) and the published monopole of the whole CMASS sample (orange), both times s^2. The published separations are rescaled by 0.9807, the ratio of the two fiducial D_V, to put them in our coordinates. Lines: our best fits with and without the acoustic feature. The grey band is outside the fit range.

Figure 13.35 sets the two measurements on one axis. Below 90 h^{-1}Mpc they agree point by point within the error bars. Both show the bump. Ours is lower between 105 and 120 h^{-1}Mpc, and beyond 150 h^{-1}Mpc our points scatter by more than the published ones, as their error bars, about twice as large, say they should. The southern cap is part of the whole sample, so the

agreement at small separations is not a test of independent data; it is a test that our coordinates, weights and estimator reproduce theirs.

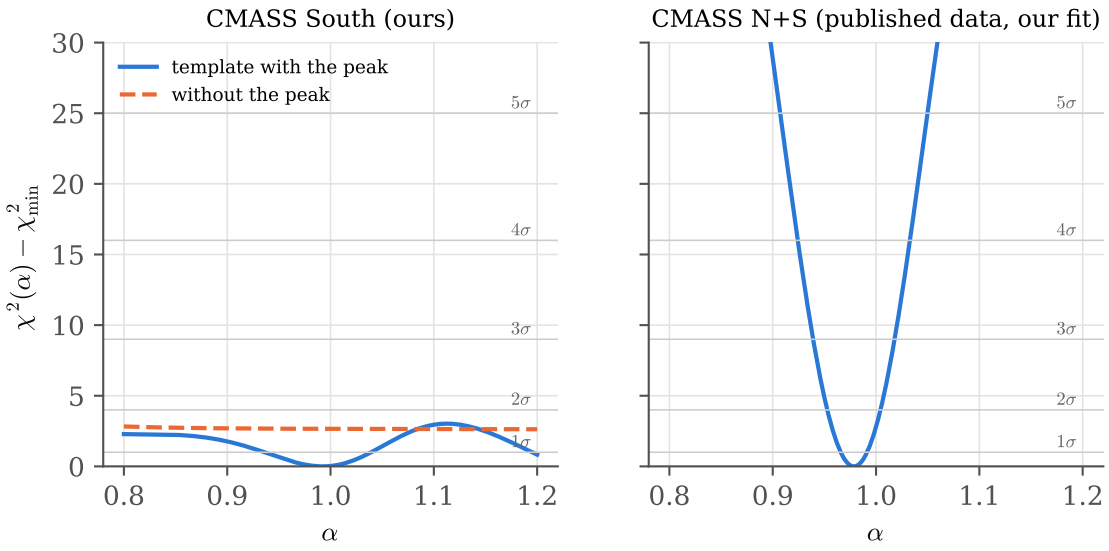


Figure 13.36: $\chi^2(\alpha) - \chi^2_{\rm min}$ for the template with the acoustic feature (solid) and without it (dashed), for our southern measurement (left) and for the published monopole of the whole sample refitted with our code (right). Grey lines: the $\Delta\chi^2 = n^2$ levels.

The published data, refitted. The right panel of Figure 13.36 is the textbook case. $\chi^2(\alpha)$ is a deep parabola with its minimum at $\alpha = 0.979 \pm 0.013$; the template without a peak is worse by $\Delta\chi^2 = 37.2$, a 6.1σ detection by the rule of (11.31). With (13.39), our template and their fiducial distance $D_V^{\rm fid} h = 1406.7\, h^{-1}$Mpc, this is $D_V/r_d = 13.89 \pm 0.18$, against their 13.87 ± 0.18. Our α differs from theirs by almost four per cent while the distances agree to a fraction of a per cent. Nothing went wrong: α is not a property of the Universe but of the pair (coordinates, template), and our template ruler in h^{-1}Mpc, 99.09 against $147.10 \times 0.7 = 102.97$, is shorter than theirs by the same fraction. Their result is reproduced by an independent code with an independent template.

Our southern measurement. The left panel is a different picture. The curve has a clear minimum at $\alpha = 0.992$, with $\chi^2_{\rm min} = 10.6$ for 14 degrees of freedom, a good fit. But it never rises far: within the grid it climbs at most 3.0 above its minimum, and the template without a peak is worse by only $\Delta\chi^2 = 2.6$. Anderson et al. separated two questions that a curve like this keeps apart [Anderson2012 §5].

- *Is there a peak?* The gap between the two curves at their minima. Ours is 2.6, so the acoustic feature is preferred at 1.6σ. As a test (Chapter 4), the null hypothesis is the smooth template, whose noise we know: if there were no acoustic feature, would a gain of 2.6 or more in χ^2 be something to expect? The tail of χ^2 with one degree of freedom beyond 2.6 is 0.11, one chance in nine, so yes, and the extra freedom of α only makes such gains more common. The same tail beyond the whole sample's 37.2 is 10^{-9}.
- *Where is it?* How far the solid curve rises away from its minimum. Near the minimum $\Delta\chi^2 < 1$ for $0.936 < \alpha < 1.038$, a local error of ± 0.051. Farther out the curve flattens into a plateau, because a template stretched far from the data's bump puts its own bump where there is none, and the smooth terms then fit almost as well as without a bump.

The plateau decides the posterior's width. At $\alpha = 0.8$ the curve is only 2.3 above its minimum, so $p(\alpha)$ there is still 0.32 of its peak value, and the standard deviation over the grid, $\sigma_\alpha = 0.101$, is set as much by the edges of the flat prior as by the data. A wider prior would give a wider error. This is what a weak detection looks like, and the right summary of our measurement is two numbers: $\alpha = 0.997 \pm 0.101$ over the prior $0.8 < \alpha < 1.2$, or ± 0.051 if one accepts that the minimum at 0.992 is the acoustic peak. Through (13.39),

$$\frac{D_V(0.565)}{r_d} = 13.88 \pm 1.41 \quad \text{(local: } \pm 0.71\text{)}, \qquad \text{Planck: } 13.92. \tag{13.40}$$

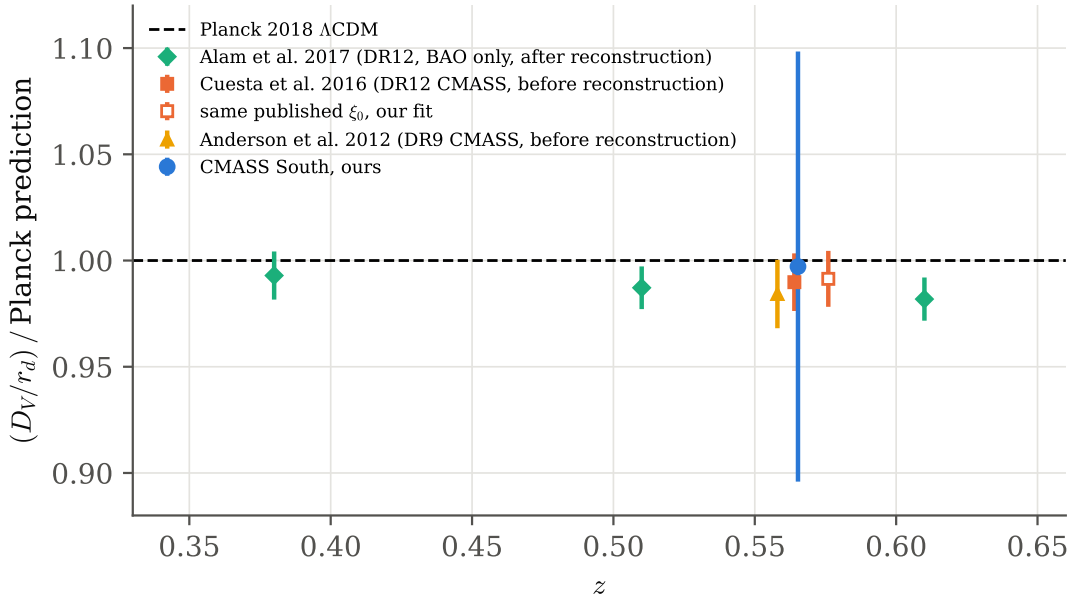


Figure 13.37: D_V/r_d divided by the Planck 2018 prediction at each redshift. BOSS results (Anderson et al. converted to CAMB's r_d; Alam et al.'s D_M and H combined into D_V), our refit of the published monopole (open square), and our southern measurement (blue, with the posterior error over $0.8 < \alpha < 1.2$). Points at $z \approx 0.57$ are spread slightly in z to be visible.

Figure 13.37 and Table 13.10 collect the numbers. For Alam et al.'s final distances we form D_V/r_d from their $D_M\, r_d^{\text{fid}}/r_d$ and $H\, r_d/r_d^{\text{fid}}$ with (11.27) and the delta-method error (11.34), using their covariance between the two; they include reconstruction and both directions, so their errors are smaller, and their bins at $z = 0.51$ and 0.61 bracket ours.

analysis	sample	z	D_V/r_d	Planck
Anderson et al. 2012	DR9 CMASS	0.57	13.79 ± 0.23	14.01
Cuesta et al. 2016	DR12 CMASS, N+S	0.57	13.87 ± 0.19	14.01
their $\hat{\xi}_0$, our fit	DR12 CMASS, N+S	0.57	13.89 ± 0.18	14.01
ours	DR12 CMASS, S	0.565	13.88 ± 1.41	13.92
Alam et al. 2017	DR12, all	0.51	12.67 ± 0.13	12.83
Alam et al. 2017	DR12, all	0.61	14.50 ± 0.15	14.76

Table 13.10: Paper against ours. All values use CAMB's r_d (Anderson et al.'s 13.44 ± 0.22 converted as in Section 13.19.3); the first four are before reconstruction and from the monopole, Alam et al.'s after reconstruction and from both directions. The last column is the Planck 2018 prediction at that z.

13.19.6 Why they differ

Three numbers need explaining: why our error is so much larger than the published one, why our detection is weaker than expected, and whether our α agrees with theirs. The distances themselves need little explaining: they agree with each other and with the Planck prediction to within two standard deviations, the BOSS values sitting one to two per cent below the prediction.

The area, and the jackknife. The southern cap is 27% of the area, so by the mode counting of Section 13.18.3 its errors on $\hat{\xi}_0$ should be 1.93 times the published ones; with the jackknife's slight excess they are 2.09 times. The significance $\sqrt{\Delta\chi^2}$ is the bump's amplitude divided by its standard error (Eq. (11.31)), so for a southern cap with an average bump it would be $6.1/2.09 = 2.9\sigma$. We found 1.6σ.

A weaker bump than average. Is 1.6 against an expected 2.9 a surprise? The estimated amplitude $\hat{A}$ of the bump scatters from survey to survey with standard deviation σ_A, so $\sqrt{\Delta\chi^2} \approx \hat{A}/\sigma_A$ scatters by about one unit. Our cap is 1.3 units below the expectation; a Gaussian falls that low or lower in about 10% of cases. The mock surveys of Section 11.4.4 showed the same spread from the other side, where the 2005 detection sat in the upper tail. Figure 13.35 shows where the deficit is: our points between 105 and 120 h^{-1}Mpc fall below the published ones, so the south's bump is narrower and lower than the whole sample's.

Why a weak bump is also badly located. The width of the χ^2 minimum is tied to the strength of the detection. Near the best fit, write the model as the smooth part plus the bump, $B^2\xi_t = B^2(\xi_{\rm nw} + \Delta)$, where $\Delta(s)$ is the acoustic feature, and suppose the smooth terms absorb the change of $\xi_{\rm nw}$ with α, so that only the bump's motion is informative. Then the two curvatures of χ^2 are

$$\begin{aligned}
\frac{1}{\sigma_\alpha^2} &\overset{(1)}{=} \Big(B^2\,\partial_\alpha\Delta\Big)^{\mathsf T}\mathbf{C}^{-1}\Big(B^2\,\partial_\alpha\Delta\Big) \overset{(2)}{=} B^4\,\big(s\Delta'\big)^{\mathsf T}\mathbf{C}^{-1}\big(s\Delta'\big),\\
\Delta\chi^2 &\overset{(3)}{=} \frac{1}{\sigma_A^2} \overset{(4)}{=} B^4\,\Delta^{\mathsf T}\mathbf{C}^{-1}\Delta,\\
\sigma_\alpha\sqrt{\Delta\chi^2} &\overset{(5)}{=} \left[\frac{\Delta^{\mathsf T}\mathbf{C}^{-1}\Delta}{(s\Delta')^{\mathsf T}\mathbf{C}^{-1}(s\Delta')}\right]^{1/2}.
\end{aligned} \tag{13.41}$$

(1) For a model linear in a parameter near its best fit, $1/\sigma^2$ is the derivative vector of the model sandwiched by $\mathbf{C}^{-1}$ (the Fisher matrix of a Gaussian mean, Section 10.3).
(2) $\partial_\alpha\Delta(\alpha s) = s\,\Delta'(\alpha s)$, evaluated at $\alpha = 1$.
(3) (11.31) with the data preferring the full amplitude.
(4) The bump's amplitude A multiplies $B^2\Delta$, so its Fisher information is that vector sandwiched by $\mathbf{C}^{-1}$.
(5) Divide the second line by the first, so that B^4 cancels, and take the square root.

The right-hand side depends only on the shape of the bump and of the covariance, not on how strong the signal is. A bump detected half as strongly is located half as precisely. The check: $\sigma_\alpha\sqrt{\Delta\chi^2}$ is 0.083 for our cap (local width) and 0.073 for the whole sample, equal within the roughness of the approximation although the covariances come from different methods. So the large error on our α is not a separate failure; it is the weak detection seen again. Had our bump been average, the local error would have been about 0.025.

Our α against theirs. Are 0.992 (ours, south) and 0.978 (the whole sample, same template) consistent? The two are not independent, since the south is inside the whole sample, so the usual rule of adding variances does not apply. Treat the whole-sample estimate as the inverse-variance average of a southern estimate α_S and a northern one α_N, independent, with variances σ_S^2 and σ_N^2, and ask for the variance of the difference $\alpha_S - \alpha_{NS}$.

$$\alpha_{NS} = \sigma_{NS}^2\Big(\frac{\alpha_S}{\sigma_S^2} + \frac{\alpha_N}{\sigma_N^2}\Big), \qquad \frac{1}{\sigma_{NS}^2} = \frac{1}{\sigma_S^2} + \frac{1}{\sigma_N^2},$$

$$\alpha_S - \alpha_{NS} \overset{(1)}{=} \sigma_{NS}^2\Big(\frac{1}{\sigma_S^2} + \frac{1}{\sigma_N^2}\Big)\alpha_S - \sigma_{NS}^2\Big(\frac{\alpha_S}{\sigma_S^2} + \frac{\alpha_N}{\sigma_N^2}\Big) = \frac{\sigma_{NS}^2}{\sigma_N^2}\,(\alpha_S - \alpha_N),$$

$$\mathrm{Var}(\alpha_S - \alpha_{NS}) \overset{(2)}{=} \frac{\sigma_{NS}^4}{\sigma_N^4}\big(\sigma_S^2 + \sigma_N^2\big) \overset{(3)}{=} \frac{\sigma_S^4}{\sigma_S^2 + \sigma_N^2} \overset{(4)}{=} \sigma_S^2 - \sigma_{NS}^2. \tag{13.42}$$

(1) Multiply α_S by $1 = \sigma_{NS}^2(1/\sigma_S^2 + 1/\sigma_N^2)$ and subtract; the α_S/σ_S^2 terms cancel.
(2) α_S and α_N are independent, so their variances add.
(3) $\sigma_{NS}^2 = \sigma_S^2\sigma_N^2/(\sigma_S^2 + \sigma_N^2)$, so $\sigma_{NS}^4/\sigma_N^4 = \sigma_S^4/(\sigma_S^2 + \sigma_N^2)^2$.
(4) $\sigma_S^2 - \sigma_{NS}^2 = \sigma_S^2\big[1 - \sigma_N^2/(\sigma_S^2 + \sigma_N^2)\big] = \sigma_S^4/(\sigma_S^2 + \sigma_N^2)$.

A subset's estimate differs from the whole sample's by noise of variance $\sigma_S^2 - \sigma_{NS}^2$, smaller than either naive sum. With the local widths, $\sqrt{0.051^2 - 0.012^2} = 0.050$, and the difference 0.014 is a 0.3σ deviation. The north and south caps really were analysed separately by BOSS and agreed within 2σ in DR9 [277].

Smaller differences. Our randoms are 5 times the galaxies, not 50 (Section 13.17.3); the shot noise of RR and DR then adds a few per cent to the variance at the acoustic scale. All analyses in Table 13.10 except Alam et al.'s are before reconstruction; reconstruction sharpens the bump and shrinks the error by about a quarter (from 0.19 to 0.14 for the whole sample, [Cuesta2016 §4.4]). The DR9 sample of Anderson et al. is a third of DR12 and partly the same galaxies, so its agreement with DR12 is again a subset comparison.

Lesson. Reproducing a BAO distance taught two things that reading the papers did not. A fitted α is meaningless without its coordinates and its template, and two analyses that disagree by four per cent in α can agree to a fraction of a per cent in D_V/r_d; equation (13.39) is what makes them comparable. And the error on a peak's position does not follow the area alone. It scales with the inverse of the detection significance, and that scatters by about one unit from sky to sky. The southern cap has 27% of the area, which makes the errors on ξ about 1.93 times larger, yet its local error on α is 0.051 against 0.012 for the whole sample, and its posterior over the prior range is wider still. A weak detection is also a poor location.

13.20 The survey window, and what the full analysis adds

The galaxy catalogue and the Planck map share one difficulty. Neither covers the whole sky evenly: the CMB is hidden behind the Galaxy and point sources, and the galaxies were observed only inside the BOSS footprint, with a density that falls with distance. For the CMB that difficulty took a whole machinery, the mode-coupling matrix of Section 8.10 and its inversion in MASTER (Section 8.11, [168]). For the galaxies we never built a coupling matrix, and yet Figure 13.35 agrees with the published measurement. Where did the window go?

13.20.1 The random catalogue is the mask

The question. A survey multiplies the true galaxy density by a window $W(\boldsymbol{x})$, the expected number of weighted galaxies per unit volume at $\boldsymbol{x}$ when there is no clustering. In our analysis $W = \bar{n}\, w_{\mathrm{FKP}}$: zero outside the footprint and the redshift slice, and shaped inside by $\bar{n}(z)$ and the FKP weight of (13.30). What does W do to the expected pair counts?

Count pairs whose separation falls in a thin shell of radius s and width δs. Write $\Theta_s(\boldsymbol{r})$ for the indicator of that shell, 1 if $s < |\boldsymbol{r}| < s + \delta s$ and 0 otherwise. The weighted galaxy density has mean $W(\boldsymbol{x})$ and pair correlation ξ (Section 11.1.2), and the randoms sample W with no correlation.

$$\begin{aligned}
\langle DD(s)\rangle &\overset{(1)}{=} \int\!\!\int W(\boldsymbol{x})\, W(\boldsymbol{y}) \left[1 + \xi(|\boldsymbol{x}-\boldsymbol{y}|)\right] \Theta_s(\boldsymbol{x}-\boldsymbol{y})\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}\boldsymbol{y} \\
&\overset{(2)}{=} \left[1 + \xi(s)\right] \int\!\!\int W(\boldsymbol{x})\, W(\boldsymbol{y})\, \Theta_s(\boldsymbol{x}-\boldsymbol{y})\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}\boldsymbol{y}, \\
\langle RR(s)\rangle &\overset{(3)}{=} \int\!\!\int W(\boldsymbol{x})\, W(\boldsymbol{y})\, \Theta_s(\boldsymbol{x}-\boldsymbol{y})\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}\boldsymbol{y}, \qquad \text{so} \qquad \frac{\langle DD(s)\rangle}{\langle RR(s)\rangle} \overset{(4)}{=} 1 + \xi(s).
\end{aligned} \tag{13.43}$$

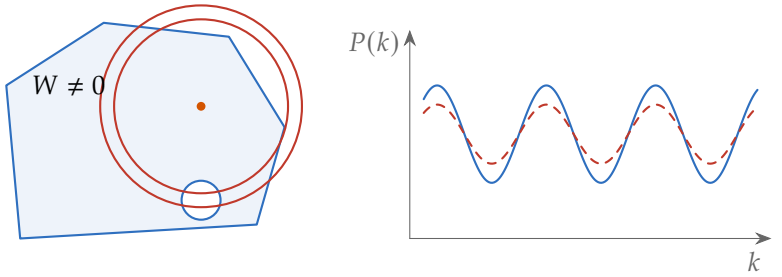
(1) The expected number of pairs in two small volumes is the product of the mean densities times $1 + \xi$ (Eq. (11.5)), summed over all pairs of volumes whose separation is in the shell.
(2) Inside a thin shell $|\boldsymbol{x} - \boldsymbol{y}| \approx s$, so ξ is the same for every pair and comes out of the integral.
(3) The randoms have the same mean density shape W and no clustering.
(4) The double integral is the same in both; it cancels. (The normalisations of (13.31) make the total pair numbers agree.)

The double integral is the window's own pair count: how many pairs at separation s the footprint can hold at all. It is the configuration-space twin of the mask's power spectrum $\mathcal{W}_\ell$ that builds the coupling matrix in [Hivon (14), (A31)]. Because the correlation function depends only on the separation, and a thin shell fixes the separation, the window multiplies each bin by its own number and mixes nothing: dividing by RR undoes it bin by bin. The randoms are the mask, and RR is its coupling "matrix", which here is diagonal. The Landy–Szalay combination (13.33) keeps this property and also cancels the leading edge terms in the variance (Section 11.2.2).

Why the power spectrum is different. The FKP estimator of Section 11.3 Fourier transforms the weighted density on a grid. Multiplication by W in configuration space is a convolution in Fourier space, so the measured power at $\boldsymbol{k}$ is the true power averaged over neighbouring wavevectors with weight $|\tilde{W}(\boldsymbol{k} - \boldsymbol{k}')|^2$. Every k-bin mixes with its neighbours, exactly as the mask mixes multipoles on the sphere, and a sharp feature such as the acoustic wiggles is smoothed (Figure 13.38). A $P(k)$ analysis must therefore convolve the model with the window, the analogue of the forward-modelling route of MASTER, before comparing it with data.

Example: how much a window smooths a wiggle. Take a wiggle $1 + A\sin(kr_d)$ and a window whose $|\tilde{W}|^2$, normalised to unit integral, is a Gaussian of width σ_k in k. The convolved wiggle is the average of $\sin\big((k-q)r_d\big)$ over $q \sim \mathcal{N}(0, \sigma_k^2)$. Write the sine as the imaginary part of $e^{i(k-q)r_d}$; the average of e^{-iqr_d} over a Gaussian is $e^{-\sigma_k^2 r_d^2/2}$ (complete the square in the exponent), a real number, so the average is $\sin(kr_d)\, e^{-\sigma_k^2 r_d^2/2}$ and the wiggle survives with amplitude $A\, e^{-\sigma_k^2 r_d^2/2}$. A survey $L \approx 1\, h^{-1}$Gpc deep has $\sigma_k \sim 1/L$, and with $r_d \approx 100\, h^{-1}$Mpc the factor is $e^{-0.005}$: a large survey barely blurs the BAO. A thin slice $100\, h^{-1}$Mpc deep in one direction has $\sigma_k r_d \sim 1$ along it and keeps only $e^{-1/2} \approx 60\%$ of the wiggle in that direction. Pair counts would see the same peak at full height.

Where the cancellation fails. Equation (13.43) assumed that the randoms trace the true window. They do not if the mean density is itself estimated from the same galaxies: the randoms' redshifts are drawn from the galaxies (Section 13.15.2), and the total number is normalised to the observed number. Both remove the large-scale fluctuations that the survey's own volume contains, the integral constraint of Section 11.2.3, which pulls $\hat{\xi}$ down by a nearly constant amount; the term a_0 in the fit (11.30) absorbs it. Ross et al. compared ways of giving the randoms their redshifts on mock catalogues and found that drawing them from the galaxies ("shuffling") biased the clustering least [Ross2012 §6.1].



shell of radius s in the footprint true (blue), window-convolved (red)

Figure 13.38: The window in two spaces. Left: in configuration space, pairs at one separation are counted in a shell; the footprint (blue, with a masked hole) cuts the shell, but it cuts the galaxy pairs and the random pairs alike, so their ratio is untouched. Right: in Fourier space the window convolves the spectrum; for a Gaussian window of width σ_k a wiggle of period $2\pi/r_d$ keeps the fraction $e^{-\sigma_k^2 r_d^2/2}$ of its amplitude (drawn for $\sigma_k r_d = 1$).

13.20.2 The full analysis, and what our shortcuts cost

Each step of our pipeline has a research-grade version. Table 13.11 lists them with the price of the shortcut; the largest prices are the area, which costs a factor of about two in every error, and the absence of mocks, which makes the error bars themselves uncertain by some tens of per cent.

step	ours	BOSS DR12, and the cost of our choice
sky	southern cap, 27% of the area	both caps [278]; errors 1.93 times larger, detection 1.6σ instead of 6.1σ
randoms	5 per galaxy in the counts	50 per galaxy [276]; a few per cent more variance
covariance	120 jackknife regions	about a thousand quick-particle-mesh (QPM) mocks (956 for CMASS) and a second set as a check; our errors uncertain by tens of per cent, Hartlap factor only a guide
statistic	monopole only	monopole and quadrupole; we measure D_V only, not D_M and H separately
reconstruction	none	displacements estimated from the density field and undone; our errors about 25% larger
template	fixed $\Sigma_{\rm nl} = 8\,h^{-1}$Mpc	damping calibrated on mocks, systematic error budget for template and fiducial cosmology
blinding	none	catalogue-level blinding, now standard [DESI2024VI §2.3.1]

Table 13.11: Our analysis of CMASS South against the published BOSS analysis.

All of it ran at laptop scale: the fully weighted pair counts took 338 s on 8 cores, and the fits

a few seconds. The full DR12 random catalogues of both caps, at 50 times the galaxies, would need more than 4 GB of downloads and several hundred times our pair-counting time, because the number of random pairs within 184 h^{-1}Mpc grows as the number of randoms times their density.

Two further summaries of the same catalogue follow from earlier chapters, and our analysis did not compute them. The power-spectrum monopole with FKP weights (Section 11.3) measures the same wiggles in Fourier space, and Section 13.20.1 says what it needs: the model convolved with the window of our footprint. The Betti curves of the smoothed galaxy density (Section 12.13.2) would test the morphology of the cosmic web, but their null distribution, like that of the CMB sky in Section 13.9, must come from mocks that pass through the same window: a mask changes the topology of every excursion set it cuts.

How this is used in real research: verification and research in a BAO analysis

Two of our checks were verifications: refitting the published monopole (Section 13.19.5) tested our fitting code against a known answer, and comparing jackknife errors with mock errors (Section 13.18.3) tested an error method against a better one. In the research analysis the same Monte Carlo machinery is the instrument. Thousands of mock surveys, each with the real footprint, density and fibre collisions, give the covariance; the same mocks measure whether reconstruction or the template shifts α, which sets the systematic error; and the fitted α of each mock gives the sampling distribution of $\hat{\alpha}$, against which non-Gaussian error behaviour like that of our Figure 13.36 is calibrated [224, 278]. DESI adds blinding: the catalogue is shifted in a hidden way until the analysis choices are frozen, so that agreement with a favoured model cannot steer them [DESI2024VI §2.3.1].

Recap

- A galaxy catalogue becomes a set of positions through a fiducial distance–redshift relation; the error made by choosing the wrong one is exactly what the dilation α measures.
- Weights restore the galaxies the survey could not observe properly (close pairs, redshift failures, stars and seeing) and the FKP weight balances signal and shot noise; the randoms carry only the FKP weight because they already trace the survey's completeness.
- The Landy–Szalay estimator with weighted counts divides out the survey window bin by bin, because in a thin shell the correlation function is constant; in Fourier space the same window convolves the spectrum and must be modelled.
- Counting pairs region by region gives every leave-one-region-out count by inclusion–exclusion, and with them the jackknife covariance $\frac{N-1}{N}\sum_k(\hat{\boldsymbol{\xi}}_{-k}-\bar{\boldsymbol{\xi}})(\hat{\boldsymbol{\xi}}_{-k}-\bar{\boldsymbol{\xi}})^{\mathsf{T}}$, exact for a mean and biased upward otherwise; on CMASS South it agrees with scaled mock errors to 1.09.
- A fitted α becomes a distance through $D_V/r_d = \alpha\, D_V^{\mathrm{fid}}/r_d^{\mathrm{tmpl}}$; the ruler's name and the coordinates must be identified before numbers are compared.
- CMASS South shows the acoustic peak at 1.6σ and gives $D_V/r_d = 13.88 \pm 1.41$, consistent with BOSS and Planck; the whole sample, refitted with our code, gives 13.89 ± 0.18 at 6.1σ, as published.
- A weak detection is also a poor location, $\sigma_\alpha\sqrt{\Delta\chi^2}$ being fixed by the shape of the bump; and a subset differs from its parent sample by noise of variance $\sigma_S^2 - \sigma_{NS}^2$.

13.21 A Higgs search on collider open data

On the fourth of July 2012 two collaborations at CERN showed, one after the other, a histogram of the invariant mass of photon pairs. Each had a smooth background falling from left to right and, near 125 GeV, a small excess that a smooth curve could not explain. Together with the four-lepton channel, those excesses were the discovery of the Higgs boson [118, 124]. In 2020 the ATLAS collaboration released part of its 2016 data at a collision energy of 13 TeV to the public, with the simulated Higgs-boson events that go with it [121, 123]. So the same search can be done from the recorded collisions to the final significance, with every step under our control. The physics of what is recorded, the photons and their momenta, the cross section and the luminosity, is set out in Chapter T4; what follows is the statistics.

What we have already seen, and what is new. The search is not new to us. In Section 4.14.3 we hunted for a bump in a simulated diphoton-like spectrum of 3000 background events with a known background shape and a Gaussian peak of width 1.5 GeV. The largest excess sat at 142.50 GeV with a local significance of 3.28σ; background-only toys turned that into a global 2.15σ, a trials factor of about 31.3, and the Gross–Vitells formula of Section 4.14.4 reproduced it from the upcrossings of a hundred toys. In Sections 4.15 and 4.16 the same kind of pseudo-spectrum, now generated with a signal, became a likelihood with nuisance parameters: q_0 for discovery, $\tilde{q}_\mu$ and CL_s for exclusion, the Asimov data set for the expected results and the Brazil band for the expected limit. The statistics is the same now. Four things are different, and each one is a real-data problem that the pseudo-experiments allowed us to ignore.

1. *Nobody knows the background shape.* In Section 4.14.3 it was the exponential we had used to generate the data. Here it is whatever the mixture of photon pairs, photon–jet and jet–jet events produces after our selection, and it must be described by a smooth function chosen from the data. A wrong choice can manufacture a bump or hide one.

2. *The signal shape comes from simulation,* and it is not a Gaussian: some photons lose energy before reaching the calorimeter, which leaves a tail on the low side of the peak.

3. *The expected signal is uncertain,* through quantities measured elsewhere (efficiencies, luminosity, the energy scale and resolution of the photons). They enter as nuisance parameters with constraint terms, as in Section 4.15.2.

4. *The sample is big.* About 76394 events lie in the window instead of 3000, so the background is measured very precisely and small imperfections of the background function matter as much as the statistical fluctuations.

The questions in words.

1. Which recorded collisions contain two well-measured photons, and what does the spectrum of their invariant mass look like?

2. How large and how wide a peak would a Standard Model Higgs boson make in that spectrum, and which smooth function can describe everything else without inventing or hiding a peak?

3. If the background alone were present, would an excess like the one in the data, or an even larger one, be something we would reasonably expect, at one mass and anywhere in the range searched?

4. How strong is the signal compared with the Standard Model prediction, and where is it?

5. At which masses can a Standard Model Higgs boson be excluded, and how does the whole result compare with what ATLAS reported in 2012?

13.21.1 From collisions to a list of masses

The files. The release contains, for the diphoton search, four files of recorded collisions (the 2016 data-taking periods A to D, together $10.06\,\text{fb}^{-1}$) and four files of simulated Higgs bosons with $m_H = 125\,\text{GeV}$ decaying to two photons, one for each way of producing it: gluon fusion (ggH), vector-boson fusion (VBF) and production together with a W or Z boson (WH, ZH). Together they are 1.9 GB. They were already reduced by ATLAS to events in which the diphoton trigger fired and two photon candidates were found [121]. Each file is a ROOT tree with one entry per collision; the script `code/ch13/31_hyy_select.py` reads the columns it needs with `uproot`. For every photon candidate the tree stores its transverse momentum p_T [the momentum component perpendicular to the beams], its pseudorapidity η [a measure of the polar angle, $\eta = -\ln\tan(\theta/2)$, zero at right angles to the beams], its azimuth ϕ around the beams, and flags describing how cleanly it was measured (Chapter T4).

Why the invariant mass. *The question:* if two photons come from the decay of one particle of mass M, what combination of their measured momenta tells us M? Energy and momentum are conserved in the decay, so the four-momentum of the parent is the sum of the two photon four-momenta, $P = p_1 + p_2$. The square of a four-momentum is the same in every frame, and in the rest frame of the parent it is M^2 (with $c = 1$). So $M^2 = (p_1 + p_2)^2$ whatever the motion of the parent. A photon is massless, so its four-momentum is fixed by (p_T, η, ϕ):

$$p = (E,\ p_T\cos\phi,\ p_T\sin\phi,\ p_T\sinh\eta), \qquad E = p_T\cosh\eta,$$

because $p_z = p_T\sinh\eta$ follows from the definition of η, and then

$$E = \sqrt{p_T^2 + p_z^2} = p_T\sqrt{1+\sinh^2\eta} = p_T\cosh\eta.$$

With the metric $p\cdot q = E_pE_q - \vec{p}\cdot\vec{q}$,

$$\begin{aligned} m_{\gamma\gamma}^2 &\overset{(1)}{=} p_1^2 + p_2^2 + 2\,p_1\cdot p_2 \overset{(2)}{=} 2\big(E_1E_2 - \vec{p}_1\cdot\vec{p}_2\big) \\ &\overset{(3)}{=} 2p_{T1}p_{T2}\big[\cosh\eta_1\cosh\eta_2 - \sinh\eta_1\sinh\eta_2 - \cos\phi_1\cos\phi_2 - \sin\phi_1\sin\phi_2\big] \\ &\overset{(4)}{=} 2p_{T1}p_{T2}\big[\cosh(\eta_1-\eta_2) - \cos(\phi_1-\phi_2)\big]. \end{aligned} \tag{13.44}$$

(1) Expand the square of the sum.
(2) A massless particle has $p^2 = E^2 - |\vec{p}|^2 = 0$.
(3) Insert the components above.
(4) The addition formulas $\cosh(a-b) = \cosh a\cosh b - \sinh a\sinh b$ and $\cos(a-b) = \cos a\cos b + \sin a\sin b$.

Only differences of η and ϕ appear: the mass does not care how fast the pair moves along the beams or which way it points around them. That is why the invariant mass, and not the energy of either photon, is where a resonance shows up as a peak; the same argument, for any pair of particles, is Section T4.3.1.

Example: one photon pair. Take a pair with $p_{T1} = 60\,\text{GeV}$, $p_{T2} = 52\,\text{GeV}$, $\Delta\eta = 1.0$ and $\Delta\phi = 2.9\,\text{rad}$. Then $\cosh 1.0 = 1.543$ and $\cos 2.9 = -0.971$, so

$$m_{\gamma\gamma}^2 = 2 \times 60 \times 52 \times (1.543 + 0.971) = 6240 \times 2.514 = 15\,688\,\text{GeV}^2, \qquad m_{\gamma\gamma} = 125.3\,\text{GeV}.$$

The photons are nearly back to back in azimuth ($\Delta\phi$ close to π), as the two decay products of a slowly moving particle must be. Nothing in this pair says whether it came from a Higgs boson: a pair of photons from ordinary quark–antiquark annihilation can have exactly these momenta. Single events are never signal or background; only the distribution of many events can be.

The selection. Most recorded events with two photon candidates contain no real photon pair at all: a jet of hadrons can contain a neutral pion whose two photons merge into one deposit. Each cut below removes a specific kind of fake or badly measured event, and each is copied from the `HyyAnalysis` example of the release.

- *Trigger*: the diphoton trigger fired.
- *Exactly two good photons*: each passes the tight identification [the shape of its shower in the calorimeter looks like one photon, not several], has $p_T > 25\,\text{GeV}$ and $|\eta| < 2.37$, outside $1.37 < |\eta| < 1.52$ where the barrel and end-cap calorimeters meet and the energy is poorly measured.
- *Isolation*: the momentum of tracks within $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2} < 0.3$ and the calorimeter energy within $\Delta R < 0.2$ around each photon are both below 6.5% of its p_T. A photon from a Higgs decay is alone; a fake photon inside a jet is surrounded by hadrons.
- *Relative momenta*: $p_{T1}/m_{\gamma\gamma} > 0.35$ and $p_{T2}/m_{\gamma\gamma} > 0.25$ for the leading and subleading photon.
- *Mass window*: $105 < m_{\gamma\gamma} < 160\,\text{GeV}$.

Why relative cuts on p_T? An absolute cut, say $p_T > 40\,\text{GeV}$ on both photons, removes almost every pair below $m_{\gamma\gamma} \approx 80\,\text{GeV}$ and fewer and fewer pairs above it, so it bends the background spectrum near the window. A cut on $p_T/m_{\gamma\gamma}$ is a cut on the decay angle, the same at every mass, so it changes the background's size but not its smoothness. A smooth background is what the fit of Section 13.22 assumes.

What survives. Of 7798424 triggered events, 979404 have exactly two good photons, 362972 of these are isolated, 322243 pass the relative momentum cuts and 76394 fall in the mass window (Figure 13.39). The four periods contribute 4227, 14804, 21876 and 35487 events, in proportion to their luminosity.

How many Higgs bosons should be among them? *The question:* the simulation produced about a million ggH events, far more than the data contain; how does one turn a selected simulated event into a number of expected real events? If a process has cross section σ [the effective area of the collision for that process, here including the probability that the Higgs boson decays to two photons] and the data correspond to an integrated luminosity L [the number of collisions per unit area that were recorded], the expected number of such events produced is $L\sigma$ (Chapter T4). The expected number selected is that times the probability ε of passing the selection, which the simulation estimates. The simulation's events carry generator weights w_j (some generators produce weighted, even negatively weighted, events), and a selected event also carries scale factors SF_j that correct the simulated efficiencies of the trigger,

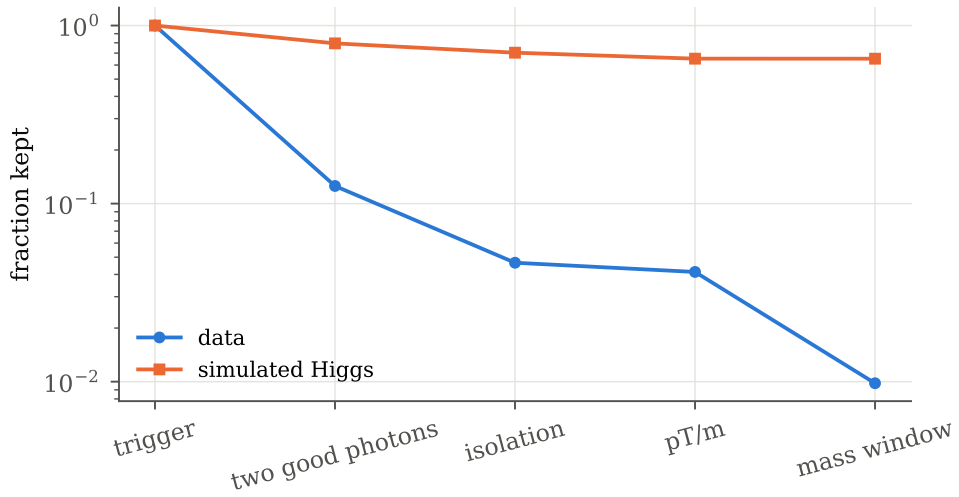


Figure 13.39: The fraction of events kept after each step of the selection, for the recorded data (blue) and for the simulated Higgs bosons (orange), both relative to the events that passed the trigger. The photon identification and the isolation remove most of the data and little of the signal: that is the purpose of each cut. The mass window removes most of the remaining data, whose spectrum extends far beyond 160 GeV.

the photon identification and the pile-up to those measured in data. As in Section T4.4.2,

$$S \overset{(1)}{=} L\,\sigma\,\varepsilon \overset{(2)}{=} L\,\sigma\,\frac{\sum_{j\in\text{selected}} w_j\,\mathrm{SF}_j}{\sum_{\text{all } j} w_j} \overset{(3)}{=} \sum_{j\in\text{selected}} \underbrace{\frac{L\,\sigma}{\sum_{\text{all } k} w_k}\,w_j\,\mathrm{SF}_j}_{\text{weight of event } j}\,. \tag{13.45}$$

(1) Expected selected events: produced events times the selection probability.
(2) The weighted fraction of simulated events that pass; with all weights one it is the plain fraction $N_{\text{sel}}/N_{\text{gen}}$. The sum in the denominator runs over all generated events, before any selection, and is stored in the files.
(3) Bring the constant inside the sum: every selected event contributes its own weight.

With $L = 10.06\,\text{fb}^{-1}$ the expected Higgs signal after all cuts is $S = 397.9$ events: 357.2 from gluon fusion, 29.9 from vector-boson fusion, 5.6 and 5.3 from production with a W or Z. A fraction 0.651 of the signal events in the files survives the selection (Figure 13.39), against 76394/7798424 of the data.

The spectrum. Figure 13.40 shows the 76394 masses in 1 GeV bins and, below them, the 397.9 expected Higgs events on their own. The data fall smoothly from about 3000 to about 600 events per GeV. The Higgs peak is 60 to 70 events per GeV high at its top, a few per cent of the background there, and no bump is visible by eye: the error bar of a single bin, $\sqrt{1600} = 40$ events, is not much smaller than the whole peak. Whether the peak is there has to be decided by the likelihood, using all the bins at once.

The full analysis, and ours. ATLAS in 2012 split its events into ten categories of different mass resolution and signal-to-background ratio, fitted each with its own background function, chosen with large simulated background samples, and combined two collision energies [118]; the later analyses of the same detector use many times more data and more categories. We keep the published selection and one inclusive spectrum, read 1.9 GB of files once, and use two

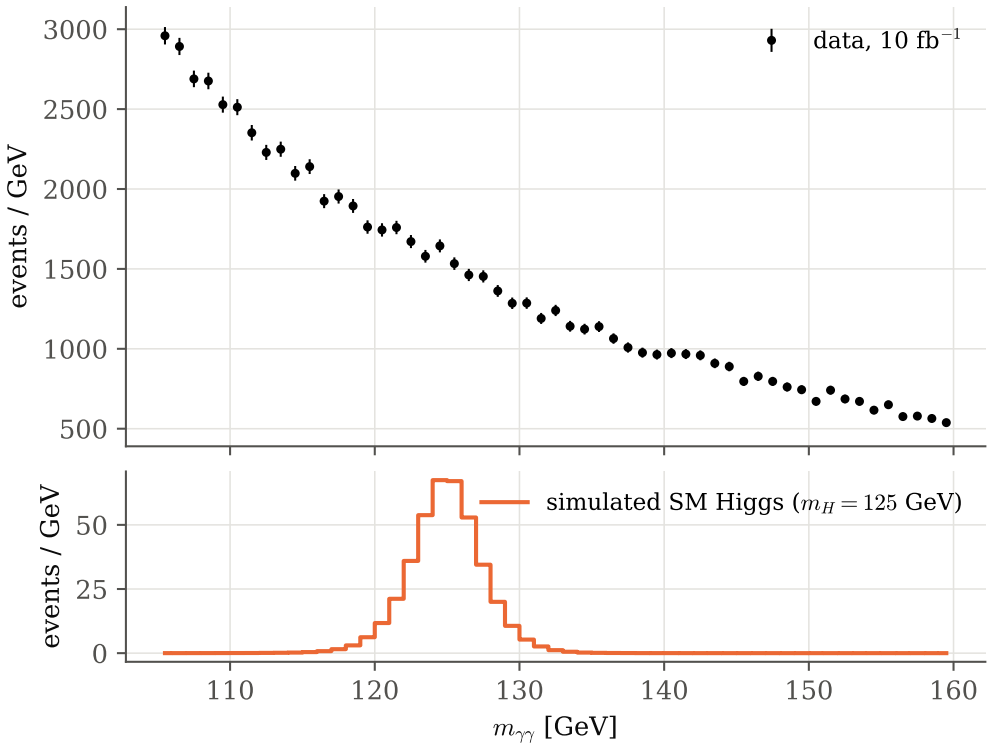


Figure 13.40: Top: the invariant mass of the 76394 selected photon pairs in 1 GeV bins, with Poisson error bars. Bottom: the Higgs signal expected in the same data from the simulation, 397.9 events in all, drawn on its own because on the scale of the top panel it would be a ripple of a few per cent.

categories only as an illustration (Section 13.25.3). The price is sensitivity, which Section 13.27 quantifies.

13.22 Shapes: the Higgs line and the background function

The likelihood of Section 4.15.1 needs two templates: the fraction f_i of the signal and the fraction g_i of the background expected in each bin i of the mass histogram, so that the mean count is $\nu_i = \mu S f_i + B g_i$ (4.88). The pseudo-experiments of Chapter 4 supplied both templates. For the data analysis we estimate the signal shape from the 397.9 simulated Higgs events of Section 13.21.1 and fit a smooth background model to the observations. Each shape can affect the inferred signal strength, so we check the two modelling choices separately.

13.22.1 The signal line shape

A Higgs boson is so narrow (its natural width is about 4 MeV, Chapter T4) that the width of the peak in $m_{\gamma\gamma}$ is entirely the detector's: the calorimeter measures each photon's energy with an error of one or two per cent, and by (13.44) $m_{\gamma\gamma} \propto \sqrt{E_1 E_2}$ inherits about half of each relative error. Many small independent errors add up to a Gaussian (the central limit theorem of Chapter 2), so the core of the peak is Gaussian. But some photons lose part of their energy before they reach the calorimeter, by converting into an electron–positron pair in the material in front of it or by leaking out of the back. Such losses only ever lower the measured mass. The peak therefore has a tail on its low side, and a Gaussian, which is symmetric, misses it.

The Crystal Ball function. The standard description, met in Section T4.3.3, is a Gaussian core glued to a power-law tail. With $t = (m - \bar{m})/\sigma$ the distance from the peak in units of the core width,

$$f_{\mathrm{CB}}(m) \propto \begin{cases} e^{-t^2/2}, & t > -\alpha, \\ A\,(B - t)^{-n}, & t \le -\alpha, \end{cases} \tag{13.46}$$

with four shape parameters: the peak position $\bar{m}$, the core width σ, the point $t = -\alpha$ where the tail takes over, and the power n of the tail. The constants A and B are not free. *The question:* what must they be for the two pieces to join smoothly at $t = -\alpha$, with the same value and the same slope? Equal slopes first, in logarithmic form (the log of each piece has a simple derivative):

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big(-\frac{t^2}{2}\Big)\Big|_{t=-\alpha} \overset{(1)}{=} \frac{\mathrm{d}}{\mathrm{d}t}\big(\ln A - n\ln(B-t)\big)\Big|_{t=-\alpha} \implies \alpha = \frac{n}{B+\alpha} \implies B \overset{(2)}{=} \frac{n}{\alpha} - \alpha,$$
$$e^{-\alpha^2/2} \overset{(3)}{=} A\,(B+\alpha)^{-n} \implies A \overset{(4)}{=} \Big(\frac{n}{\alpha}\Big)^n e^{-\alpha^2/2}. \tag{13.47}$$

(1) Equal values and equal slopes means equal logarithmic slopes. On the left the derivative is $-t = \alpha$; on the right it is $n/(B - t) = n/(B + \alpha)$.
(2) Solve for B.
(3) Equal values at $t = -\alpha$.
(4) $B + \alpha = n/\alpha$ by the previous line.

For the normalisation we need the area of each piece, in units of σ. The core is a piece of a Gaussian, $\int_{-\alpha}^{\infty} e^{-t^2/2}\mathrm{d}t = \sqrt{\pi/2}\,[1 + \mathrm{erf}(\alpha/\sqrt{2})]$. The tail is a power law, which integrates in one line:

$$\int_{-\infty}^{-\alpha} A(B-t)^{-n}\mathrm{d}t \overset{(1)}{=} A\Big[\frac{(B-t)^{1-n}}{n-1}\Big]_{-\infty}^{-\alpha} \overset{(2)}{=} \frac{A}{n-1}\Big(\frac{n}{\alpha}\Big)^{1-n} \overset{(3)}{=} \frac{n}{\alpha(n-1)}\,e^{-\alpha^2/2}. \tag{13.48}$$

(1) The antiderivative of $(B - t)^{-n}$ is $(B - t)^{1-n}/(n - 1)$; the minus signs from the chain rule and from $1 - n$ cancel.
(2) For $n > 1$ the term at $t \to -\infty$ vanishes; at $t = -\alpha$, $B - t = n/\alpha$.
(3) Insert A from (13.47): $(n/\alpha)^n(n/\alpha)^{1-n} = n/\alpha$.

The tail needs $n > 1$ to have a finite area at all. As $n \to \infty$ the tail area $\to e^{-\alpha^2/2}/\alpha$, which is the large-$\alpha$ estimate of the Gaussian's own tail beyond $-\alpha$ (Chapter 4): a very steep power law is a Gaussian tail again. The library `code/ch13/lib_hyy.py` writes the cumulative distribution of (13.46) from these two areas, so that the expected signal in every bin is an exact difference of two cumulative values.

Fitting it to the simulation. The simulated events carry the weights w_j of (13.45). A weighted maximum-likelihood fit maximises $\sum_j w_j \ln p(m_j \mid \bar{m}, \sigma, \alpha, n)$, where p is (13.46) normalised over the window 105–160 GeV: each simulated event counts as many times as the real events it stands for. The fit to all production modes together gives

$$\bar{m} = 125\,\mathrm{GeV}, \qquad \sigma = 2.467\,\mathrm{GeV}, \qquad \alpha = 1.88, \qquad n = 10.3,$$

and a plain Gaussian fitted the same way has width 2.6 GeV (Figure 13.42). The Gaussian is wider because it has to stretch its symmetric shape to cover the low tail, which also makes it too wide at the top. The Crystal Ball puts 3.3% of the signal more than 2σ below the peak, against 2.3% for a Gaussian. Above the peak the simulation shows a small tail of its own, which the one-sided function misses; later analyses use a Crystal Ball with a power law on both sides. Its effect here is a few events out of 397.9.

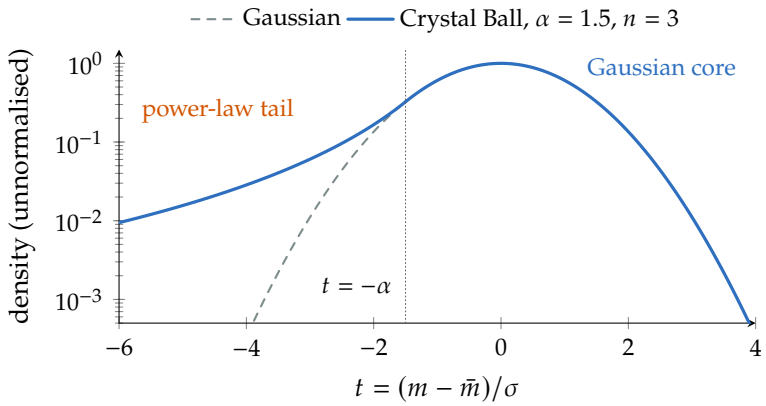


Figure 13.41: The Crystal Ball line shape on a logarithmic scale. To the right of $t = -\alpha$ it is the Gaussian; to the left it continues as a power law that joins with the same value and slope, (13.47), and falls far more slowly than the Gaussian (dashed). The excess area on the left is the events whose photons lost energy before being measured.

Example: the full width at half maximum. A peak is often quoted by its full width at half maximum (FWHM), because it can be read off a histogram without a fit. For a Gaussian, half maximum is where $e^{-t^2/2} = \frac{1}{2}$, so $t = \pm\sqrt{2\ln 2} = \pm 1.177$ and

$$\text{FWHM} = 2\sqrt{2\ln 2}\,\sigma = 2.355\,\sigma.$$

The Crystal Ball has the same half-maximum points as long as the tail starts beyond them, $\alpha > 1.177$, as it does here. The numerical FWHM of our fit is 5.81 GeV, $2.35\,\sigma$. ATLAS in 2012 quoted a FWHM of 3.9 GeV with $\sigma_{\text{CB}} = 1.6$ GeV for its inclusive sample [AadHiggs2012 §5.4]: a ratio of 2.4, slightly above 2.355 because their model adds a wider Gaussian (under 10% of the events) to the Crystal Ball and the inclusive sample mixes categories of different widths; for single categories they quote a ratio of typically 2.3. Their peak was narrower than ours by a factor $2.467/1.6 \approx 1.5$, a difference we return to in Section 13.27.

Example: the resolution depends on the photons. Fitted separately, the pairs in which both photons are central ($|\eta| < 0.75$) and unconverted have $\sigma = 1.825$ GeV, the rest $\sigma = 2.58$ GeV. Central unconverted photons cross the least material and land in the best part of the calorimeter. Section 13.25.3 uses this difference.

Other masses. The search tests masses m_H from 110 to 150 GeV, but the simulation exists only at 125 GeV. The calorimeter's relative resolution changes little over this range, so we scale both the peak position and the width with the mass, $\bar{m}(m_H) = \bar{m}\, m_H/125$ and $\sigma(m_H) = \sigma\, m_H/125$, and keep α and n. ATLAS interpolates between simulations at several masses [118].

13.22.2 Choosing the background function: the spurious signal

The background is mostly pairs of real photons from quark–antiquark annihilation and gluon fusion, plus photon–jet and jet–jet events where a jet passed as a photon [AadHiggs2012 §5]. No calculation predicts their sum after our cuts to a fraction of a per cent, which is the precision 76394 events demand. So the background is described by a smooth function with free parameters, fitted to the data together with the signal. The plain claim is: *any function that fits the data well will do*. The everyday question behind it is "does the function fit the data?", and we can answer it with the goodness-of-fit test of Section 4.8.

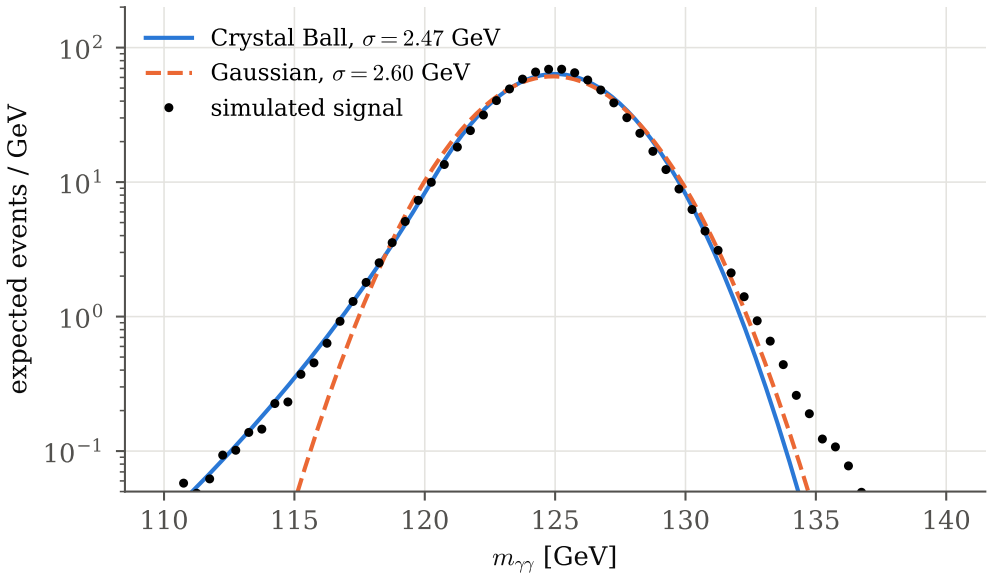


Figure 13.42: The simulated Higgs signal (points, expected events per GeV for 10 fb^{-1}, logarithmic scale) with the fitted Crystal Ball (solid) and Gaussian (dashed). The Gaussian falls far too fast below 118 GeV, where the photons that lost energy end up; the Crystal Ball's power law follows them.

Does each function fit? The script `code/ch13/32_hyy_model.py` fits eight families to the spectrum with no signal: exponentials of polynomials of degree one to four (Exp1 to Exp4, $g(m) \propto \exp \sum_{k=1}^{K} a_k u^k$ with u the mass mapped onto $[-1, 1]$), a power law m^{-a} (Pow) and Bernstein polynomials of degree three to five (Bern3 to Bern5, $g \propto \sum_{j=0}^{K} c_j \binom{K}{j} x^j (1-x)^{K-j}$ with x the mass mapped onto $[0, 1]$). The deviance $D = 2\sum_i [v_i - n_i + n_i \ln(n_i/v_i)]$ in 1 GeV bins is the likelihood-ratio goodness-of-fit statistic of Section 4.8.3, approximately χ^2 with $55 - 1 - K$ degrees of freedom.

	Exp1	Pow	Exp2	Bern3	Exp3	Bern4	Exp4	Bern5
free shape parameters K	1	1	2	3	3	4	4	5
deviance	78.3	63.2	58.4	54.7	54.5	54.2	54.3	53.4
p-value	0.013	0.16	0.252	0.337	0.342	0.318	0.315	0.308
max $\lvert S_{\text{spur}} \rvert$ (events)	707.9	245.8	273.8	98.1	218.6	79.3	40.8	15.6
max $\lvert S_{\text{spur}} \rvert / \sigma_S$	4.43	2.17	2.42	0.71	1.19	0.41	0.29	0.06

Only the single exponential is rejected at the one-per-cent level. Every other family fits.

The hidden assumption. "Fits the data" answers a question about all 55 bins at once. The question that matters for a search is narrower: *if there were no signal at all, would this function, fitted together with a signal template, return a signal of zero?* A function can follow the data within the errors everywhere and still be wrong by a few hundred events in a way that looks like a peak, because a deviation spread over a 6 GeV wide region of 1527 background events per GeV is a fraction of a per cent of the background there, below what the goodness of fit can see. The test that asks the right question is the **spurious-signal** test.

Working definition 13.3 (spurious signal)

Take a background spectrum with no signal in it, as realistic as possible (a "truth"). Fit it with the candidate background function plus a signal at mass m_H. The fitted signal $S_{\text{spur}}(m_H)$ [events the function invents, or hides when negative] is the spurious signal. ATLAS required $|S_{\text{spur}}|$ below 20% of the statistical error σ_S of the fitted signal at every m_H in 110–150 GeV [AadHiggs2012 §5.5]. *Operationally:* fit the truth's exact expectation (an Asimov spectrum), so that the result is the bias alone, with no fluctuation.

Where does a truth come from? ATLAS used large samples of simulated background. The open-data release has none for this selection, so we take as truths the two most flexible families, Bern5 and Exp4, each fitted to the whole data spectrum *together with* a 125 GeV signal, and keep only their background part. (Fitted without the signal they would have absorbed the real Higgs boson in the data, which is what we must not call background; their fitted signals were 496 and 462 events.) Each candidate is then fitted, with a signal, to each truth's Asimov spectrum scaled to 76394 events, at every m_H from 110 to 150 GeV in 1 GeV steps, skipping the truth's own family.

Example: the exponential of a quadratic. Exp2 fits the data with $p = 0.252$. Against the truths it invents up to 273.8 events, 2.42 times the statistical error σ_S of the signal: a function that passes the goodness-of-fit test would, on a background-only spectrum, produce a bump of more than two standard deviations by its shape alone. Figure 13.43 shows why: its spurious signal oscillates with m_H, positive where the truth bends above the Exp2 shape and negative where it bends below. A rigid function cannot follow a bend, so it uses the signal template to do it.

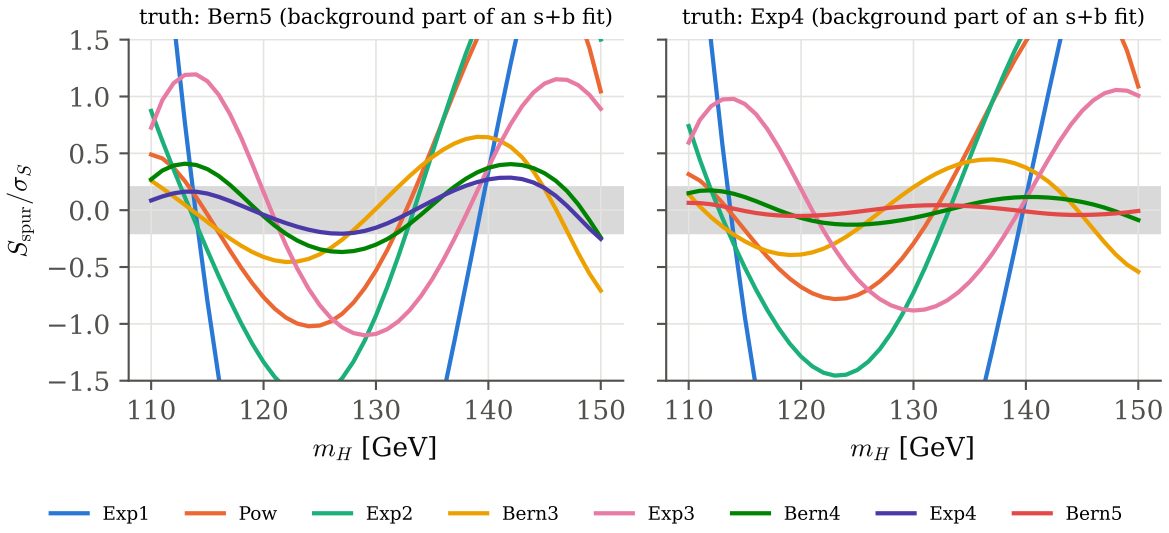


Figure 13.43: The spurious signal of each candidate background family, in units of the statistical error of the fitted signal, against the hypothesised Higgs mass. Left: the truth is the background part of a Bern5 fit to the data; right: of an Exp4 fit. The grey band is the ±20% criterion. The rigid families (Exp1, Pow, Exp2) swing far outside it; only the most flexible ones stay close to zero everywhere.

The verdict, and its limits. Only Bern5 passes, with $|S_{\text{spur}}| \leq 15.6$ events, $0.06\,\sigma_S$; Bern4 reaches $0.41\,\sigma_S$ against the Bern5 truth. We use Bern5 from here on. Two honest caveats.

The truths are themselves fits to the same data, so the test checks the families against each other rather than against an independent truth, and Bern5 has been tested only against Exp4. And with only the data, a more flexible truth would always make the next family fail. So, as ATLAS did, the spurious signal also enters the likelihood as a nuisance parameter: a term $\theta_{\text{ss}} S_{\text{spur}} f_i(m_H)$ added to the mean, with $S_{\text{spur}} = 15.6$ events and a unit Gaussian constraint on θ_{ss}. "The background model might be wrong by this much" becomes a systematic uncertainty of the signal.

The price of flexibility. A flexible function invents no signal, but it can also absorb part of a real one: a fifth-degree polynomial over 55 GeV can make a broad wiggle that resembles a 6 GeV-wide peak. How much sensitivity does that cost? *The question in words:* how much smaller is the expected q_0 when the background shape is fitted than when it is known? By Wald's approximation (Section 4.16.2), the expected q_0 for a signal μ is $q_{0,\text{A}} = \mu^2/V_{\mu\mu}$, with $V_{\mu\mu}$ the variance of $\hat{\mu}$ when the background parameters are fitted too. With the background known, the variance is the smaller $1/F_{\mu\mu}$, where F is the Fisher matrix. Then

$$\frac{q_{0,\text{A}}^{\text{fitted}}}{q_{0,\text{A}}^{\text{known}}} \overset{(1)}{=} \frac{\mu^2/V_{\mu\mu}}{\mu^2 F_{\mu\mu}} \overset{(2)}{=} \frac{1}{F_{\mu\mu}(\mathbf{F}^{-1})_{\mu\mu}} \overset{(3)}{=} 1 - \rho^2, \tag{13.49}$$

(1) The Asimov q_0 with the profiled and with the conditional variance.
(2) The variance of $\hat{\mu}$ with all parameters fitted is the $\mu\mu$ element of the inverse Fisher matrix (Chapter 10).
(3) The identity $1/F_{\mu\mu} = V_{\mu\mu}(1 - \rho^2)$ of Chapter 3, used in (4.93), with ρ^2 the fraction of the variance of $\hat{\mu}$ shared with the background parameters.

The background parameters play exactly the role the systematic uncertainty played in (4.93): whatever part of the signal's information they share is lost. On the Asimov data set of our spectrum with a $\mu = 1$ Higgs signal at 125 GeV (`code/ch13/33_hyy_search.py`):

background	known	Bern1	Bern2	Bern3	Bern4	Bern5	Bern6
expected $Z = \sqrt{q_{0,\text{A}}}$	3.39	2.88	2.87	2.51	2.5	2.16	2.01
$1 - \rho^2$	1	0.79	0.74	0.55	0.55	0.41	0.35

The expected significance falls with every added degree, and the second row explains the first: each Z is close to the known-background value times $\sqrt{1 - \rho^2}$. This is the trade-off ATLAS describes as a "compromise between limiting the size of a potential bias while retaining good statistical power" [AadHiggs2012 §5.5]: rigid functions bias the signal, flexible ones dilute it. The narrower the peak, the less a smooth polynomial can imitate it, which is one reason good mass resolution is worth so much.

13.23 Is there a peak? The discovery test

The spectrum of Figure 13.40 holds 76394 photon pairs, and a Standard Model Higgs boson would add about 397.9 of them, piled within a few GeV of 125 GeV. The general problem is old and simple to state: a small number of extra events is spread over a few bins of a histogram that already contains many events; how should the counts be combined to decide whether the extra events are there? Before writing the full likelihood, the smallest version of the problem shows what matters.

The smallest example: one window. Count the events in a mass window and compare the count with the background expected in it. If the window holds s signal and b background events with b large, the count fluctuates by $\sqrt{b}$, and the excess is worth about $s/\sqrt{b}$ standard deviations (Eq. (4.108) with $s \ll b$). The window's width is the only dial, and it acts on both numbers at once. *The naive choice:* use the whole window 105–160 GeV, which certainly contains all the signal. Then $s \approx 398$ and $b \approx 76\,000$, so

$$\frac{s}{\sqrt{b}} \approx \frac{398}{\sqrt{76\,000}} = \frac{398}{276} = 1.44.$$

Most of those 76 000 events are far from 125 GeV and carry no information about the peak, but each adds its fluctuation. *A better choice:* keep only $m_H \pm k\sigma$ around the peak, with $\sigma = 2.467$ GeV the width of the signal from Section 13.22.1. A Gaussian peak puts a fraction $\mathrm{erf}(k/\sqrt{2})$ of its events inside, while the background in the window is its density β times the width $2k\sigma$. Near 125 GeV the background density is $\beta = 1527$ events per GeV. The ratio $\mathrm{erf}(k/\sqrt{2})/\sqrt{2k}$ is largest near $k = 1.4$, where the window keeps 84% of the signal, so

$$\frac{s}{\sqrt{b}} = \frac{0.84 \times 398}{\sqrt{1527 \times 2.8 \times 2.47}} = \frac{334}{103} = 3.25.$$

Cutting away the empty regions more than doubled the significance. *The price:* the number b under the peak must now be known, and it is not measured inside the window, where the signal sits; it has to be interpolated from the sidebands with a background function. Every uncertainty of that interpolation lowers the 3.25. The lesson in one line: *the events near the peak carry the information, and the background under the peak is what we pay to know.* A likelihood does better than any window: it weights every bin by how much signal it can hold, fits the background under the peak together with the signal, and can ask the discovery question of Section 4.16 at every mass from 110 to 150 GeV.

13.23.1 The likelihood of the spectrum

The mean count in each bin. The histogram now has 220 bins of 0.25 GeV between 105 and 160 GeV, fine enough that the shape of the peak (about ten bins wide) is not smeared by the binning. Each count n_i is Poisson, as in Section 4.15.1, with a mean built from four pieces:

$$\nu_i = \underbrace{\mu\,(1+\delta_y)^{\theta_y}\, s_i(m_H;\theta_\sigma,\theta_m)}_{\text{Higgs signal}} + \underbrace{\theta_{\text{ss}}\, S_{\text{spur}}\, f_i(m_H)}_{\text{spurious signal}} + \underbrace{\sum_{j=0}^{5} c_j\, b_{j,5}(m_i)}_{\text{background}}. \tag{13.50}$$

Each piece, one at a time:

- $s_i(m_H)$ is the number of Higgs events the simulation predicts in bin i for a Standard Model Higgs boson of mass m_H: the 397.9 events of Section 13.21.1 distributed with the Crystal Ball shape of (13.46). The **signal strength** μ [the observed signal divided by the Standard Model prediction] is the parameter of interest: $\mu = 0$ is no Higgs boson, $\mu = 1$ is the Standard Model.
- θ_y moves the expected number of signal events by a relative $\delta_y = 10\%$ per unit. It stands for everything that scales the signal as a whole: the luminosity, the photon identification and isolation efficiencies. ATLAS quoted 8–11% for the identification and 1.8–3.6% for the luminosity [AadHiggs2012 §5.6]. The power form $(1+\delta_y)^{\theta_y}$ keeps the yield positive for any θ_y.

- θ_σ changes the width of the peak by a factor $(1+\delta_\sigma)^{\theta_\sigma}$ with $\delta_\sigma = 14\%$, the resolution uncertainty of the 2012 analysis [AadHiggs2012 §5.6].
- θ_m shifts the peak by a relative $\delta_m \theta_m$, with $\delta_m = 0.0086$, the photon energy-scale uncertainty that the release stores for every photon, averaged over the selected signal events (a 1% shift of both photon energies moves $m_{\gamma\gamma} \propto \sqrt{E_1 E_2}$ by 1%).
- $\theta_{\text{ss}} S_{\text{spur}} f_i$ is the spurious signal of Section 13.22.2: a signal-shaped term of size $S_{\text{spur}} = 15.6$ events, with f_i the signal shape normalised to one.
- $b_{j,5}(m)$ are the six Bernstein basis polynomials of degree five chosen in Section 13.22.2, and the c_j are free. They describe the background everywhere, under the peak included, without any separate sideband.

Constraints. The four θ's are calibrations measured outside this spectrum. As in Section 4.15.2, each one comes with an auxiliary measurement a_k whose likelihood is a Gaussian of unit width, so that $\theta_k = \pm 1$ is one standard deviation of that calibration, and $a_k = 0$ for the real data. The whole likelihood is (4.89), and the function we minimise is

$$-2 \ln \mathcal{L}(\mu, \boldsymbol{c}, \boldsymbol{\theta}) = 2 \sum_{i=1}^{220} \left[\nu_i - n_i \ln \nu_i \right] + \sum_k \left(\theta_k - a_k \right)^2 + \text{const}, \tag{13.51}$$

with eleven parameters: μ, six c_j and four θ_k. The first sum is -2 times the log of the Poisson probabilities, without the $\ln n_i!$ that does not depend on the parameters; the second is -2 times the log of the four Gaussian constraints. `code/ch13/lib_hyy.py` codes (13.50) and (13.51) in its class `Search` and minimises with MINUIT through `iminuit`.

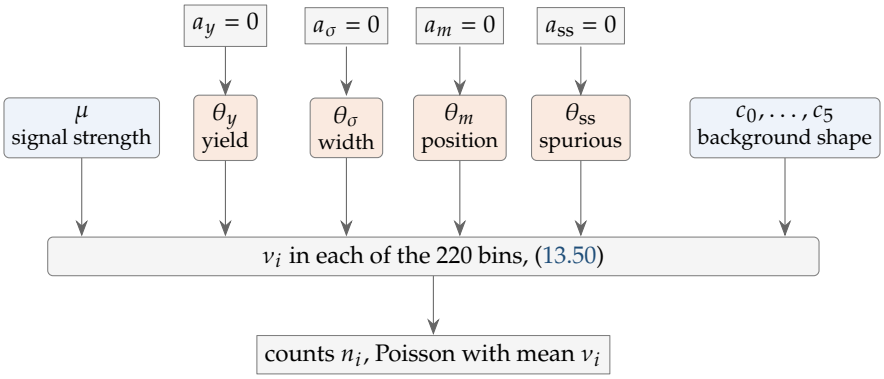


Figure 13.44: The structure of the diphoton likelihood. The signal strength μ and the six background coefficients vary freely in the fit, while a Gaussian constraint of unit width connects each of the four nuisance parameters (orange) to its auxiliary measurement a_k. All eleven parameters together set the mean count of every bin, and the observed counts are Poisson about those means.

13.23.2 The scan: local significance

The question at one mass. Fix a mass, say $m_H = 125\,\text{GeV}$. *If there were no Higgs boson ($\mu = 0$), would an excess like the one in the data, or a larger one, be something we would reasonably expect?* The null hypothesis is "background only", because its sampling distribution is the one we can work out: with $\mu = 0$, the statistic q_0 of (4.98) follows the half-χ^2_1 law of (4.101) for a large sample. So the procedure is: fit (13.51) twice, once with μ free and once with $\mu = 0$, the background and nuisance parameters re-fitted each time; take q_0 as twice the difference of the two minima of $-\ln \mathcal{L}$ when $\hat{\mu} > 0$ (and 0 otherwise); then

$$p_0 = \text{P}(q_0 \geq q_{0,\text{obs}} \,|\, \mu = 0) = 1 - \Phi\left(\sqrt{q_{0,\text{obs}}}\right), \qquad Z = \Phi^{-1}(1 - p_0) = \sqrt{q_{0,\text{obs}}}, \tag{13.52}$$

from (4.101) with $\mu' = 0$. The significance Z is simply the square root of q_0: that is what makes q_0 convenient.

The scan. We do not know m_H, so `code/ch13/33_hyy_search.py` repeats the test at every m_H from 110 to 150 GeV in steps of 0.5 GeV. Figure 13.46 shows the result. The largest excess is at $m_H = 124.5$ GeV, with

$$q_{0,\mathrm{obs}} = 7.6, \qquad Z_{\mathrm{loc}} = 2.77, \qquad p_0 = 0.002846, \qquad \hat{\mu} = 1.28.$$

The word *local* is a warning: this is the p-value of a test made at one mass chosen in advance, while we chose 124.5 GeV after looking at all of them. The correction is the subject of Section 13.24. With all four nuisance parameters held at their nominal values the largest Z is 2.77: the calibrations cost almost nothing here, because a 10% uncertainty on the yield or a 14% uncertainty on the width changes q_0 only at second order when the signal is a small fraction of the data. Away from the largest excess the two curves of Figure 13.46 do separate, most visibly near 140 GeV. There the energy-scale parameter θ_m can slide the signal template a few GeV towards the excess at 143 GeV, paying for it with its constraint term θ_m^2. A nuisance parameter that moves the peak competes with the scanned mass itself; at the top of an excess it has nowhere better to go, which is why the largest q_0 is unchanged. The other excesses in the scan are what background fluctuations look like: the largest one more than 8 GeV away from the peak is at 143 GeV with $Z = 2.71$.

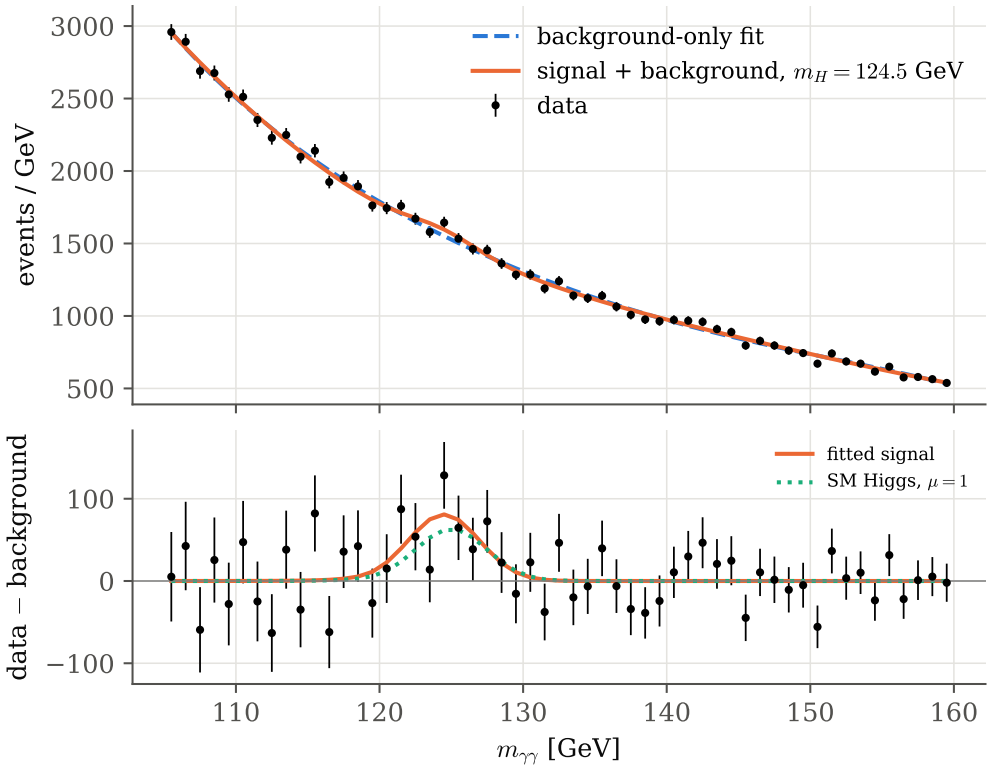


Figure 13.45: Top: the data in 1 GeV bins with the background-only fit (dashed) and the best signal-plus-background fit at $m_H = 124.5$ GeV (solid). Bottom: the same data with the fitted background subtracted. The fitted signal (orange) and the Standard Model expectation (dotted) are of similar size; the residuals far from the peak scatter about zero as Poisson noise of about $\pm\sqrt{1600} = \pm 40$ events per GeV.

Example: the excess in events. At 124.5 GeV the fit gives $\hat{\mu} = 1.28$, that is $1.28 \times 397.9 \approx 510$ extra events spread over the peak, against a background of about 1527 events per GeV. The bottom panel of Figure 13.45 shows them as a bump of about 80 events per GeV at its top on residuals that scatter by ±40: a bump no single bin could establish, which the likelihood finds by adding the evidence of all the bins it covers.

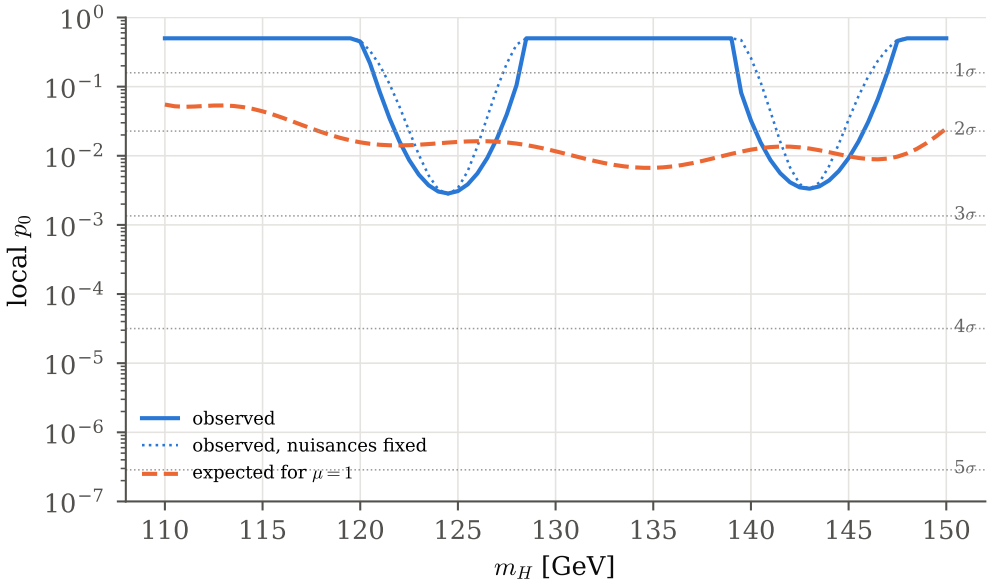


Figure 13.46: The local p-value p_0 of the background-only hypothesis against the tested mass (solid), with the nuisance parameters fixed at their nominal values (dotted) and the median expected for a Standard Model Higgs boson at each mass (dashed, from Asimov data sets). The horizontal lines mark 1σ to 5σ. The observed curve dips where the data have an excess, and the expected curve says how deep the dip should be if the Standard Model Higgs boson were at that mass.

13.23.3 How large an excess should we expect?

The question. A local Z of 2.77 is neither a discovery nor nothing. To judge it we need to know what a Standard Model Higgs boson would typically produce in this data set. *If the truth were $\mu = 1$ at 125 GeV, what q_0 would we typically see?* The answer is the Asimov data set of Section 4.16.3: replace the counts by their expected values $n_i = b_i + s_i$, with b_i the fitted background, and compute q_0 on them. By (4.107) the result $q_{0,\mathrm{A}}$ is the median of q_0 under $\mu = 1$, so $\sqrt{q_{0,\mathrm{A}}}$ is the median significance. The script does this at each mass, which gives the dashed curve of Figure 13.46; at 125 GeV,

$$Z_{\mathrm{exp}} = \sqrt{q_{0,\mathrm{A}}} = 2.15.$$

The data show 2.74 at 125 GeV, and $\hat{\mu} = 1.28$ at the peak: a little more signal than the Standard Model predicts, within the fluctuations Section 13.25 quantifies.

Where the expected significance comes from. *The question:* which features of the data set the number 2.15? Start from the simplest case, the background known exactly and no nuisance parameters. The bins are independent, so $\ln \mathcal{L}$ is a sum over bins and so is q_0. In each bin the

Asimov value is the counting result (4.108). With $x_i = s_i/b_i$, which is below 0.1 in every bin,

$$q_{0,\mathrm{A}} \overset{(1)}{=} 2\sum_i \left[(s_i+b_i)\ln\!\left(1+\frac{s_i}{b_i}\right) - s_i\right] \overset{(2)}{=} 2\sum_i \left[b_i(1+x_i)\left(x_i - \frac{x_i^2}{2} + \frac{x_i^3}{3} - \cdots\right) - b_i x_i\right]$$
$$\overset{(3)}{=} 2\sum_i b_i\left[\frac{x_i^2}{2} - \frac{x_i^3}{6} + \cdots\right] \overset{(4)}{\approx} \sum_i \frac{s_i^2}{b_i}. \tag{13.53}$$

(1) (4.108) applied bin by bin and summed, since q_0 is a sum of independent bins when the background is known.
(2) Write $s_i = b_i x_i$ and expand $\ln(1+x) = x - x^2/2 + x^3/3 - \cdots$.
(3) Multiply out: $(1+x)(x - x^2/2 + x^3/3) = x + x^2/2 - x^3/6 + \cdots$, and the leading x cancels against $-b_i x_i$.
(4) Keep the first term, $b_i x_i^2 = s_i^2/b_i$; the next is smaller by a factor $x_i/3$.

Each bin contributes the square of its own counting significance $s_i/\sqrt{b_i}$, and squares add. This is the window of the opening example with every bin given its own weight: a bin with no signal contributes nothing (instead of adding noise), and the bins at the top of the peak contribute most. On our spectrum $\sqrt{\sum_i s_i^2/b_i} = 3.41$.

The same in closed form. A peak of S events with a Gaussian shape of width σ, on a background of density β events per GeV that is nearly flat across it, has $s_i = S\,\varphi(m_i)\,\Delta$ and $b_i = \beta\Delta$ in bins of width Δ, with φ the Gaussian density. For bins much narrower than σ the sum becomes an integral:

$$q_{0,\mathrm{A}} \overset{(1)}{=} \sum_i \frac{S^2\varphi(m_i)^2\Delta^2}{\beta\Delta} \overset{(2)}{=} \frac{S^2}{\beta}\int \varphi(m)^2\,\mathrm{d}m \overset{(3)}{=} \frac{S^2}{\beta}\cdot\frac{1}{2\pi\sigma^2}\int e^{-(m-m_H)^2/\sigma^2}\mathrm{d}m$$
$$\overset{(4)}{=} \frac{S^2}{\beta}\cdot\frac{\sqrt{\pi}\,\sigma}{2\pi\sigma^2} = \frac{S^2}{2\sqrt{\pi}\,\sigma\,\beta}. \tag{13.54}$$

(1) Insert s_i and b_i into (13.53).
(2) One factor Δ is left: $\sum_i g(m_i)\Delta \to \int g\,\mathrm{d}m$.
(3) The square of $\varphi(m) = e^{-(m-m_H)^2/2\sigma^2}/(\sqrt{2\pi}\sigma)$.
(4) The Gaussian integral $\int e^{-u^2/a^2}\mathrm{d}u = \sqrt{\pi}\,a$ with $a = \sigma$.

Equation (13.54) says the peak acts like a counting window of effective width $2\sqrt{\pi}\,\sigma = 3.54\,\sigma$ holding all S events: $Z^2 = S^2/(\beta \times 3.54\sigma)$. It shows the three dials of every resonance search at once. The significance grows like the signal S, falls like the square root of the background density, and falls like the square root of the mass resolution. Halving σ is worth as much as halving the background density β: either one doubles Z^2.

Example: our numbers. With $S = 397.9$ events in the window, $\sigma = 2.467$ GeV and $\beta = 1527$ per GeV,

$$q_{0,\mathrm{A}} \approx \frac{398^2}{2\times 1.772\times 2.467\times 1527} = \frac{158\,400}{13\,350} = 11.9, \qquad Z \approx 3.44,$$

the value 3.44 printed by the script, against 3.41 from the exact sum with the Crystal Ball shape. The tail of the Crystal Ball puts a few per cent of the signal where the background is large, which costs a little.

From 3.4 to 2.15: what the fit costs. The full Asimov value at 125 GeV is 2.15, well below 3.41. The difference is the price paid at the end of the opening example: the background under the peak is not known but fitted. By (13.49), fitting six background coefficients together with μ keeps a fraction $1 - \rho^2 = 0.41$ of $q_{0,\mathrm{A}}$, and

$$3.41 \times \sqrt{0.41} \approx 2.16,$$

the expected significance with the degree-five background in the table of Section 13.22.2. The nuisance parameters take the rest of the way to 2.15. So most of the loss is the background function, and very little is the calibrations. A real analysis fights this loss by splitting the events into categories with different σ and β (Section 13.25.3) and by using a background function no more flexible than the spurious-signal test demands.

Watch out

Z_{exp} is the *median* significance under $\mu = 1$, not a prediction of what this data set will show. The observed q_0 fluctuates about $q_{0,\mathrm{A}}$ with a spread of about $2\sqrt{q_{0,\mathrm{A}}}$ (Eq. (4.100): $\sqrt{q_0} = \hat{\mu}/\sigma_\mu$ is Gaussian with unit width), so with $Z_{\mathrm{exp}} = 2.15$ a true Standard Model Higgs boson gives Z anywhere between about 1.2 and 3.2 in two thirds of repetitions.

13.24 Anywhere in the range: checking the formulas and looking elsewhere

Two claims of Section 13.23 rest on formulas rather than on the data. The local p-value 0.002846 came from the half-χ^2 law (4.101), which Wald's approximation guarantees only for a large sample and a well-behaved fit, and here the fit has eleven parameters and a background that bends. And the excess at 124.5 GeV was picked as the largest of 81 tested masses, so its local p-value is not the probability that background alone produces such an excess *somewhere*. Both are questions about the sampling distribution of q_0 under the background-only hypothesis, and both can be answered by simulating the experiment many times.

What we saw in the toy bump hunt, and what is different now. In Section 4.14.3 a simulated spectrum of 3000 events with a known exponential background was scanned for a Gaussian peak of width 1.5 GeV over 50 GeV. Its largest bump had $Z_{\mathrm{loc}} = 3.28$; 20000 background-only scans made that a global 2.15σ, a trials factor of 31.3, and the Gross–Vitells formula (4.85), calibrated on 100 scans, reproduced it. The method is the same now. Three things differ, and each changes the size of the correction. The peak is wider, $\sigma = 2.467$ GeV against 1.5, and the range is shorter, 40 GeV against 50: there are fewer independent places for a fluctuation to appear, roughly $40/2.5 = 16$ against $50/1.5 = 33$. The background shape is fitted in every toy, with six coefficients, so a toy that fluctuates in a smooth, broad way is absorbed by the background rather than counted as a bump. And the local significance is lower, 2.77 against 3.28, and by (4.86) the trials factor grows with the local significance. All three point to a smaller trials factor than the 31.3 of the toy.

How the toys are made. `code/ch13/34_hyy_toys.py` draws each background-only toy as 220 Poisson counts whose means are the background-only fit to the data, b_i. Every toy is then analysed exactly as the data: the degree-five Bernstein background is refitted, and q_0 is computed at all 81 masses. Fitting 81 masses for each of thousands of spectra with MINUIT would take hours, but the model with the four nuisance parameters fixed is linear in its seven free parameters (μ and the six c_j). For a mean linear in the parameters, $\ln \mathcal{L} = \sum_i (n_i \ln \nu_i - \nu_i)$

is concave, so Newton's method converges from any reasonable start, and it can be run on 250 spectra at once as matrix operations. At the largest excess the nuisance parameters changed Z by less than 0.01 (Section 13.23.2), and the largest q_0 of a scan is what the global p-value counts, so fixing them in the toys leaves it unchanged. Scanned with this fitter, the data give $q_{0,\text{obs}} = 7.7$ at 124.5 GeV, essentially the 7.6 of the full fit.

13.24.1 Is the half-χ^2 law right?

The question: if there were no Higgs boson, how would q_0 at one fixed mass, 125 GeV, be distributed over repetitions of the experiment? The half-χ^2 law makes three checkable predictions. Half of the toys should have $\hat{\mu} \le 0$ and so $q_0 = 0$ exactly; the toys give 0.508. The fraction with $q_0 > 4$ (a local 2σ) should be $\Phi(-2) = 0.0228$; the toys give 0.0228. The fraction with $q_0 > 9$ (a local 3σ) should be $\Phi(-3) = 0.0013$; the toys give 0.0012. With 4000 toys the statistical error of a fraction p is $\sqrt{p(1-p)/4000}$, about 0.002 at 0.023 and 0.0006 at 0.0013, so the agreement is as good as the toys can tell. Figure 13.47 (left, blue) shows the whole tail.

The same question under the signal hypothesis tells us whether the Asimov expected significance is the median, as (4.107) says. Toys drawn with $\mu = 1$ at 125 GeV have a median $\sqrt{q_0}$ of 2.15, against the Asimov value 2.16, and the central 68% lie between 1.14 and 3.13: a spread of one unit in Z, as the pitfall of Section 13.23.3 said. Only a fraction 0.002 of experiments of this size would see a Standard Model Higgs boson at 5σ. With 76 000 events the asymptotic formulas are doing exactly what they promise; they were derived for this situation.

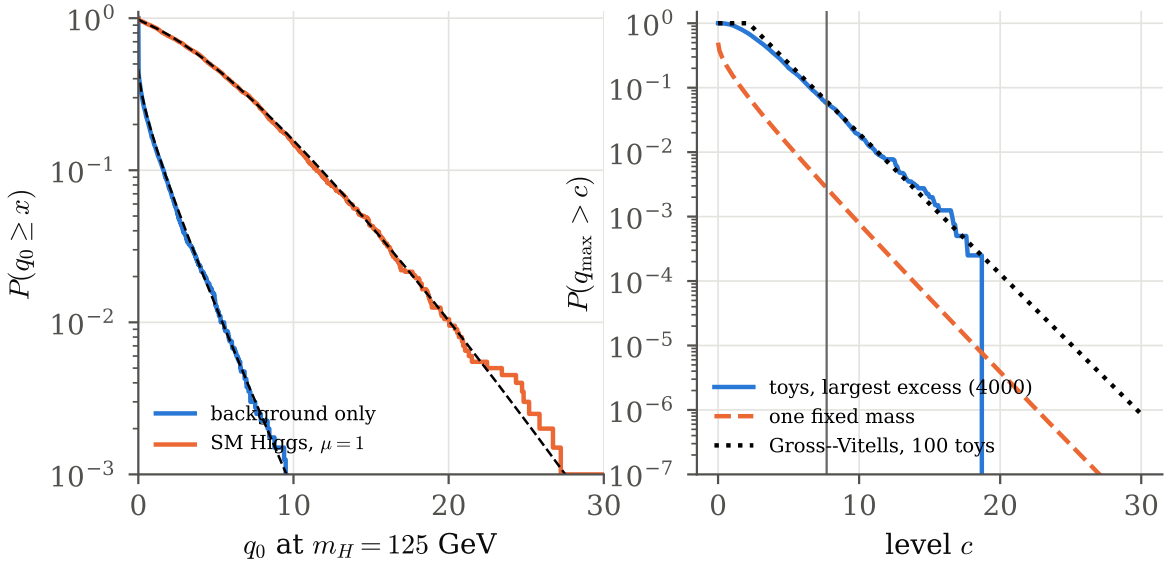


Figure 13.47: Left: the probability that q_0 at 125 GeV exceeds x, from toys with background only (blue) and with a Standard Model Higgs boson (orange), against the asymptotic formulas (black dashed). The step at $x = 0$ in the blue curve is the half of the toys with $\hat{\mu} \le 0$. Right: the probability that the largest q_0 anywhere in 110–150 GeV exceeds a level c (blue, 4000 toys), the probability at one fixed mass (orange dashed), and the Gross–Vitells formula calibrated on the first 100 toys (black dotted). The vertical line is the excess in the data.

13.24.2 The global p-value

The question. *If there were no Higgs boson, how often would an excess as large as the one we saw appear somewhere between* 110 *and* 150 *GeV?* The event is now "the largest q_0 in the scan is at

least $q_{0,\mathrm{obs}} = 7.7''$. Each toy gives one largest q_0, so the answer is a count: 229 of the 4000 background-only toys have $q_{\max} \geq 7.7$, so

$$p_{\mathrm{glob}} = 0.0572 \pm 0.0037,$$

with the binomial error $\sqrt{p(1-p)/N}$ of Chapter 4.

The same from a hundred toys. For a stronger excess the counting would need millions of toys. The Gross–Vitells formula (4.85) needs only the mean number of upcrossings of a low level, which a few scans measure well. Counting how often each of the first 100 toy scans crosses $c_0 = 0.5$ upwards gives $\mathbb{E}[N(c_0)] = 2.14$ (all 4000 toys: 2.06). Then, at $c = q_{0,\mathrm{obs}}$,

$$\begin{aligned} p_{\mathrm{glob}} &\overset{(1)}{=} p_{\mathrm{loc}}(c) + \mathbb{E}[N(c_0)]\, e^{-(c-c_0)/2} \overset{(2)}{=} \Phi\big(-\sqrt{7.7}\big) + 2.14\, e^{-(7.7-0.5)/2} \\ &\overset{(3)}{=} 0.0613, \qquad Z_{\mathrm{glob}} = 1.54. \end{aligned} \tag{13.55}$$

(1) The Gross–Vitells formula (4.85): the chance of being above c at the start of the range plus the expected number of upward crossings of c, scaled down from the reference level c_0 by $e^{-(c-c_0)/2}$.
(2) Insert the observed q_0 and the measured $\mathbb{E}[N(c_0)]$.
(3) Evaluate, and convert to a significance with $Z = \Phi^{-1}(1-p)$.

The two routes agree within the error of the toy count, and the right panel of Figure 13.47 shows that they agree over the whole curve, not only at the observed value. With the full likelihood (all nuisance parameters free) the local significance is 2.77, and the same formula gives $Z_{\mathrm{glob}} = 1.53$.

The trials factor. The ratio of the two p-values is $p_{\mathrm{glob}}/p_{\mathrm{loc}} = 22.2$. Its asymptotic form (4.86), $1 + \sqrt{2\pi}\, \mathcal{N}_1 Z$ with $\mathcal{N}_1 = \mathbb{E}[N(c_0)]e^{c_0/2} = 2.75$, gives 20.1. Both are below the 31.3 of the toy bump hunt, for the three reasons met there: a wider peak and a shorter range, a fitted background and a lower local significance. The number $\mathcal{N}_1 = 2.75$, against 3.60 in Section 4.14.3, measures how many independent places the scan offers. It is smaller, as expected, but not half as small as the crude count of range over width (16 against 33) suggested: that count is only a rough guide, and the upcrossings measure the real number directly.

Example: how often background alone gives a local 3σ somewhere. At a fixed mass a local 3σ excess has probability 0.0013. Somewhere in the range it is far more common: the toys give 0.032 and the formula 0.032, so roughly one search in thirty of this size shows a local 3σ bump with no Higgs boson at all. That is why particle physics asks for 5σ before it says "discovery": a 3σ excess in a scan is a reason to look again with more data, not a result.

The data's second bump. The scan of Figure 13.46 has a second excess, at 143 GeV with local $Z = 2.71$, almost as large as the one at the Higgs mass. Nothing is expected there, and the global p-value is exactly the tool that tells us not to be impressed: an excess of that size somewhere in the range happens in about 0.0572 of background-only experiments. The data alone cannot say which of the two bumps is the Higgs boson. What singles out 125 GeV is outside this spectrum: the 2012 discovery in this and other decay channels, and the expected size of the signal.

Which range? The global p-value depends on the range searched, and the range has to be fixed before the data are seen. In 2012 ATLAS quoted both: a global 5.1σ for 110–600 GeV and

5.3σ for 110–150 GeV, from a local 5.9σ in all channels combined [AadHiggs2012 §9]. For us, in 2016 data, the mass was already known to a fraction of a GeV, and the honest test is the local one at 125 GeV: $Z = 2.74$. Had we searched only 120–130 GeV, the expected number of upcrossings would be a quarter of that for 40 GeV, and the formula gives $Z_{\text{glob}} = 2.1$.

How this is used in real research: Toys in a real analysis

The LHC experiments compute local significances with the asymptotic formulas, check them with toys at a few masses as above, and quote global significances from the Gross–Vitells formula or a variant with several search dimensions (mass and width) [15, 27]. Toys are run in full only where the formulas are in doubt: few events, a parameter near a boundary, or a statistic far in its tail. Each toy there is a full fit with every nuisance parameter, and the auxiliary measurements a_k are drawn afresh in each toy, which our statistics-only toys skip.

13.25 How strong, and where? Measuring μ and m_H

Once a signal is taken to be present, the question changes from "is it there?" to "how much of it is there, and where?". The first is a test with one answer, yes or no at a chosen level; the second is a measurement with an interval. The data are the same 76394 masses, the likelihood is the same (13.51), and the tool is the profile likelihood ratio $\lambda(\mu)$ of (4.92) read as a curve rather than at one point.

13.25.1 The signal strength

The question. *Which values of μ at $m_H = 125$ GeV are compatible with the data?* For each trial μ we fit everything else (the six background coefficients and the four nuisance parameters) and record how much worse the best fit is than the overall best: $-2\ln\lambda(\mu)$. By Wald's approximation (4.100), $-2\ln\lambda(\mu) \approx (\mu - \hat{\mu})^2/\sigma_\mu^2$, a parabola that reaches 1 at $\hat{\mu} \pm \sigma_\mu$. So the values where the curve crosses 1 bound the 68% interval, and this remains true when the curve is not a parabola, because a reparametrisation that would make it one does not move the crossing points (Section 4.11.5).

The result. Figure 13.48 (left) shows the curve. It crosses 1 at

$$\hat{\mu} = 1.28\,^{+0.55}_{-0.49} \qquad (m_H = 125\,\text{GeV}).$$

The Standard Model value $\mu = 1$ lies inside; $\mu = 0$, no Higgs boson, is at $-2\ln\lambda = 2.74^2 \approx 7.5$, which is the q_0 of Section 13.23.2 read from the other side. The interval is wider above than below. The yield enters as $\mu(1+\delta_y)^{\theta_y}$, so the uncertainty it carries is a fixed fraction of μ: for large μ it is larger in absolute terms.

How much of it is statistics? Repeating the scan with the four nuisance parameters fixed at their nominal values gives the dashed curve, crossing 1 at ± 0.46. That is the uncertainty we would have with perfect calibrations: the statistical part. If the statistical and the systematic fluctuations of $\hat{\mu}$ are independent and Gaussian, their variances add, and the systematic part follows by subtraction:

$$\sigma_{\text{tot}}^2 \overset{(1)}{=} \sigma_{\text{stat}}^2 + \sigma_{\text{syst}}^2 \quad\Longrightarrow\quad \sigma_{\text{syst}} \overset{(2)}{=} \sqrt{\sigma_{\text{tot}}^2 - \sigma_{\text{stat}}^2} = \sqrt{0.52^2 - 0.46^2} \approx 0.24. \tag{13.56}$$

(1) The variance of a sum of independent fluctuations is the sum of their variances.
(2) Solve for the systematic part, with $\sigma_{\text{tot}} = 0.52$ the half-width of the total interval.

So $\hat{\mu} = 1.28 \pm 0.46\,(\text{stat}) \pm 0.24\,(\text{syst})$. A 10% yield uncertainty alone would give $0.10 \times 1.28 \approx 0.13$; the width and the position of the peak add the rest, because a wider or shifted template changes how many events the fit assigns to the peak. Statistics dominate: this measurement would improve with more data long before better calibrations mattered.

What the nuisance parameters did. At the best fit the four nuisance parameters sit at $\hat{\theta}_y = 0.00$ and $\hat{\theta}_{\text{ss}} = 0.00$ (to two decimals), $\hat{\theta}_\sigma = 0.03$ and $\hat{\theta}_m = -0.3$, all well within one standard deviation of their calibrations. The data cannot distinguish a larger yield from a larger μ, so they leave θ_y where its constraint puts it. Only θ_m moves: the data's peak sits a little below the simulated one. A nuisance parameter that the fit pushes far from zero would be a sign that a calibration and the data disagree, and would have to be investigated before quoting anything.

Example: compared with the Standard Model and with 2012. Our $\hat{\mu} = 1.28 \pm 0.52$ differs from the Standard Model value by $(1.28 - 1)/0.52 \approx 0.5$ standard deviations. ATLAS measured $\hat{\mu} = 1.8 \pm 0.5$ in the diphoton channel in 2012 [AadHiggs2012 Table 7], a value 1.6 standard deviations above one. The two results differ by $(1.8 - 1.28)/\sqrt{0.5^2 + 0.52^2} \approx 0.7$ standard deviations: entirely compatible. The high 2012 value was partly the luck of the experiment that made the discovery: a search that happens to fluctuate upwards is the one that crosses 5σ first.

13.25.2 The mass

The question. *Which Higgs masses are compatible with the data?* The same recipe applies with m_H in the place of μ: for each trial m_H from 120 to 130 GeV, fit everything else, μ included, and plot $-2\ln\lambda(m_H)$ (Figure 13.48, right). It crosses 1 at

$$\hat{m}_H = 124.5 \pm 1.4\,\text{GeV} \quad \text{(all nuisances)}, \qquad \hat{m}_H = 124.5\,^{+0.8}_{-0.9}\,\text{GeV} \quad \text{(nuisances fixed)}.$$

Here the systematic part is larger than for μ. The photon energy scale is known to $\delta_m = 0.0086$, which at 125 GeV is ± 1.07 GeV, and a shift of the energy scale moves the peak exactly as a shift of m_H does, so the data cannot separate them. In quadrature, $\sqrt{0.85^2 + 1.07^2} = 1.37$, the total.

How precisely should a peak locate a mass? The statistical ± 0.85 GeV is far smaller than the width $\sigma = 2.467$ GeV of the peak. *The question:* how does the position error depend on the peak? The Fisher information on the position m_0 of a peak with $s_i = S\varphi(m_i - m_0)\Delta$ on a flat background $b_i = \beta\Delta$ is $\sum_i (\partial s_i/\partial m_0)^2/b_i$, the same kind of sum as (13.53) with the derivative of the signal in the place of the signal (Chapter 10). For a Gaussian peak, $\partial\varphi/\partial m_0 = \big((m - m_0)/\sigma^2\big)\varphi$:

$$\begin{aligned} \frac{1}{\sigma_m^2} &\overset{(1)}{=} \frac{S^2}{\beta}\int \frac{(m-m_0)^2}{\sigma^4}\,\varphi(m-m_0)^2\,\mathrm{d}m \overset{(2)}{=} \frac{S^2}{\beta}\cdot\frac{1}{2\pi\sigma^6}\int u^2 e^{-u^2/\sigma^2}\mathrm{d}u \\ &\overset{(3)}{=} \frac{S^2}{\beta}\cdot\frac{1}{2\pi\sigma^6}\cdot\frac{\sqrt{\pi}\,\sigma^3}{2} = \frac{S^2}{4\sqrt{\pi}\,\sigma^3\beta} \overset{(4)}{=} \frac{q_{0,\text{A}}}{2\sigma^2}. \end{aligned} \tag{13.57}$$

(1) The sum over narrow bins becomes an integral, as in (13.54).
(2) Write $u = m - m_0$ and square $\varphi = e^{-u^2/2\sigma^2}/(\sqrt{2\pi}\sigma)$.
(3) The second moment of a Gaussian: $\int u^2 e^{-u^2/a^2}\mathrm{d}u = \sqrt{\pi}\,a^3/2$ with $a = \sigma$.
(4) Compare with $q_{0,\text{A}} = S^2/(2\sqrt{\pi}\sigma\beta)$ of (13.54).

So $\sigma_m = \sqrt{2}\,\sigma/Z$, with $Z = \sqrt{q_{0,\mathrm{A}}}$ the significance of the peak: a peak located to a fraction of its width, and the fraction is $\sqrt{2}$ over its significance. A barely significant peak locates nothing; a 5σ peak locates its mass to 0.28 of its width. For the signal the data show, $\hat{\mu} = 1.28$ times the known-background $Z = 3.39$, this gives $\sqrt{2} \times 2.467/(1.28 \times 3.39) \approx 0.80\,\mathrm{GeV}$, close to the ± 0.85 of the profile. (The fitted background costs little here: a smooth polynomial cannot imitate the antisymmetric change that shifting a peak makes.)

Example: the 2012 mass. For the diphoton channel in 2012, $\sigma_{\mathrm{CB}} = 1.6\,\mathrm{GeV}$ and a local $Z = 4.5$ [AadHiggs2012 Table 7] give $\sigma_m \approx \sqrt{2} \times 1.6/4.5 = 0.50\,\mathrm{GeV}$; the statistical error ATLAS quoted for the combined mass, dominated by the diphoton and four-lepton channels, was 0.4 GeV [AadHiggs2012 §9]. Their peak was narrower and more significant, so their statistical error was smaller than our $\pm 0.85\,\mathrm{GeV}$ by almost a factor of two.

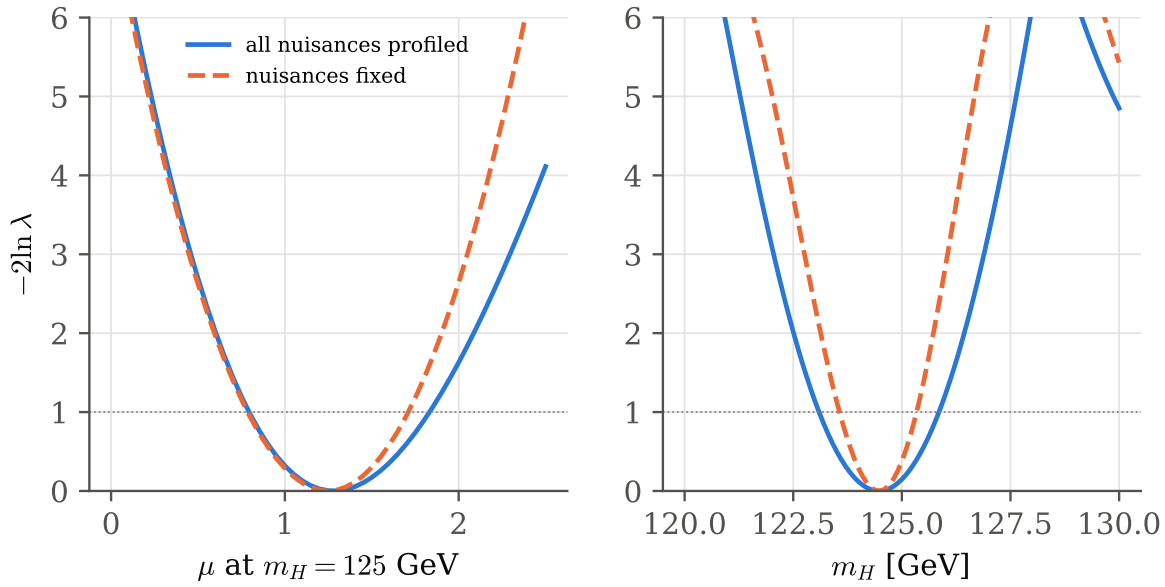


Figure 13.48: Left: $-2\ln\lambda$ as a function of the signal strength at $m_H = 125\,\mathrm{GeV}$, with all nuisance parameters profiled (solid) and fixed (dashed). The crossings of the dotted line at 1 bound the 68% intervals. Right: the same for the Higgs mass. The solid curve is much wider than the dashed one for the mass, because the photon energy scale shifts the peak exactly as a different mass would; for μ the two curves are close, because the measurement is limited by statistics.

13.25.3 Splitting the events into categories

The opening example of Section 13.23 showed that bins with more signal per background carry more information. The same holds for groups of events. In Section 13.22.1 the pairs with both photons central and unconverted had a narrower peak, $\sigma = 1.825\,\mathrm{GeV}$, than the rest, $\sigma = 2.58\,\mathrm{GeV}$. If we fit the two groups as separate spectra, each with its own background, the narrow peak is no longer diluted by the wide one.

Why splitting cannot hurt. *The question:* if two independent spectra share one μ, how do their significances combine? The likelihood is the product of the two (the rule by which Junk combined the LEP search channels [30, (1)]), so $\ln\mathcal{L}$ and q_0 are sums, and (13.53) becomes a sum over categories c and their bins. Take two categories with S_1, S_2 signal and B_1, B_2 background in comparable windows. Split, $Z^2 = S_1^2/B_1 + S_2^2/B_2$; merged, $Z^2 = (S_1 + S_2)^2/(B_1 + B_2)$. Their

difference is

$$\frac{S_1^2}{B_1}+\frac{S_2^2}{B_2}-\frac{(S_1+S_2)^2}{B_1+B_2} \overset{(1)}{=} \frac{S_1^2B_2(B_1+B_2)+S_2^2B_1(B_1+B_2)-(S_1+S_2)^2B_1B_2}{B_1B_2(B_1+B_2)}$$
$$\overset{(2)}{=} \frac{S_1^2B_2^2-2S_1S_2B_1B_2+S_2^2B_1^2}{B_1B_2(B_1+B_2)} \overset{(3)}{=} \frac{(S_1B_2-S_2B_1)^2}{B_1B_2(B_1+B_2)} \geq 0. \tag{13.58}$$

(1) Bring the three terms to the common denominator $B_1B_2(B_1+B_2)$.
(2) Expand: the terms $S_1^2B_1B_2$ and $S_2^2B_1B_2$ cancel against those of $-(S_1^2+2S_1S_2+S_2^2)B_1B_2$.
(3) A perfect square.

Splitting gains exactly when the categories have different signal-to-background ratios, $S_1/B_1 \neq S_2/B_2$, and never loses (as long as the background in each category is determined as well as before, which with fitted backgrounds is only approximately true).

Example: two categories in our data. The central unconverted category holds 8582 events and 58.8 expected Higgs events with $\sigma = 1.825$ GeV; the rest holds 67812 events and 339.1 Higgs events with $\sigma = 2.58$ GeV. Near 125 GeV the fitted background densities are about $\beta_1 \approx 171$ and $\beta_2 \approx 1360$ events per GeV (the signal yields 58.8 and 339.1 divided by β are 0.344 and 0.25 GeV). With (13.54) for each,

$$Z_1^2 = \frac{58.8^2}{2\times 1.772\times 1.825\times 171} = 3.1, \qquad Z_2^2 = \frac{339.1^2}{2\times 1.772\times 2.58\times 1360} = 9.2,$$

so $Z = \sqrt{12.4} = 3.52$ split, against 3.44 merged. The full fit with fitted backgrounds (statistical part only) gives an expected 2.26 split against 2.16 merged, and on the data 2.99 against 2.71, with $\hat{\mu} = 1.31$ against 1.25 (Figure 13.49). The gain is small because only one Higgs event in seven lands in the good category. ATLAS in 2012 used ten categories built from the photon positions, conversions and the transverse momentum of the pair, and Section 13.27 shows what that bought.

13.26 Where can a Higgs boson be excluded?

Before July 2012 the main output of every Higgs search was not a significance but a list of excluded masses. By then the searches at LEP, the Tevatron and the first LHC data had excluded a Standard Model Higgs boson at 95% confidence everywhere below 600 GeV except for some regions between 116 and 127 GeV [AadHiggs2012 §1]. An exclusion answers the opposite question to a discovery. The hypothesis under test is now the signal: *if a Higgs boson with strength μ existed at mass m_H, would data as background-like as ours, or more so, be something we would reasonably expect?* If not, that μ is excluded at that mass.

What we saw in Chapter 4, and what is new. In Section 4.13.4 the question was asked of a single count, with CL_{s+b} and CL_s computed from Poisson sums. In Sections 4.16.4 and 4.16.5 the same ideas were carried to a binned likelihood with nuisance parameters: the statistic $\tilde{q}_\mu$ (4.99), its asymptotic distribution (4.104), the CL_s upper limit (4.111) and the expected band (4.112), all computed on a simulated spectrum. The formulas are the same now. What is new is a scan of 41 masses on real data, where the observed limit can be compared with what the experiment should have achieved, and a check of the asymptotic CL_s limit against toys.

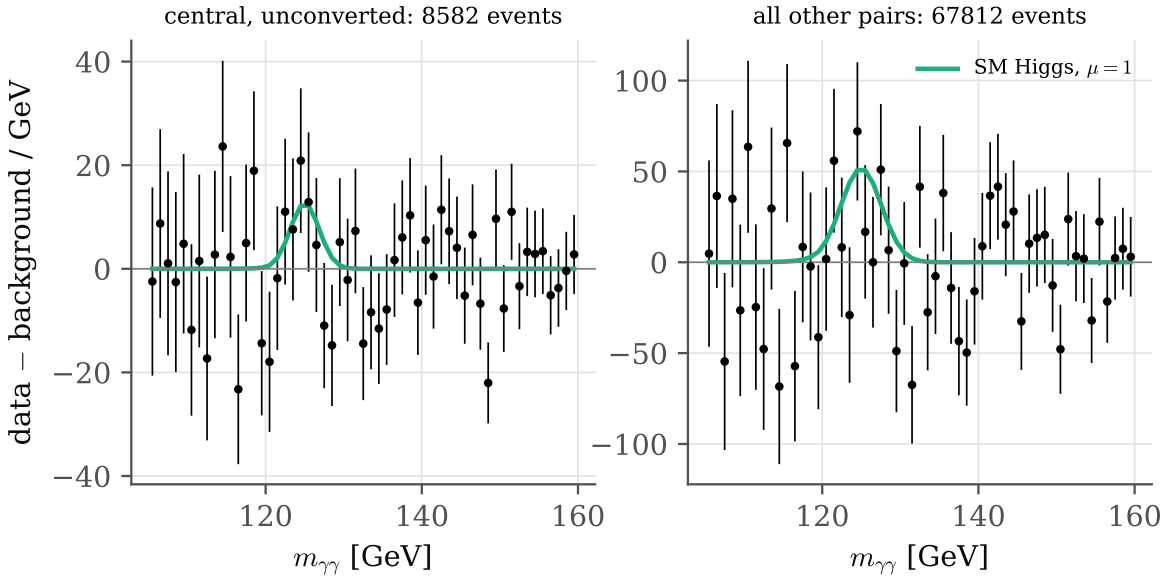


Figure 13.49: The data minus the fitted background in the two categories, with the Standard Model Higgs signal expected in each (orange). The central unconverted pairs (left) are fewer, but their peak is narrower and stands higher above the noise relative to its size; the rest (right) holds most of the signal on a much larger background.

The statistic and the two p-values. At each mass m_H and each trial μ the script computes $\tilde{q}_\mu$ from two fits of (13.51): with μ fixed and with μ free (or at 0 if the free fit gives $\hat{\mu} < 0$). It is large when the data have too few events for a signal of strength μ, and zero when they have more than μ predicts, because an upper limit should not be strengthened by an excess. Two p-values follow from the observed value, one under each hypothesis:

- $\mathrm{CL}_{s+b} = \mathrm{P}(\tilde{q}_\mu \geq \tilde{q}_{\mu,\mathrm{obs}} \,|\, \mu)$: how often a world with the signal gives data at least this background-like;
- $\mathrm{CL}_b = \mathrm{P}(\tilde{q}_\mu \geq \tilde{q}_{\mu,\mathrm{obs}} \,|\, 0)$: how often a world without it does.

Both come from (4.104), with $\mu' = \mu$ and $\mu' = 0$. Excluding μ whenever $\mathrm{CL}_{s+b} < 0.05$ would be a valid test, but it has a flaw that Read described with two experiments [31]: when the background fluctuates down, an experiment can exclude signals so small that it could never have seen them, and an experiment with more background can then quote the stronger limit. The CL_s method divides by CL_b,

$$\mathrm{CL}_s = \frac{\mathrm{CL}_{s+b}}{\mathrm{CL}_b} < \alpha = 0.05, \tag{13.59}$$

which leaves the test unchanged where the two hypotheses predict very different data ($\mathrm{CL}_b \approx 1$) and makes exclusion harder where they predict nearly the same data (CL_b small) [30, 31]. The price is over-coverage: a true signal is excluded in fewer than 5% of experiments, as Figure 4.35 showed for counts.

The one number the formulas need. The asymptotic distributions contain one unknown, the standard deviation σ of $\hat{\mu}$. *The question:* how is it found without simulation? By (4.107), with the background-only Asimov data set ($n_i = b_i$, the true $\mu' = 0$), $\sigma_{\mathrm{A}}^2 = \mu^2/\tilde{q}_{\mu,\mathrm{A}}$: fit the expected background spectrum with μ fixed, see how much worse it fits than with μ free, and read off the width. Because σ depends a little on μ (the nuisance parameters scale with the signal), the

script computes it at $\mu = 0.25, 0.5, 1$ and 2 and interpolates. At 125 GeV, $\sigma_A = 0.491$ at $\mu = 1$; the statistical-only estimate of Section 13.25.1 was 0.46, and the difference is the systematic part. The observed limit μ_{up} is then the μ at which (13.59) becomes an equality, found by root finding.

What should we have excluded? An observed limit means little without the limit the experiment expected. *The question:* if there is no Higgs boson at this mass, what limit would the experiment typically report, and in what range? Without signal, $\hat{\mu} \sim \mathcal{N}(0, \sigma^2)$, and the CL_s limit is an increasing function of $\hat{\mu}$; so its quantiles are the limits at the quantiles of $\hat{\mu}$, which is (4.112):

$$\mu_{up}(N) = \sigma_A \left[N + \Phi^{-1}\left(1 - \alpha\, \Phi(N)\right) \right], \qquad N = -2, -1, 0, 1, 2.$$

The median $N = 0$ gives $\sigma_A \Phi^{-1}(1 - \alpha/2) = 1.96\, \sigma_A$. At 140 GeV, $\sigma_A = 0.477$ and the expected median limit is $0.93 = 1.94\, \sigma_A$, the small difference from 1.96 coming from the change of σ_A with μ.

Example: the $+1\sigma$ edge of the band. For $N = 1$: $\Phi(1) = 0.841$, so $1 - 0.05 \times 0.841 = 0.958$ and $\Phi^{-1}(0.958) = 1.73$. The edge is at $\sigma_A(1 + 1.73) = 2.73\, \sigma_A$, which at 140 GeV is about $2.73 \times 0.477 = 1.30$. Without the division by CL_b the edge would be at $\sigma_A(1 + 1.645) = 2.65\, \sigma_A$: when the data fluctuate up, CL_s and CL_{s+b} differ little. For $N = -2$ the difference is large. $\Phi(-2) = 0.0228$, so $\Phi^{-1}(1 - 0.05 \times 0.0228) = \Phi^{-1}(0.99886) = 3.05$ and the edge is at $1.05\, \sigma_A$, where CL_{s+b} alone would give $(-2 + 1.645)\sigma_A < 0$, an exclusion of every signal including zero. That negative limit is exactly what CL_s was built to forbid.

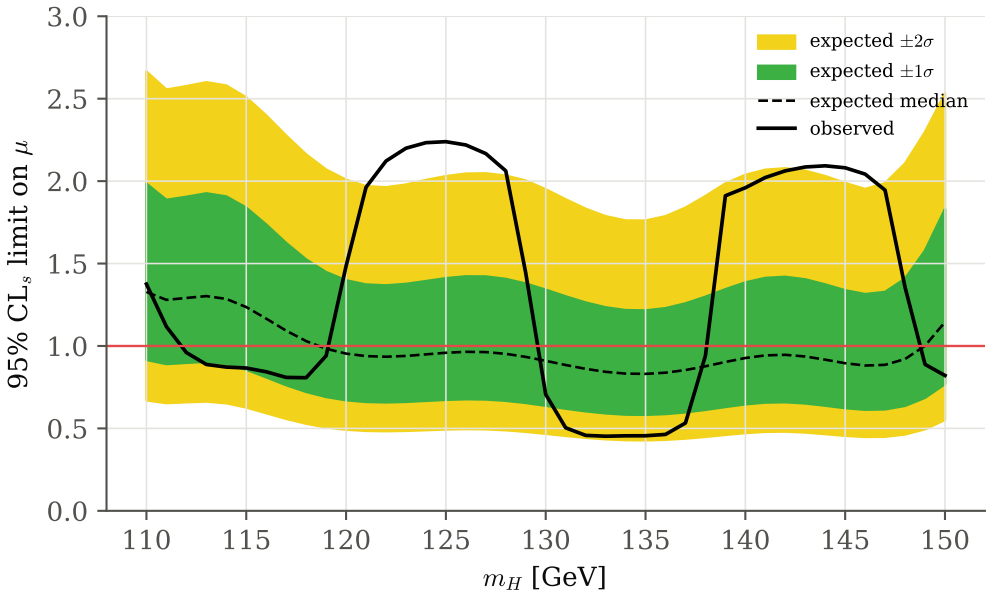


Figure 13.50: The 95% CL_s upper limit on the signal strength μ against the Higgs mass. The dashed line is the median expected without a Higgs boson, the green and yellow bands the ranges expected in 68% and 95% of background-only experiments, and the solid line the observed limit. Where the solid line is below the red line at $\mu = 1$, a Higgs boson with the expected signal is excluded. The observed limit leaves the band upwards near 125 GeV, where the data contain the Higgs boson, and near 143 GeV, where they contain a fluctuation.

The result. Figure 13.50 shows the observed limit and the expected band from 110 to 150 GeV. The median expected limit lies between 0.83 and 1.33, highest at the low end of the range,

where the background is largest, and at the edges of the fitted window, where the polynomial background is most flexible. So with 10 fb^{-1} at 13 TeV the diphoton channel alone expects to exclude a Higgs boson of the Standard Model size over most of the range, but only barely: the median is close to one. The observed limit excludes $\mu = 1$ at 19 of the 41 masses, in the ranges 112–119, 130–138, 149–150 GeV, and leaves a gap around 125 GeV where the observed limit, 2.24, is more than twice the expected 0.96. That gap is the excess of Section 13.23 seen from the other side: a limit cannot exclude a signal that is there. The other gap, near 143 GeV, is the second bump of Section 13.24.2; there the observed limit at 140 GeV is 1.96 against an expected 0.93. A limit curve on real data is read with the p-value scan beside it: the first says what the data rule out, the second what they might contain.

What "$\mu = 1$ at another mass" means here. The simulation exists only at 125 GeV, so at every mass we used the same expected yield, 397.9 events, with the shape scaled as in Section 13.22.1. The true Standard Model yield changes with mass: above about 130 GeV both the production cross section and the diphoton branching ratio fall. Our μ away from 125 GeV is therefore measured relative to a signal of the 125 GeV size, not to the Standard Model at that mass; ATLAS used simulations at many masses [AadHiggs2012 §5.4] and expected to exclude 110–140 GeV in 2012, observing 112–123 and 132–143 GeV [AadHiggs2012 Table 7]. The pattern is the same as ours: an exclusion on both sides of the peak and a hole where the boson is.

13.26.1 Are the asymptotic limits right?

The question: does the CL$_s$ computed from the half-χ^2-type formulas agree with CL$_s$ computed from the actual distribution of $\tilde{q}_\mu$? At 140 GeV, `code/ch13/34_hyy_toys.py` draws 1000 spectra with signal strength μ and 1000 with background only, for each μ on a grid, computes $\tilde{q}_\mu$ for each with the background refitted, and counts the fractions above the observed value: those fractions are CL$_{s+b}$ and CL$_b$ directly [30, (3)–(6)]. (These toys hold the four nuisance parameters fixed, as in Section 13.24; the statistical-only σ_A at 140 GeV is 0.438.)

Figure 13.51 (left) compares the two distributions of $\tilde{q}_\mu$ for $\mu = 0.5$ with the asymptotic curves of (4.104): under the signal hypothesis the half-χ^2 for small values and a shifted Gaussian tail for large ones, under background only a distribution pushed to larger values. The toys follow the curves. The right panel shows CL$_s$ as a function of μ; the toy points scatter about the asymptotic line, and the 95% limits are

$$\mu_{\text{up}}^{\text{toys}} = 1.05, \qquad \mu_{\text{up}}^{\text{asymptotic}} = 1.07.$$

The two agree to within the scatter of the toy points. Both are about half the limit 1.96 that the full likelihood gives at the same mass. That difference is not an error of the formulas but the energy-scale parameter: with θ_m free, a signal hypothesised at 140 GeV can slide towards the excess at 143 GeV (Section 13.23.2), so the data look less background-like for it and its exclusion is weaker. A comparison of toys with asymptotics must use one and the same likelihood, as it does here. With 76 000 events, about 1500 per GeV under the peak, the asymptotic formulas are accurate to the precision a few thousand toys can measure. They would not be for a search with a handful of events in the signal region, which is where the toy method of [30, 31] came from: the LEP Higgs searches, with a few candidates each.

13.27 Reproducing the 2012 diphoton result

The 2012 ATLAS paper has met us before. In Section T4.6.2 its Table 4 gave the number of expected diphoton events, and from $N_S = 110.5$, $\sigma \times \text{BR} = 50$ fb and $L = 5.9$ fb^{-1} we

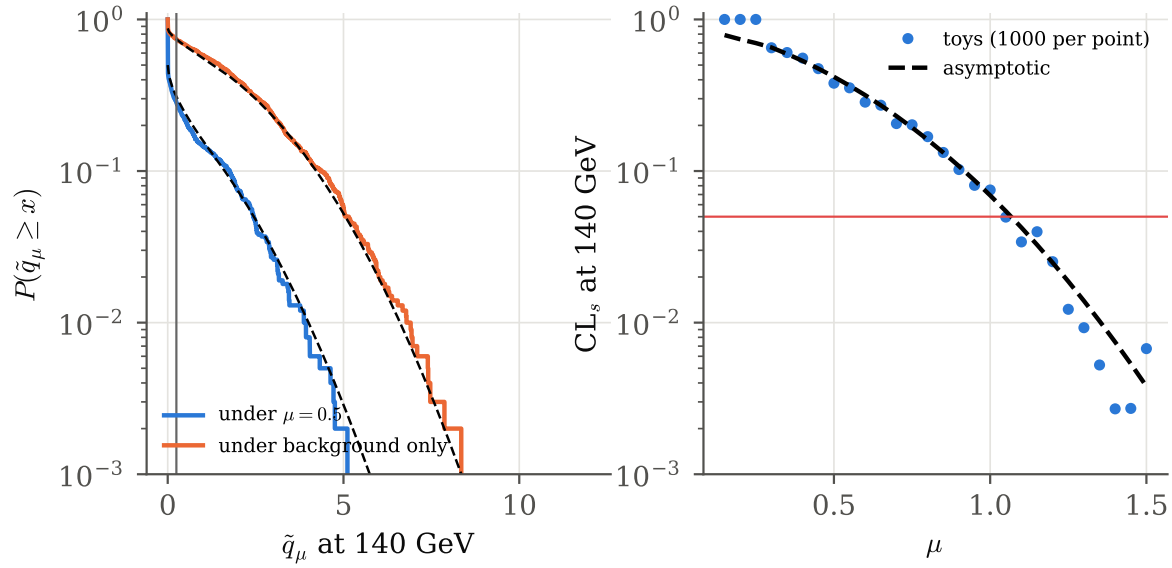


Figure 13.51: Left: the distribution of $\tilde{q}_\mu$ at 140 GeV for $\mu = 0.5$, from toys with that signal (blue) and with background only (orange), against the asymptotic formulas (black dashed); the vertical line is the value in the data. Right: CL_s against μ from the toys (points, 1000 toys per hypothesis and μ) and from the asymptotic formulas (dashed); the limit is where they cross 0.05 (red line).

recomputed a selection efficiency of 0.375 at 8 TeV; Section T4.6.4 went through its local and global significances; and Section T4.6.5 noted that the open-data files store a gluon-fusion $\sigma \times \mathrm{BR}(\gamma\gamma)$ twice as large as the 2012 value. Those were counts of events. The same table now answers a different question, how significant those events were expected to be, and it lets us set the 2012 result and ours side by side ingredient by ingredient. The efficiency below combines both collision energies, so it differs slightly from the 8 TeV value.

13.27.1 Their question

In the first half of 2012 the Higgs boson was the last missing particle of the Standard Model, and only a narrow window of masses near 125 GeV had not yet been excluded (Section 13.26). ATLAS asked whether the diphoton spectrum, combined over the 2011 data at 7 TeV and the first 2012 data at 8 TeV, showed a peak in that window that the background could not explain, and if so how strong it was [118]. Their answer in the diphoton channel alone was a local excess of 4.5σ at 126.5 GeV where 2.5σ was expected, with $\hat{\mu} = 1.8 \pm 0.5$ [AadHiggs2012 Table 7]. Ours, from 10 fb^{-1} of 2016 data at 13 TeV, is 2.74σ at 125 GeV where 2.15 was expected, with $\hat{\mu} = 1.28 \pm 0.52$. *Why did a search with about the same luminosity, at a higher energy where the Higgs boson is produced more than twice as often, expect less?* The answer is in the ingredients, and one formula can compare them.

13.27.2 Their setup

The 2012 analysis used 4.8 fb^{-1} at 7 TeV and 5.9 fb^{-1} at 8 TeV. The photon selection is close to ours (two isolated photons, relative momentum cuts of 0.35 and 0.25), with a fit range of 100–160 GeV. The essential difference is that the events were split into ten categories, by where the photons hit the calorimeter (central, rest, the transition region), whether they converted, the transverse momentum p_{Tt} of the pair relative to its thrust axis, and a category with two forward jets for vector-boson fusion [AadHiggs2012 §5.3]. For each category they published three

numbers per collision energy in their Table 4: the number of data events N_D in 100–160 GeV, the expected number of signal events N_S at $m_H = 126.5$ GeV, and the full width at half maximum of the signal. Their sums are:

	7 TeV	8 TeV	together
luminosity (fb^{-1})	4.8	5.9	10.7
$\sigma \times \mathrm{BR}(H \to \gamma\gamma)$ (fb)	39	50	
data events N_D, 100–160 GeV	23 788	35 251	59039
expected signal N_S	79.6	110.5	190.1
FWHM, inclusive (GeV)	3.9		$\sigma \approx 1.66$

The background in each category was a fourth-order Bernstein polynomial, an exponential of a second-order polynomial or an exponential, chosen by the spurious-signal test of Section 13.22.2 [AadHiggs2012 §5.5].

13.27.3 Their key equation, derived

ATLAS computed their expected significance $E(Z_l)$ from Asimov data sets with the full likelihood. With only Table 4 in hand we cannot redo that, but everything needed for the leading-order expected significance is there. *The question:* what expected significance do the published numbers imply? Three results already derived combine into one formula:

1. a Gaussian peak of S events and width σ on a background of density β per GeV has $q_{0,\mathrm{A}} = S^2/(2\sqrt{\pi}\,\sigma\beta)$, (13.54);
2. independent categories (and independent collision energies) add their $q_{0,\mathrm{A}}$, Section 13.25.3;
3. fitting the background shape keeps a fraction $1 - \rho^2$ of it, (13.49).

So

$$Z^2_{\mathrm{exp}} \approx (1-\rho^2)\sum_c \frac{S_c^2}{2\sqrt{\pi}\,\sigma_c\,\beta_c}, \qquad \sigma_c = \frac{\mathrm{FWHM}_c}{2\sqrt{2\ln 2}}, \tag{13.60}$$

with the sum over the ten categories at each energy, twenty terms in all.

The symbol map. Their N_S is our S_c. Their FWHM becomes our σ_c through the Gaussian relation of Section 13.22.1; we use it for both energies, since Table 4 gives it only at 8 TeV. Their N_D counts events in the whole 60 GeV window, while β_c is the density *under the peak*. *How do we get one from the other?* The spectrum falls roughly exponentially, by a factor of about 3.9 from 100 to 160 GeV in their Fig. 4, so its decay length is $\lambda = 60/\ln 3.9 \approx 44$ GeV. For an exponential the ratio of the density at the peak, 26.5 GeV into the window, to the mean density is

$$\frac{\beta(126.5)}{N_D/60} \overset{(1)}{=} \frac{e^{-26.5/\lambda}}{\frac{1}{60}\int_0^{60} e^{-x/\lambda}\mathrm{d}x} \overset{(2)}{=} \frac{60\,e^{-26.5/\lambda}}{\lambda\,(1-e^{-60/\lambda})} \overset{(3)}{=} \frac{60\times 0.548}{44\times 0.744} = 1.00. \tag{13.61}$$

(1) Density at $x = 26.5$ GeV over the mean density of the window, both for $e^{-x/\lambda}$; the normalisation cancels.
(2) The integral of an exponential.
(3) $e^{-26.5/44} = 0.548$ and $1 - e^{-60/44} = 1 - 1/3.9 = 0.744$.

The peak happens to sit where the density equals the window's mean, so $\beta_c = N_{D,c}/60$ GeV to within a per cent. Finally their $E(Z_l)$ is our $Z_{\mathrm{exp}} = \sqrt{q_{0,\mathrm{A}}}$, and their μ is ours.

The background factor. The last ingredient, $1-\rho^2$, depends on the background function and the window. The script computes it from the Fisher matrix (13.49) for a Gaussian peak of the inclusive width on an exponential background: for their fourth-order Bernstein polynomial over 60 GeV it is 0.7, and for our fifth-order one over 55 GeV with our wider peak, 0.41. A narrower peak is harder for a polynomial to imitate, so it loses less.

13.27.4 What we compute

`code/ch13/35_hyy_atlas2012.py` types in Table 4, evaluates each of the twenty terms of (13.60) and the inclusive version (one category with the totals), and evaluates the same formula on our spectrum with $S = 397.9$, the FWHM of our fitted Crystal Ball, 5.81 GeV, and $\beta = 1527$ per GeV. It also reads the cross section times branching ratio stored in our four simulation files, to separate the signal per fb^{-1} into "how many Higgs bosons decayed to two photons" and "what fraction of them we selected".

13.27.5 Side by side

	ATLAS 2012 (7 + 8 TeV)	open data (13 TeV)
luminosity (fb^{-1})	10.7	10.06
$\sigma \times \mathrm{BR}(H \to \gamma\gamma)$ (fb)	39, 50	114
Higgs bosons to $\gamma\gamma$ produced	482	1149
selection efficiency	0.39	0.35
expected signal S per fb^{-1}	17.8	39.5
background under the peak, per GeV per fb^{-1}	92	152
mass resolution σ (GeV)	1.66	2.467
categories	10 per energy	1
$1-\rho^2$ of the background fit	0.7	0.41
Z_{exp}, formula, inclusive, $1-\rho^2=1$	2.5	3.44
Z_{exp}, formula, categories, $1-\rho^2=1$	2.93	
Z_{exp}, formula with $1-\rho^2$	2.45	2.21
Z_{exp}, full Asimov fit	2.5 (quoted)	2.15
Z_{obs} (local)	4.5	2.74
$\hat{\mu}$	1.8 ± 0.5	1.28 ± 0.52

The expected significance. From nothing but Table 4, (13.60) gives 2.45 for ATLAS against the 2.5 they quoted, and 2.21 for us against our full Asimov value 2.15. The leading-order formula with the background factor captures what the full likelihood does to within a few per cent. The comparison of the two "inclusive" entries, 2.5 against 2.93, shows what the ten categories were worth to ATLAS; for us, two categories gained only from 2.16 to 2.26 (Section 13.25.3).

Growth with luminosity. At a fixed analysis, S and β both grow in proportion to the luminosity L, so by (13.54) $Z_{\mathrm{exp}} \propto S/\sqrt{\beta} \propto \sqrt{L}$. A 5σ expected significance needs $L(5/Z_{\mathrm{exp}})^2$: 43 fb^{-1} with the 2012 diphoton analysis, 54 fb^{-1} with ours. Figure 13.52 shows both lines with the two observations. ATLAS did not wait for 43 fb^{-1}: the discovery came from combining the diphoton excess with the four-lepton and WW channels, 5.9σ locally from 4.9σ expected [AadHiggs2012 Table 7], and from an upward fluctuation that made the diphoton excess 1.6 standard deviations larger than the Standard Model predicted.

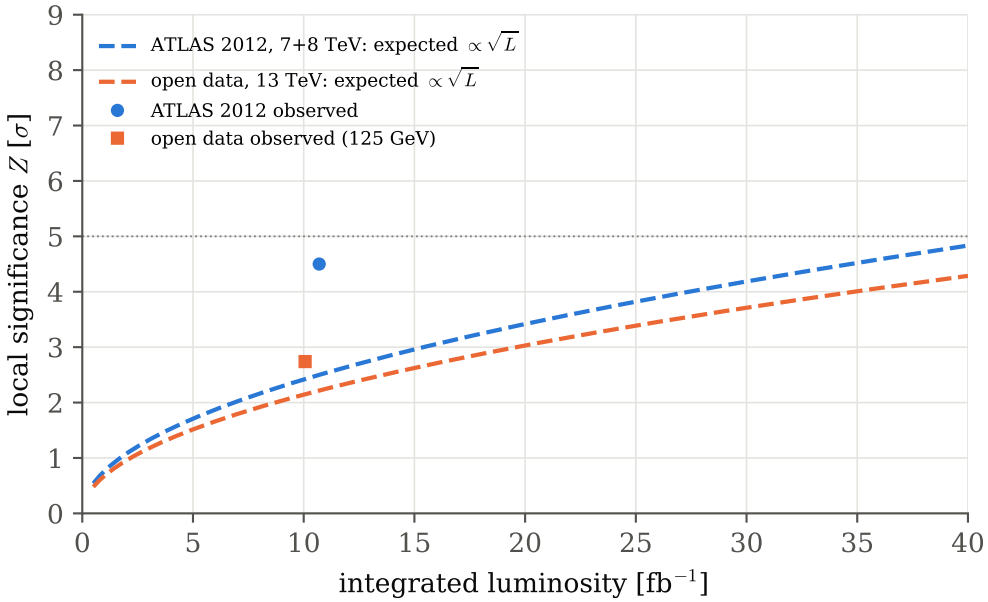


Figure 13.52: The expected local significance of the diphoton search against integrated luminosity, for the 2012 ATLAS analysis (blue dashed, scaled from its quoted 2.5σ at 10.7 fb^{-1}) and for ours (orange dashed, from 2.15σ at 10.06 fb^{-1}), both growing as $\sqrt{L}$. The points are the observed local significances. Both observations lie above their expectation: in 2012 by two units of Z, in our data by about half a unit.

13.27.6 Why they differ

Each ingredient of (13.60) can be compared separately, and the formula says how much each one matters: $Z \propto \sqrt{1-\rho^2}\, S/\sqrt{\sigma\beta}$ for one category. Compared at equal luminosity, ours over theirs, one factor at a time:

- *More Higgs bosons per* fb^{-1}*: a factor* $39.5/17.8 = 2.22$. At 13 TeV the cross section times branching ratio of all four production modes together is 114 fb, 2.5 times that at 7–8 TeV (gluon fusion alone is 102 fb, Section T4.6.5), because at a higher energy a Higgs boson of fixed mass needs gluons carrying a smaller fraction of the proton's momentum, and such gluons are more numerous (Chapter T4). Our selection keeps a fraction 0.35 of the produced diphoton decays against ATLAS's 0.39 (they quote about 40% [AadHiggs2012 §5.4]), which takes back a little.
- *More background per* fb^{-1}*: a factor* $(152/92)^{-1/2} = 0.78$. The diphoton continuum grows with energy as well, and the open-data photons pass a looser isolation: the calorimeter energy is summed in a cone of $\Delta R < 0.2$ and compared with 6.5% of p_T, where the 2012 analysis used a cone of 0.4 and a fixed 4 GeV [AadHiggs2012 §5.1], and a smaller cone lets more jets that fake photons through.
- *A wider peak: a factor* $(2.467/1.66)^{-1/2} = 0.82$. Our width is what the released simulation gives; the release does not document the photon energy calibration in enough detail to trace the difference further.
- *No categories: a factor* $2.5/2.93 = 0.85$. The ten categories raised ATLAS's leading-order Z by separating the few events with a narrow peak and little background from the many with a wide peak and much background, exactly as (13.58) promises.
- *A more flexible background: a factor* $\sqrt{0.41/0.7} = 0.77$. A degree-five polynomial over 55 GeV

against a degree-four one over 60 GeV, and a wider peak that the polynomial imitates more easily (Section 13.22.2).

The product is $2.22 \times 0.78 \times 0.82 \times 0.85 \times 0.77 = 0.93$. The two predictions of (13.60) per unit of luminosity, $2.21/\sqrt{10.06} = 0.70$ against $2.45/\sqrt{10.7} = 0.75$, are in exactly that ratio, and so, to the precision of the formula, are the quoted expectations. The budget explains the opening puzzle: the higher energy more than doubled the signal, and four smaller losses, each between 15 and 25%, took all of it back. The two largest, the background function and the background density, are the cheapest to recover: categories with less background and a narrower peak would improve both.

The *observed* values differ for a different reason. ATLAS's diphoton excess was a 1.6σ upward fluctuation of the signal on top of an expected 2.5σ; ours is a fluctuation of about half a standard deviation.

The lesson. Reading the paper gives one number, 2.5σ expected; rebuilding it shows that the number is the product of five ingredients, each of which can be measured separately and traded against the others: signal per luminosity, background density under the peak, mass resolution, how finely the events are sorted by their quality, and how flexible the background function has to be. A resonance search is designed by improving the weakest of these, and the formula $Z^2 \approx (1-\rho^2)\sum_c S_c^2/(2\sqrt{\pi}\sigma_c\beta_c)$ tells which one that is before any data are taken.

13.28 The full analysis, and what our shortcuts cost

Every number of this search came from a laptop-sized computation: 1.9 GB of files read once, a few thousand toy spectra fitted by Newton's method, and each script ran within minutes on four cores. The statistics is the same as in the collaboration's analysis, but several pieces were simplified. Each simplification has a cost we can name, and for most of them the cost was measured along the way.

	full analysis	ours, and the cost
events	all channels, full photon calibration, many times more data in later years	10 fb^{-1} of released data, diphoton only; $Z \propto \sqrt{L}$ (Figure 13.52)
categories	ten in 2012, more later, each with its own background function	one (two as an example); the ten categories were worth 2.5 → 2.93 in 2012
signal shape	Crystal Ball plus Gaussian, simulated at many masses	Crystal Ball at 125 GeV, scaled to other masses; μ away from 125 GeV relative to a fixed yield
background choice	spurious-signal test on large simulated background samples	truths fitted to the same data (Section 13.22.2); the test only compares families
nuisance parameters	dozens: each efficiency, energy scale and resolution term, theory	four, with sizes taken from 2012 and from the release
toys	full fits with all nuisance parameters, auxiliary measurements redrawn	statistical part only, linear fits; nuisances changed Z by less than 0.01 here
global significance	Gross–Vitells in mass (and width), combined channels	Gross–Vitells in mass, checked with 4000 toys

Verification and research. Two kinds of simulation appeared in this search, and they did different jobs. The toys of Sections 13.24.1 and 13.26.1 *verified* known analytic results: the half-χ^2 law of q_0, the Asimov median, the asymptotic CL_s. With 76 000 events the formulas passed, so the toys added confidence but no new number. The background-only scans of Section 13.24.2 were the *research instrument*: no closed formula gives the global p-value of a scan over a fitted, bending background, and the Gross–Vitells formula itself needs the upcrossing count that only simulated scans provide. The same division runs through every capstone of this chapter: simulation checks a formula where one exists and replaces it where none does.

Recap

- From the released 13 TeV files, the HyyAnalysis selection keeps 76394 photon pairs in 105–160 GeV, where a Standard Model Higgs boson would add 397.9 events; the invariant mass $2p_{T1}p_{T2}[\cosh\Delta\eta-\cos\Delta\phi]$ is where it would appear as a peak.
- The signal is a Crystal Ball of width 2.467 GeV fitted to weighted simulation; the background a fifth-order Bernstein polynomial, the least flexible family whose spurious signal stays below 20% of the statistical error. Flexibility costs a fraction ρ^2 of $q_{0,\mathrm{A}}$, here more than half.
- The expected significance of a peak is $Z^2 \approx (1-\rho^2)\sum_c S_c^2/(2\sqrt{\pi}\sigma_c\beta_c)$, and its mass precision $\sigma_m = \sqrt{2}\,\sigma/Z$; splitting into categories never loses, by $(S_1B_2-S_2B_1)^2/(B_1B_2(B_1+B_2))$.
- The data show a local 2.77σ excess at 124.5 GeV where 2.15σ was expected at 125 GeV; toys confirm the asymptotic laws, and the global significance over 110–150 GeV is about 1.54σ, with a trials factor of 22.2 and a background bump of similar local size at 143 GeV.
- $\hat{\mu} = 1.28 \pm 0.46\,(\mathrm{stat}) \pm 0.24\,(\mathrm{syst})$ and $\hat{m}_H = 124.5 \pm 1.4$ GeV, the mass error dominated by the photon energy scale.
- CL_s excludes a signal of the Standard Model size at 112–119, 130–138, 149–150 GeV and not near 125 GeV, where the observed limit is 2.24 against an expected 0.96.
- Rebuilt from its Table 4, the 2012 ATLAS expected 2.5σ follows from the same formula; our larger signal per fb^{-1} is outweighed by more background, a wider peak, the absence of categories and a more flexible background function.

13.29 A small twist in the polarization of the sky

In Section 13.1 we took the SMICA temperature map and asked how strong its fluctuations are on each angular scale. The pipeline we built for that question handles a great deal: the mask that hides our Galaxy, the beam and pixels that blur the map, the noise, which drops out when the two half-mission maps are cross-correlated, and simulated skies that tell us how much every number would scatter. All of it was aimed at a spectrum whose *size* we wanted to know, and whose value is large: thousands of $\mu\mathrm{K}^2$ at the acoustic peaks.

Now the question is different. The same SMICA files also contain the polarization of the sky, and in the polarization there is a spectrum that the standard model says must vanish on average. If something turned the polarization of every CMB photon by the same small angle on its way to us, that spectrum would no longer vanish: it would become a faint copy of a spectrum we know well, about one per cent of it for an angle of a third of a degree. We are no longer measuring how big something is. We are looking for a faint copy of a known pattern in a place where nothing should be, and asking how much of that pattern is there.

A picture on the wall. Picture a wall covered in wallpaper with vertical stripes, some bold and some faint, and a photograph of it taken with the camera turned slightly. In the photograph every stripe leans a little to one side. The lean is obvious where the stripes are bold and invisible where the wall is plain, because a plain wall has nothing to lean. So the way to measure the tilt is to compare two pictures: the pattern of "lean" and the pattern of "stripe". If the first is a scaled copy of the second, the scale factor is the tilt. Notice also what the photograph cannot tell us. A crooked camera and crooked wallpaper give the same photograph. Only something in the picture that we know was hung straight, a door frame say, can tell the two apart: the door frame leans only if the camera is crooked.

The same picture on the sky. Recall from Chapter T1 the physics in one breath. CMB photons are linearly polarized, and at each point of the sky the polarization is a short headless stick (Section T3.3), described by the Stokes parameters Q and U with $Q + iU = Pe^{2i\psi}$, where ψ is the angle of the stick. The pattern of sticks splits into two kinds. **E modes** are patterns that look the same in a mirror (sticks lined up along or across the direction in which the polarization changes, like the stripes). **B modes** are patterns with a handedness (sticks at 45° to that direction, like swirls). The physics of the early Universe, as far as we know, does not prefer left to right, so on average there is no correlation between E and B, or between temperature and B: $C_\ell^{EB} = C_\ell^{TB} = 0$. Now suppose every stick is turned by the same angle β between the last scattering surface and us, as a slowly varying pseudoscalar field coupled to photons would do [62]. That is our crooked wallpaper. Some of every E pattern becomes B, and Chapter T1 derives what happens to the spectra:

$$C_\ell^{EB,o} = \tfrac{1}{2}\sin(4\beta)\,\big(C_\ell^{EE} - C_\ell^{BB}\big), \qquad C_\ell^{TB,o} = \sin(2\beta)\,C_\ell^{TE}, \tag{13.62}$$

where "o" means observed and the spectra on the right are those before the rotation. The crooked camera is a detector whose polarization directions are all turned by an angle α from where we think they are: it turns every stick in the map, CMB or not, and its effect on (13.62) is identical, with β replaced by $\alpha + \beta$. The door frame is the polarized emission of our own Galaxy, which is turned by α but not by β, because it has not travelled across the Universe [76].

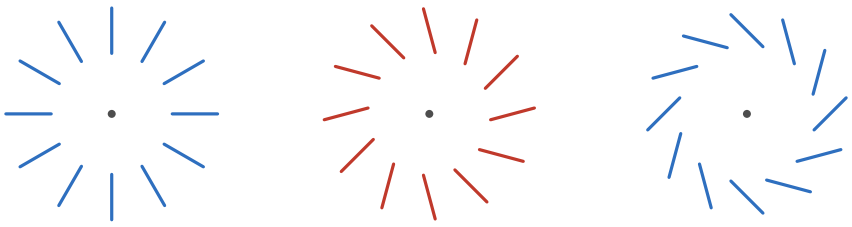


Figure 13.53: Polarization sticks on a ring around a point. Left: a pattern of E type, sticks pointing away from the centre, which looks the same in a mirror. Right: a pattern of B type, every stick at 45°, which has a handedness. Middle: the E pattern with every stick turned by the same angle (exaggerated here to 15°): it has become mostly E with a little B, the fraction $\sin 2\beta$ of the amplitude that (13.62) turns into the EB correlation.

The story in symbols. Take the smallest case: one multipole ℓ, a sky with E and no B, and a rotation by β. For each m the rotation mixes the two coefficients as $E' = \cos 2\beta\, E - \sin 2\beta\, B$ and $B' = \sin 2\beta\, E + \cos 2\beta\, B$ (Chapter T1), so with $B = 0$,

$$\big\langle E'B'^{*}\big\rangle \overset{(1)}{=} \cos 2\beta\, \sin 2\beta\, \big\langle |E|^2 \big\rangle \overset{(2)}{=} \tfrac{1}{2}\sin 4\beta\, C_\ell^{EE} \;\approx\; 2\beta\, C_\ell^{EE}. \tag{13.63}$$

(1) Substitute $E' = \cos 2\beta\, E$ and $B' = \sin 2\beta\, E$ (the terms in B vanish) and take the mean.

(2) $2\sin x\cos x = \sin 2x$ with $x = 2\beta$, and $\langle|E|^2\rangle = C_\ell^{EE}$; for a small angle $\sin 4\beta \approx 4\beta$.

The numbers make the difficulty plain. At $\ell = 1000$ the fiducial E spectrum is $D_\ell^{EE} = 42.1\,\mu\mathrm{K}^2$ (with $D_\ell = \ell(\ell+1)C_\ell/2\pi$ as always), so an angle of 0.3° makes $D_\ell^{EB} = 0.44\,\mu\mathrm{K}^2$, about one per cent of it. The dial is the angle, and the obvious estimator is the ratio at one multipole, $\hat\beta_\ell = \widehat{C}_\ell^{EB}/2\widehat{C}_\ell^{EE}$. It is hopeless. At $\ell = 1000$ the noise power of one SMICA half-mission map in polarization is about 4 times the E signal, the multipole has about 1766 independent modes on the kept sky, and the scatter of that single ratio is about 3° (the estimate is made in `44_angle.py` from the variance of a product of two Gaussian coefficients, the formula of Section 13.31). The good choice is the wallpaper's: compare the whole pattern of $\widehat{C}_\ell^{EB}$ across a thousand and more multipoles with the whole pattern of $\widehat{C}_\ell^{EE} - \widehat{C}_\ell^{BB}$, giving most weight where E is strong and the noise is weak. That buys a factor of order $\sqrt{1000}$ and more, and its price is that we need the covariance of the $\widehat{C}_\ell^{EB}$ across multipoles, which only simulated skies supply. Lesson: a tiny effect that is a known pattern times one number is measured by fitting the pattern, never by reading one point.

A picture: a projection onto a template. Write the measured EB spectrum, bin by bin, as one long list $\boldsymbol{e}$, and the pattern it should copy, $C^{EE} - C^{BB}$, as a second list $\boldsymbol{D}$. If the sky is rotated, $\boldsymbol{e} \approx 2\beta\,\boldsymbol{D}$ plus noise. Estimating β is asking how much of the direction $\boldsymbol{D}$ is in $\boldsymbol{e}$: a projection, with the noise deciding which components of the projection to trust. The same estimator wears many names. It is the linear least-squares fit of one amplitude (Section 3.10.2); it is the matched filter that finds a gravitational-wave chirp of known shape in noisy data (Section 10.8.3); it is template fitting wherever a signal's shape is known and its strength is not. Linear algebra is the natural language because spectra, templates and weights are lists of numbers and the estimator is a weighted dot product.

The route is then clear. The polarization maps are brought to laptop size and their spectra computed through the SMICA confidence mask; the mask turns out to be unable to fake a rotation; the estimator is tested on six hundred simulated skies in which a known angle was hidden, then applied to the real sky and compared with the published measurements. The same spectra finally answer a second question: do they show the one B-mode signal that the standard model guarantees, the lensing of E into B?

The event. The data are the SMICA Q and U maps of the 2018 release [19, 186], from the same files as the temperature map of Section 13.1.1, in their two half-mission versions. Our event is a number computed from them: the angle $\hat\beta$ that best turns the measured $\widehat{C}_\ell^{EE} - \widehat{C}_\ell^{BB}$ and $\widehat{C}_\ell^{TE}$ into the measured $\widehat{C}_\ell^{EB}$ and $\widehat{C}_\ell^{TB}$.

What could have produced a nonzero angle? Before computing it, we list the situations that could make $\hat\beta$ differ from zero, one at a time, each with its reason.

1. *A rotation β of the CMB polarization in flight.* This is the effect we are after. It produces exactly the pattern (13.62).

2. *A miscalibration α of the detectors' polarization angles.* Each *Planck* polarimeter's orientation on the sky was measured on the ground, to about 0.3° [214]. A common offset turns every stick in the map, so it produces the same pattern, with $\alpha + \beta$ in place of β. SMICA combines many detectors and frequencies with fixed weights, so the map carries one effective α, a weighted average of theirs.

3. *Polarized dust in our Galaxy with an EB correlation of its own.* If the dust filaments are slightly twisted relative to the magnetic field that aligns them, the dust has $C^{EB} \neq 0$ [62, 72]. SMICA removes most of the dust, but not all of it near the plane.
4. *Leakage of temperature into polarization by the beams.* An asymmetric beam turns some of the bright temperature map into spurious polarization, which can correlate with T and E.
5. *Leakage of E into B by the mask.* Cutting the sky mixes E and B. It adds scatter but cannot produce the pattern (13.62) on average.
6. *Chance.* With no rotation at all, the noise and the CMB itself make $\widehat{C}_\ell^{EB}$ scatter around zero, and some realisations lean one way. This is the explanation whose sampling distribution we can produce, by simulating skies.

The first two cannot be told apart with one CMB map, whatever we do: they make the same pattern from the same sky. The third is tested by changing how much of the sky near the Galactic plane we keep; the fifth and sixth by simulations; the fourth we can only bound by quoting the published analyses.

The questions, in words.

(a) What are the EE, BB, EB, TE and TB cross-spectra of the two half-mission maps through the SMICA polarization mask, and can the mask fake a rotation?

(b) If the sky were not rotated, how much would an estimate of the angle scatter? If it were rotated by a known angle, would the estimator find it, on average, and would its error bar cover the truth as often as it claims?

(c) Which angle does the real sky prefer, and would an angle that large, or larger, be something we would reasonably expect if the true angle were zero?

(d) How does it compare with the published measurements, and what exactly has been measured?

(e) Do the same spectra show the B modes that lensing must make?

The inputs, and what would break them. Two inputs are *physical*. (i) Outside the mask, the polarization is a Gaussian CMB with the fiducial spectra, possibly rotated by one angle, plus noise; dust residuals would break this, and would show up as an angle that depends on the mask. (ii) The noise of the two halves is independent, so the cross-spectra carry no noise bias. Two are *observational*: (iii) the beam is the same for E and B and for every direction on the sky, so it multiplies each coefficient by a number and cannot rotate anything; (iv) the angle is the same everywhere on the sky. One is *mathematical*: (v) the angle is small, so $\sin 4\beta \approx 4\beta$; at $\beta = 1^\circ$ the error of this approximation is $1 - \sin(4^\circ)/(4^\circ \cdot \pi/180) \approx 8 \times 10^{-4}$ of the angle, far below anything we can measure.

13.29.1 The polarization maps at laptop size

The SMICA files of Table 13.1 hold, besides the temperature, the Stokes Q and U maps (columns `Q_STOKES` and `U_STOKES`) and two confidence masks, one for temperature and one for polarization (`PMASK`, field 4 of the full-mission file). SMICA builds its polarization map the way it builds the temperature map: it combines the frequency maps multipole by multipole, separately for E and for B, with weights chosen so that the CMB passes and the Galaxy is suppressed [19, §3]. The confidence mask marks the pixels where that cleaning is trusted. It is

generous: it keeps a fraction 0.929 of the sky, far more than the temperature common mask of Section 13.2.

Noise first. The polarization maps differ from the temperature map in one way that shapes everything below. Outside the mask, the root mean square of Q and U in a pixel of the full-mission map is 28.6 μK, and 41.7 μK in a half-mission map, while the CMB polarization itself has a root mean square of a few μK. Pixel by pixel, the maps are noise. The CMB polarization is there only statistically, in the spectra, where thousands of modes are averaged. That is why Figure 13.54a looks like static and why the polarized intensity becomes structured only after smoothing over a degree (Figure 13.54b), where the noise averages down.

The degrade. We bring the maps to $N_{\mathrm{side}} = 512$ exactly as in (13.1), with one change: Q and U are transformed together with the spin-2 harmonic transform, which turns them into the E and B coefficients, and the pixel window used is the polarization one. A HEALPix pixel averages Q and U over its area in the local frame of each point, and the window this produces for E and B differs slightly from the temperature window; `hp.pixwin(..., pol=True)` returns both. At $\ell = 1000$ the polarization window of $N_{\mathrm{side}} = 512$ is 0.835 and the beam is 0.826; the beam supplied for polarization (`POL_BEAM`) is the same 5′ effective beam as for temperature.

Filling the holes with zeros. In Section 13.1.1 the holes of the temperature map were filled with the inpainted CMB before the transform, because the Galactic plane and the point sources in the temperature map are thousands of μK bright and the band limit would ring them out over the kept sky. In polarization the masked pixels are not much brighter than the kept ones: the largest polarized intensity inside the holes of the full map is 356 μK, against 306 μK outside, both set by noise spikes. So the danger of ringing is small, and we simply set $Q = U = 0$ in the masked pixels at full resolution before the transform. That guarantees that nothing from the regions SMICA does not trust can be carried into the kept sky. The mask is degraded as in Section 13.1.1: a pixel at $N_{\mathrm{side}} = 512$ is kept only if all sixteen of its small pixels are kept, which leaves a fraction 0.925.

Listing 13.10: The spin-2 degrade: zero the masked pixels, transform Q and U to E and B, swap the polarization pixel windows, synthesise (`code/ch13/41_pol_data.py`)

```
fl[2:] = pw512[2:] / pw2048[2:]
out = {}
for kind in ("full", "hm1", "hm2"):
    src = lp.raw_path(kind)
    q = hp.read_map(src, field=1, dtype=np.float64) * 1e6
    u = hp.read_map(src, field=2, dtype=np.float64) * 1e6
    p = np.hypot(q, u)
    tag = {"full": "Full", "hm1": "HmOne", "hm2": "HmTwo"}[kind]
    nums[f"ThirteenEPmaxIn{tag}"] = f"{p[hole].max():.0f}"
    nums[f"ThirteenEPmaxOut{tag}"] = f"{p[~hole].max():.0f}"
    nums[f"ThirteenEPrmsOut{tag}"] = f"{np.sqrt(np.mean(p[~hole] ** 2) / 2):.1f}"
    q[hole] = 0.0
    u[hole] = 0.0
    alm = hp.map2alm([np.zeros_like(q), q, u], lmax=L, iter=1, pol=True)
    del q, u, p
    aE, aB = hp.almxfl(alm[1], fl), hp.almxfl(alm[2], fl)
    _, Q, U = hp.alm2map([alm[0] * 0, aE, aB], N, lmax=L, pol=True)
    out[f"{kind}_Q"], out[f"{kind}_U"] = Q.astype(np.float32), U.astype(np.float32)
    print(f"{kind}: done ({time.time() - t0:.0f} s)", flush=True)
```

For each map the loop converts Q and U from K to μK, records the largest polarized intensity inside and outside the mask, zeroes the masked pixels, and calls `hp.map2alm` with `pol=True`,

which returns the T, E and B coefficients (the temperature slot is filled with zeros because the temperature maps of Section 13.1.1 are reused). `hp.almxfl` multiplies E and B by the ratio of polarization pixel windows, and `hp.alm2map` with `pol=True` synthesises Q and U at $N_{\mathrm{side}} = 512$. The three maps took 65 s on the cluster and are cached in `data/ch13/`.

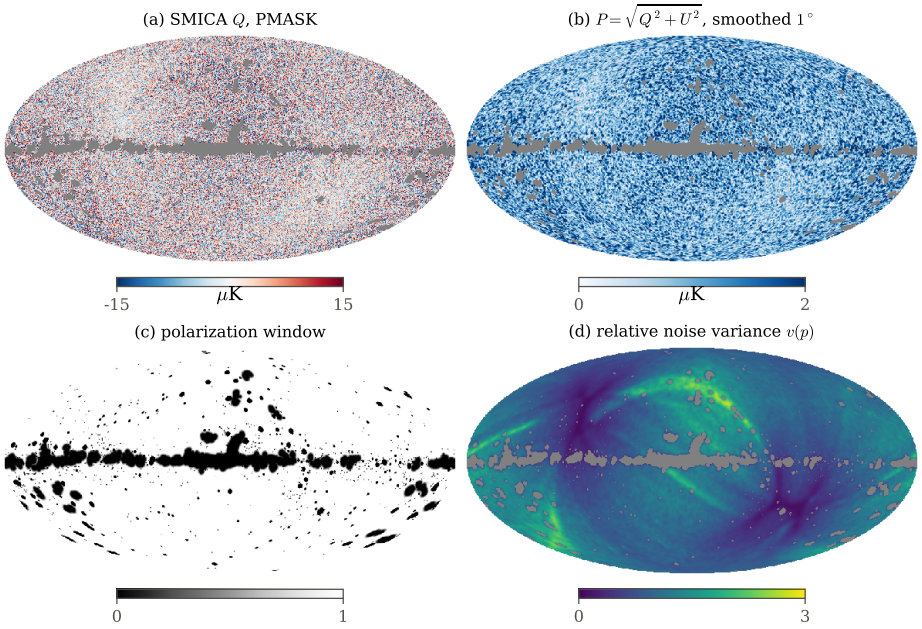


Figure 13.54: (a) The SMICA Q map at $N_{\mathrm{side}} = 512$, outside the polarization confidence mask (grey): at the pixel level it is dominated by noise. (b) The polarized intensity after smoothing Q and U over 1°, where the noise has averaged down and the CMB polarization becomes visible. (c) The window of the analysis, the confidence mask with tapered edges (Section 13.30.1). (d) How the noise variance varies over the sky relative to its mean, measured from the half-difference map (Section 13.30.4): the scanning strategy of *Planck* leaves the regions near the ecliptic poles less noisy.

13.30 Polarization spectra through a mask

For the temperature map of Section 13.3 the route from a masked map to a spectrum had three steps: multiply by a tapered window, transform, and undo on average what the window did, with the MASTER coupling matrix. For polarisation, the window also mixes E and B modes. This matters because the angle we want to measure itself converts E into B. The toy skies of Section 10.17 let us distinguish these effects by examining what masking does to the cross-spectrum used to estimate the angle.

13.30.1 The window

The degraded confidence mask has 1009 holes, of which 25 are larger than 20 deg^2; we split them as in Section 13.2.1 and taper the Galactic holes over 2° and the small ones over 0.5° with the cosine profile (13.2) (Figure 13.54c). The window keeps $f_{\mathrm{sky}} = 0.925$ of the sky; the mean of W^2 is 0.873 and the effective sky fraction that counts modes, $\langle W^2\rangle^2/\langle W^4\rangle$ (Section 8.9), is 0.882. The temperature spectra TE and TB use the temperature window of Section 13.2.2 for T and this window for E and B; nothing below requires the two windows to be equal.

For the Galactic check we also build a *strict* window, the product of the polarization window and the temperature window. It removes in addition everything the temperature common mask

removes, a broader band around the plane and many more source holes, and keeps $f_{\rm sky} = 0.757$ ($\langle W^2\rangle^2/\langle W^4\rangle = 0.683$). Dust residuals, if they matter, are stronger in the main window.

13.30.2 Why a mask turns E into B

The toy skies of Section 10.17 met this problem first. Whether the sticks around a point form a spoke-and-ring pattern (E) or a swirl (B) is not a property of one pixel; it depends on how the sticks change across a neighbourhood, so cutting the map changes the answer near every edge. The picture there was a field of wheat combed in neat rows and photographed through a window frame: a row that stops abruptly at the frame is, for anyone who describes the photograph by its waves, a bundle of waves pointing in slightly different directions, and for those directions part of the combed pattern reads as a swirl, although nothing in the field has changed. Cutting one E wave with a straight edge gave the size of the effect (Eq. (10.105)): a sharp edge reads as B a strip about one wavelength wide along the cut, a smooth taper spreads the cut over its own width and leaks far less at wavelengths shorter than that width, and in neither case is the leaked B correlated with E on average. On a band-shaped mask the simulated skies there showed exactly this: the 5° taper pushed the leak well below the lensing B at $\ell \approx 300$, while the sharp band still leaked a sizeable fraction of it (Figure 10.46). The B created this way is the **E-to-B leakage**. Because E is a hundred times stronger than the lensing B in the sky, even a small leaked fraction can be larger than the real B. What the band mask cannot tell us is how much leaks through the SMICA window, with its narrow Galactic cut and hundreds of small holes; that we measure.

Example: how much leaks? The simulated skies of Section 13.31.2 let us measure it, because there we can observe the E part of a sky alone, with no B and no noise, and look at the B that comes out. Through the *binary* confidence mask the leaked B power is, in the median bin, 0.7% of the E power; through the tapered window it is 0.43% (Figure 13.55). The taper helps here by only a factor of about 2, much less than for the temperature coupling of Section 8.10.4, and at the lowest multipoles not at all (in the first bin, $50 \le \ell < 80$, the two are 2.2% and 2.41%). The reason is the geometry of this mask: it removes only a narrow Galactic band and a few hundred small holes, so most of the leakage comes from long edges and from the large-scale shape of the cut, which a taper of 2° does not smooth. The cut-wave estimate of Section 10.17 puts numbers on this: a taper of width $\theta_{\rm apo}$ multiplies the leak of a sharp edge by $\pi^2/(2k\theta_{\rm apo})$ (Eq. (10.107)), about a fifth near $\ell \approx 750$ for the 2° Galactic taper but about three quarters for the 0.5° tapers of the small holes, and it gains nothing on scales longer than the taper; with hundreds of small holes, the overall gain stays small. Even so, a leaked half per cent of E is comparable to the lensing B itself, which is why the leaked power must be subtracted before B modes can be measured (Section 13.34). Research codes go further and use *pure* B estimators, which build the window's derivatives into the transform so that a pure E pattern gives no B at all on the cut sky [215]. The angle does not need them, because a rotation passes through the mask unchanged.

13.30.3 A rotation passes through the mask unchanged

The question. The window mixes E and B, and the rotation mixes E and B. Can the first fake or spoil the second? The toy skies of Section 10.17.1 settled it, and the reason carries over to any window. The window multiplies $Q + iU$ at each pixel by its own real weight W; a uniform rotation multiplies it by $e^{2i\beta}$, the same number at every pixel; and two multiplications can be done in either order. So rotating the sky and then masking it gives the same map as masking it and then rotating (Figure 13.56). The spin-2 transform is linear, so the constant $e^{2i\beta}$ goes

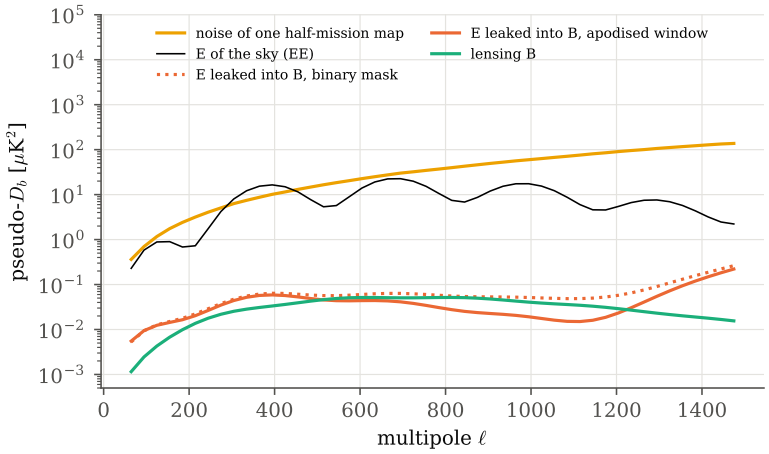


Figure 13.55: What a pseudo-BB spectrum of one half-mission map is made of, averaged over the simulated skies. The noise (yellow) is above everything else except the E power of the sky (black) near its first peak, $\ell \approx 400$. Far below, the share of E that the binary mask turns into B (dotted orange), the smaller share left by the tapered window (solid orange; the two differ little below $\ell \approx 600$), and the B modes made by lensing (green), seen through the same window.

straight through it, and the coefficients $\widetilde{E}_{\ell m}$, $\widetilde{B}_{\ell m}$ of the masked map, the **pseudo-coefficients**, turn exactly as the full-sky coefficients do (Chapter T1):

$$\widetilde{E}'_{\ell m} = \cos 2\beta\, \widetilde{E}_{\ell m} - \sin 2\beta\, \widetilde{B}_{\ell m}, \qquad \widetilde{B}'_{\ell m} = \sin 2\beta\, \widetilde{E}_{\ell m} + \cos 2\beta\, \widetilde{B}_{\ell m}. \tag{13.64}$$

The beam and the pixel window multiply E and B by the same number T_ℓ (input (iii)), so they commute with the rotation as well. Whatever the window has done, the rotation acts on the pseudo-coefficients as if there were no window.

The pseudo-spectra, $\widetilde{C}_\ell^{XY} = \frac{1}{2\ell+1}\sum_m \mathrm{Re}\,\widetilde{X}_{\ell m}\widetilde{Y}^*_{\ell m}$, therefore obey the rotation formulas of Chapter T1, now with tildes: substituting (13.64) into the definition and multiplying out, exactly as for the full sky, gives

$$\widetilde{C}_\ell^{EB\prime} = \tfrac{1}{2}\sin 4\beta\,\big(\widetilde{C}_\ell^{EE} - \widetilde{C}_\ell^{BB}\big) + \cos 4\beta\, \widetilde{C}_\ell^{EB}, \qquad \widetilde{C}_\ell^{TB\prime} = \sin 2\beta\, \widetilde{C}_\ell^{TE} + \cos 2\beta\, \widetilde{C}_\ell^{TB}, \tag{13.65}$$

the second because T is not rotated. These are the formulas (13.62) again, but for the *pseudo*-spectra, and they hold sky by sky.

The last term on average. What remains is $\widetilde{C}_\ell^{EB}$ of the unrotated sky. As on the toy skies, the masked EB is fed by the true EB alone, never by EE or BB [169]: the leak adds B power but no correlation with E. A sky with $C_\ell^{EB} = 0$ therefore has $\langle\widetilde{C}_\ell^{EB}\rangle = 0$, and likewise for TB. Combining,

$$\big\langle\widetilde{C}_\ell^{EB\prime}\big\rangle = \tfrac{1}{2}\sin 4\beta\,\big\langle\widetilde{C}_\ell^{EE} - \widetilde{C}_\ell^{BB}\big\rangle, \qquad \big\langle\widetilde{C}_\ell^{TB\prime}\big\rangle = \sin 2\beta\,\big\langle\widetilde{C}_\ell^{TE}\big\rangle. \tag{13.66}$$

We check this on six hundred simulated skies in Section 13.31.3: with no rotation, the estimated angle averages to zero.

This is why no spin-2 version of the MASTER correction of Section 13.3.4 appears in the angle measurement: the leaked E sits in both $\widetilde{C}_\ell^{EE}$ and $\widetilde{C}_\ell^{BB}$ of the template, and the angle is read from pseudo-spectra directly. The leakage does still cost something: the leaked E in $\widetilde{B}$ adds to the variance of $\widetilde{C}_\ell^{EB}$, in the same way that noise does. On the toy skies this mattered only for a map deep enough to see below the lensing B (Section 10.17.3); SMICA's polarization is far noisier, and Figure 13.55 shows that its noise is so much larger than the leakage through the tapered window that the leakage costs essentially nothing.

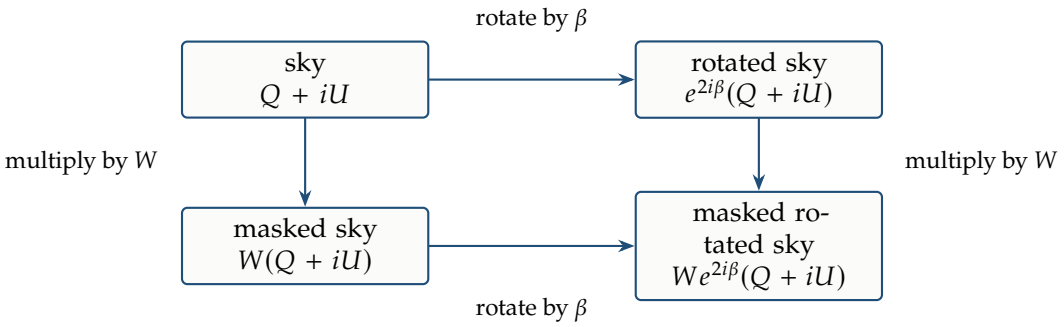


Figure 13.56: Both routes from the sky (top left) to the masked rotated sky (bottom right) give the same map, because the rotation multiplies every pixel by the same number $e^{2i\beta}$ and the window multiplies every pixel by its own real weight. On the bottom row the rotation therefore acts on the pseudo-coefficients exactly as it acts on the full-sky coefficients.

13.30.4 Cross-spectra of the halves, and the noise

No noise bias. As in Section 13.3.2, the noise of the two half-mission maps is independent, so a spectrum formed from one coefficient of each half has no noise term on average. For a parity-odd spectrum there are two such products, $\widetilde{E}^{(1)}\widetilde{B}^{(2)*}$ and $\widetilde{E}^{(2)}\widetilde{B}^{(1)*}$, and both have the same mean. We average them,

$$\widehat{C}_\ell^{EB} = \frac{1}{2\ell+1}\sum_m \tfrac{1}{2}\,\mathrm{Re}\big(\widetilde{E}^{(1)}_{\ell m}\widetilde{B}^{(2)*}_{\ell m} + \widetilde{E}^{(2)}_{\ell m}\widetilde{B}^{(1)*}_{\ell m}\big), \tag{13.67}$$

and the same for TE and TB; for EE and BB there is only one product. Each half is rotated by the same angle, so (13.65) holds term by term for (13.67). All spectra are then binned in 48 bins of $\Delta\ell = 30$ between $\ell = 50$ and 1489, as plain averages of D_ℓ over the bin. The range follows the published analyses, which start at $\ell = 51$ because the largest scales of the *Planck* polarization maps carry systematic residuals, and stop near $\ell = 1500$ [62]; at $N_{\mathrm{side}} = 512$ we cannot go much further anyway.

Listing 13.11: The symmetrised cross-spectra of two half-mission maps (`code/ch13/lib_biref.py`)

```
def xcl(a, b):
    """Symmetrised cross-spectrum 1/2 (a1 b2* + b1 a2*) given pairs a=(a1,a2), b=(b1,b2)."""
    return 0.5 * (hp.alm2cl(a[0], b[1]) + hp.alm2cl(b[0], a[1]))


def cross_spectra(h1, h2):
    """h = (T, E, B) pseudo-a_lm of one half.  Binned cross-spectra of the two halves."""
    T, E, B = zip(h1, h2)
    return {"EE": binned(hp.alm2cl(E[0], E[1])), "BB": binned(hp.alm2cl(B[0], B[1])),
            "EB": binned(xcl(E, B)), "TE": binned(xcl(T, E)), "TB": binned(xcl(T, B))}
```

`xcl` forms the average of the two cross products in (13.67), and `cross_spectra` returns the five binned spectra from the pseudo-coefficients (T, E, B) of each half. The pseudo-coefficients themselves come from `hp.map2alm` with `pol=True` applied to WQ and WU.

The noise model. We build it as in Section 13.4.1, now for Q and U. The half-difference $d = (\mathrm{hm1} - \mathrm{hm2})/2$ contains no sky. Its local variance $v(p)$, the average of d_Q^2 and d_U^2 smoothed over 2° and divided by its mean over the kept sky, says how the noise level varies over the sky (Figure 13.54d; at most 2.7 times the mean). The E and B spectra of $d/\sqrt{v}$, divided by $\langle W^2\rangle$ and averaged over 31 neighbouring multipoles, say how it varies with scale. A simulated

half-mission noise map is then $\sqrt{2}\sqrt{v}$ times a Gaussian polarization field with those two spectra, the $\sqrt{2}$ derived in Section 13.4.1. The noise is almost the same in E and in B (their ratio is 0.96 at $600 < \ell < 900$), as it should be for noise that knows nothing about the patterns on the sky.

Example: the noise level in familiar units. Noise that is white on the sky has a flat spectrum, $N_\ell = \sigma^2_{\rm pix}\Omega_{\rm pix}$, and experiments quote $\sqrt{N_\ell}$ in μK arcmin: the noise in a pixel of one square arcminute. Dividing the measured noise spectrum of one half-mission map by the squared beam and pixel window and averaging over $600 < \ell < 900$ gives 91 μK arcmin. For comparison, the temperature noise of Section 13.4.1 stayed below the sky at every multipole we used; here the noise power of one half is above the E signal at almost every multipole we use (the exception is the first E peak near $\ell \approx 400$), and about 1471 times the lensing B near $\ell = 980$ (Figure 13.55). Every result below is limited by this noise.

13.31 The estimator, tested on skies where we know the angle

A carpenter who is handed a new ruler measures something of known length before trusting it on the table she is building. The estimator of the angle is our ruler, and the object of known length is a simulated sky: we rotate it by an angle we choose, pass it through everything the real sky went through, and see what the estimator says. One such sky tells us little, because one estimate scatters. Hundreds of them tell us three things: whether the estimator is right on average (its *bias*), how much it scatters (its *error*), and whether the error bar it reports for each sky contains the truth as often as it should (its *coverage*). The estimator comes first, from least squares; six hundred skies then give the three numbers.

The same carpentry was done once already, on a toy. In Section 10.15 a Fisher forecast told us how the error on the angle shrinks with lower noise, more sky and more multipoles, before any sky existed; in Section 10.17 simulated toy skies checked that forecast and the estimator's bias, scatter and coverage, mask leakage included. What is new here is that every ingredient is the real one: SMICA's noise, measured from its own half-missions, SMICA's mask, and a covariance that has to be estimated rather than written down. So the toy's conclusions are a prediction, and the simulated skies below test whether they survive contact with the real pipeline.

13.31.1 Least squares for one angle

The model. Stack the binned parity-odd spectra into one list, $\boldsymbol{e} = (\widehat{D}^{EB}_b, \widehat{D}^{TB}_b)$, with 2×48 entries, and the spectra they should copy into a second list of the same length, $\boldsymbol{D} = (\widehat{D}^{EE}_b - \widehat{D}^{BB}_b, \widehat{D}^{TE}_b)$. For a small angle, (13.66) says that on average $\boldsymbol{e} = 2\beta\boldsymbol{D}$, since $\frac{1}{2}\sin 4\beta \approx 2\beta$ and $\sin 2\beta \approx 2\beta$. The data scatter around that mean with a covariance $\mathbf{C}$, so we write

$$\boldsymbol{e} = 2\beta\,\boldsymbol{D} + \boldsymbol{n}, \qquad \boldsymbol{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}). \tag{13.68}$$

This is the linear model with one parameter of Section 3.10.2. The template $\boldsymbol{D}$ is made of the *measured* spectra, not of a theory curve, as in the term of [62] that carries the rotation of everything in the map: it is then automatically seen through the same window, beam and pixels as $\boldsymbol{e}$, and no theory enters at all.

The best angle. The likelihood of (13.68) is largest where $\chi^2(\beta) = (\boldsymbol{e} - 2\beta\boldsymbol{D})^{\mathsf{T}}\mathbf{C}^{-1}(\boldsymbol{e} - 2\beta\boldsymbol{D})$ is smallest. Setting its derivative to zero,

$$0 = \frac{\mathrm{d}\chi^2}{\mathrm{d}\beta} \overset{(1)}{=} -4\,\boldsymbol{D}^{\mathsf{T}}\mathbf{C}^{-1}\big(\boldsymbol{e} - 2\beta\boldsymbol{D}\big)$$
$$\Longrightarrow \qquad \hat{\beta} \overset{(2)}{=} \frac{\boldsymbol{D}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{e}}{2\,\boldsymbol{D}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{D}}. \tag{13.69}$$

(1) Differentiate the quadratic form: $\mathrm{d}(\boldsymbol{r}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{r})/\mathrm{d}\beta = 2(\mathrm{d}\boldsymbol{r}/\mathrm{d}\beta)^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{r}$ for the symmetric $\mathbf{C}^{-1}$, with $\boldsymbol{r} = \boldsymbol{e} - 2\beta\boldsymbol{D}$ and $\mathrm{d}\boldsymbol{r}/\mathrm{d}\beta = -2\boldsymbol{D}$.
(2) Expand, $\boldsymbol{D}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{e} = 2\beta\,\boldsymbol{D}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{D}$, and divide by $2\boldsymbol{D}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{D}$.

Its error. $\hat{\beta}$ is a fixed linear combination $\boldsymbol{w}^{\mathsf{T}}\boldsymbol{e}$ of the data, with $\boldsymbol{w} = \mathbf{C}^{-1}\boldsymbol{D}/(2\boldsymbol{D}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{D})$, so

$$\mathrm{Var}\,\hat{\beta} \overset{(1)}{=} \boldsymbol{w}^{\mathsf{T}}\mathbf{C}\,\boldsymbol{w} \overset{(2)}{=} \frac{\boldsymbol{D}^{\mathsf{T}}\mathbf{C}^{-1}\mathbf{C}\,\mathbf{C}^{-1}\boldsymbol{D}}{4\,(\boldsymbol{D}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{D})^2} \overset{(3)}{=} \frac{1}{4\,\boldsymbol{D}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{D}}, \qquad \sigma_\beta = \frac{1}{2\sqrt{\boldsymbol{D}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{D}}}. \tag{13.70}$$

(1) The variance of a linear combination $\boldsymbol{w}^{\mathsf{T}}\boldsymbol{e}$ of a vector with covariance $\mathbf{C}$ (Section 3.6.1); $\boldsymbol{D}$ is treated as fixed.
(2) Substitute $\boldsymbol{w}$; $\mathbf{C}^{-1}$ is symmetric.
(3) $\mathbf{C}^{-1}\mathbf{C} = \mathbb{1}$, and one factor $\boldsymbol{D}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{D}$ cancels.

The template is itself measured and noisy, which (13.70) ignores. Whether that matters is one of the things the simulated skies decide, since they carry a noisy template too.

Example: one bin, then two. With a single bin, $\boldsymbol{e} = e$, $\boldsymbol{D} = D$ and $\mathbf{C} = \sigma^2$, (13.69) is $\hat{\beta} = e/2D$, the naive ratio of Section 13.29, and its error is $\sigma/2|D|$. With two independent bins,

$$\hat{\beta} \overset{(1)}{=} \frac{D_1 e_1/\sigma_1^2 + D_2 e_2/\sigma_2^2}{2\,(D_1^2/\sigma_1^2 + D_2^2/\sigma_2^2)} \overset{(2)}{=} \frac{w_1\hat{\beta}_1 + w_2\hat{\beta}_2}{w_1 + w_2}, \qquad w_b = \frac{D_b^2}{\sigma_b^2}, \tag{13.71}$$

(1) (13.69) with $\mathbf{C} = \mathrm{diag}(\sigma_1^2, \sigma_2^2)$.
(2) Write $D_b e_b = D_b^2 \cdot e_b/D_b = 2D_b^2\hat{\beta}_b$, with $\hat{\beta}_b = e_b/2D_b$ the angle from bin b alone.

The estimator is the weighted mean of the single-bin angles of Section 3.7, each weighted by one over its own variance, $(2|D_b|/\sigma_b)^2$: a bin counts in proportion to the square of how much E it has, compared with its noise. The wallpaper of Section 13.29 again: bold stripes count, a plain wall does not.

Example: the scatter of a cross-spectrum. The single-multipole numbers of Section 13.29 came from the variance of one cross product. Take one coefficient $x = E^{(1)}_{\ell m}$ from the first half and one $y = B^{(2)}_{\ell m}$ from the second, independent, with zero mean and variances $\sigma_x^2 = C_\ell^{EE} + N_\ell$ and $\sigma_y^2 = C_\ell^{BB} + N_\ell$ (sky plus the noise of one half). Then

$$\mathrm{Var}(xy) \overset{(1)}{=} \mathbb{E}[x^2y^2] - \big(\mathbb{E}[xy]\big)^2 \overset{(2)}{=} \mathbb{E}[x^2]\;\mathbb{E}[y^2] - 0 \overset{(3)}{=} \sigma_x^2\sigma_y^2. \tag{13.72}$$

(1) The definition of the variance.
(2) For independent x and y the mean of a product is the product of the means, applied to x^2, y^2 and to x, y (whose means are zero).
(3) Both means are zero, so $\mathbb{E}[x^2]$ is the variance.

Averaging over the $(2\ell+1)$ values of m, of which a fraction $f_{\rm sky}^{\rm eff}$ is effectively independent on the cut sky, divides this by $(2\ell+1)f_{\rm sky}^{\rm eff}$. With the noise far above C_ℓ^{BB}, the error of $\widehat{C}_\ell^{EB}$ is set by the noise times the E power plus noise: which is why the leakage, a small addition to σ_y^2, costs so little.

The covariance. The bins of $\widehat{D}^{EB}$ and $\widehat{D}^{TB}$ are correlated with each other, because both use the same B coefficients, and neighbouring bins are weakly correlated through the window. No formula gives this covariance exactly for our window and our noise, so we take it from simulated skies with no rotation, as in Section 13.4.3, and multiply its inverse by the Hartlap factor $(n-p-2)/(n-1)$ of Section 8.13.2. With $p = 2\times 48$ and $n = 400$ skies the factor is 0.76. The remaining 200 skies are kept apart and used only to test the estimator: a sky that helped build **C** would be a slightly too easy test object.

13.31.2 The simulated skies

The rule of Section 13.4.2 holds: whatever is done to the data is done to every simulated sky.

Drawing T and E together. The temperature and E polarization of the CMB are correlated: both are made by the same density waves at last scattering, and $C_\ell^{TE} \neq 0$. So for each (ℓ, m) we need two Gaussian coefficients with variances C_ℓ^{TT} and C_ℓ^{EE} and covariance C_ℓ^{TE}. With two independent unit Gaussians z_1, z_2 (complex for $m>0$, real for $m=0$, as in Section 8.5), set

$$a^T = \sqrt{C_\ell^{TT}}\, z_1, \qquad a^E = \frac{C_\ell^{TE}}{\sqrt{C_\ell^{TT}}}\, z_1 + \sqrt{C_\ell^{EE} - \frac{(C_\ell^{TE})^2}{C_\ell^{TT}}}\, z_2. \tag{13.73}$$

It has the right second moments:

$$\begin{aligned} \left\langle a^T a^{E*} \right\rangle &\overset{(1)}{=} \sqrt{C_\ell^{TT}}\, \frac{C_\ell^{TE}}{\sqrt{C_\ell^{TT}}} \left\langle |z_1|^2 \right\rangle \overset{(2)}{=} C_\ell^{TE}, \\ \left\langle |a^E|^2 \right\rangle &\overset{(3)}{=} \frac{(C_\ell^{TE})^2}{C_\ell^{TT}} + C_\ell^{EE} - \frac{(C_\ell^{TE})^2}{C_\ell^{TT}} \overset{(4)}{=} C_\ell^{EE}. \end{aligned} \tag{13.74}$$

(1) Multiply out; the term with $\langle z_1 z_2^* \rangle$ vanishes because z_1, z_2 are independent with zero mean.
(2) $\langle |z_1|^2 \rangle = 1$.
(3) The two squared coefficients, since the cross term vanishes for the same reason.
(4) The two fractions cancel.

This is the Cholesky factor of the 2×2 covariance matrix, the recipe of Section 2.12.1. B is drawn independently from the lensing B spectrum C_ℓ^{BB} of the fiducial model; EB and TB are zero.

One simulated sky.

1. Draw a^T, a^E, a^B for $\ell \le 1535$ from the CAMB fiducial spectra [8] with (13.73), multiply by the transfer functions, and make the temperature map and two polarization maps, one from E alone and one from B alone.
2. Make two half-mission noise maps for T (the model of Section 13.4.1) and for Q, U (Section 13.30.4).

3. Compute the pseudo-coefficients of the signal maps and of the noise maps separately, through the same windows as the data.
4. For each injected angle $\beta \in \{0, 0.3°, 1°\}$, rotate the signal pseudo-coefficients with (13.64), add the noise, and compute the five binned cross-spectra.

Step 4 is where Section 13.30.3 pays off a second time. Rotating the pseudo-coefficients is *exactly* the same as rotating the sky before observing it, so one simulated sky yields three rotated skies for the price of one, with identical CMB and noise: the differences between them are due to the rotation alone. The noise is not rotated. Keeping the E part and the B part of the sky apart costs one more transform and gives, for free, the leakage and the lensing spectra of Figure 13.55, and a version of every sky with the lensing B removed.

Listing 13.12: One simulated polarized sky, observed like the real one, rotated three ways (`code/ch13/43_pol_sims.py`)

```
def one_sky(rng, with_binary):
    aT, aE, aB = lb.correlated_alm(cls, L, rng)
    tsky = hp.alm2map(hp.almxfl(aT, TT_), N, lmax=L)
    qE, uE = spin_synth(hp.almxfl(aE, TP), zero)
    qB, uB = spin_synth(zero, hp.almxfl(aB, TP))
    sT = hp.map2alm(WT * tsky, lmax=L, iter=0)
    sE = spin_anal(WP, qE, uE)                  # pseudo (E, B) of the E part: B here is leakage
    sB = spin_anal(WP, qB, uB)                  # pseudo (E, B) of the lensing B part
    cE, cB = spin_anal(WC, qE + qB, uE + uB)      # the stricter window, beta = 0
    cEl = spin_anal(WC, qE, uE)
    noise, cnoise = [], []
    for _ in range(2):
        nt = sqvT * np.sqrt(2.0) * hp.alm2map(cs.synalm(ngT, L, rng), N, lmax=L)
        nq, nu = spin_synth(cs.synalm(ngE, L, rng), cs.synalm(ngB, L, rng))
        nq, nu = np.sqrt(2.0) * sqvP * nq, np.sqrt(2.0) * sqvP * nu
        nE, nB = spin_anal(WP, nq, nu)
        noise.append((hp.map2alm(WT * nt, lmax=L, iter=0), nE, nB))
        cnoise.append(spin_anal(WC, nq, nu))
    out = {}
    sigE, sigB = sE[0] + sB[0], sE[1] + sB[1]
    for i, beta in enumerate(BETAS):
        rE, rB = lb.rotate(sigE, sigB, beta)
        h = [(sT + n[0], rE + n[1], rB + n[2]) for n in noise]
        s = lb.cross_spectra(*h)
        for k, v in s.items():
            out[f"b{i}_{k}"] = v
    h = [(sT + n[0], sE[0] + n[1], sE[1] + n[2]) for n in noise]     # no lensing B in the sky
    for k, v in lb.cross_spectra(*h).items():
        out[f"nolens_{k}"] = v
```

`lb.correlated_alm` implements (13.73); `spin_synth` and `spin_anal` are healpy's spin-2 transforms. The variables `sE` and `sB` hold the pseudo-coefficients (E and B) of the E part and of the B part of the sky; their sum is the observed sky. The loop over `BETAS` is step 4. One sky takes 6.2 s on eight cores; we ran 600 of them, in chunks of 25 cached in `data/ch13/`.

13.31.3 What the simulations say

Is the estimator unbiased? On the 200 test skies with no rotation the joint estimator averages 0.000°, with a Monte Carlo error of 0.004° on that average (the scatter divided by $\sqrt{200}$). With an injected 0.3° the average error is 0.000° ± 0.004°, and with 1° it is 0.001° ± 0.004° (Figure 13.57). In each case the question is whether an average as far from zero as this one, or farther, would

be something we would reasonably expect from an unbiased estimator with this Monte Carlo error; the answer is read off by comparing each bias with its error. With no rotation, this also confirms (13.66): the mask leaks E into B, but no EB appears on average.

Is the error bar right? For EB alone, the scatter of the estimate over the test skies with no rotation is 0.048°, and the error bar the estimator reports, (13.70) averaged over the skies, is 0.048°: a ratio of 1.00. TB alone scatters by 0.087°, almost twice as much: both copy their template with the same factor 2β, but the TB template, the TE correlation, is small compared with the scatter of $\widehat{C}_\ell^{TB}$, which is the polarization noise multiplied by the large temperature power. For the joint estimate the reported error is 0.045°, smaller than that of EB alone, as adding information should make it. But the joint estimate actually scatters by 0.050°, a ratio of 1.11, and no better than EB alone. Each of these scatters is itself estimated from 200 skies, so it carries a Monte Carlo error of $1/\sqrt{2(200-1)}$, about five per cent; a ratio is meaningful only when it differs from one by more than that.

Why the joint error bar is too small. Formula (13.70) assumes that $\mathbf{C}$ is the true covariance. Ours is estimated from $n = 400$ skies, and its inverse, although corrected on average by the Hartlap factor, is noisy element by element. Noisy weights are not the optimal weights, and an estimator with slightly wrong weights scatters more than the optimal one: the extra variance is a factor of about $1 + p/n$ for p bins [281]. For the joint estimate, $p = 2 \times 48$ and $\sqrt{1 + p/n} = 1.11$, which the measured ratio matches within its five per cent Monte Carlo error; for EB alone the factor is 1.06, too close to one for two hundred skies to resolve. With four hundred skies, doubling the number of bins costs about as much as the TB information gains. So for the real sky we quote the joint angle with the error the test skies give it, their scatter 0.05°, not the formula's.

Does the interval cover the truth? A 68% interval $\hat\beta \pm \sigma_\beta$ should contain the true angle in 68% of the skies, and $\hat\beta \pm 1.96\sigma_\beta$ in 95%. With 200 skies a coverage fraction has a binomial error of about 3 percentage points near 68% and 2 near 95%. For EB alone the reported intervals cover the truth in 68% and 95% of the skies, as they should. For the joint estimate with the formula's error bar they cover it in 65%, 65% and 64% of the skies for the three angles, and in 92%, 92% and 92% at $1.96\sigma_\beta$. An error bar that is too small by the factor 1.11 expected from the noise of the covariance should cover the truth in the fraction of a Gaussian within $1/1.11 = 0.90$ of its standard deviation, 63%, and within $1.96/1.11 = 1.76$, 92%; the measured fractions are to be read against those two numbers and against 68 and 95, with the binomial errors just quoted. Two hundred skies make coverage a blunt instrument: it can expose an error bar that is wrong by a tenth only at the edge of its resolution, which is why the direct comparison of the scatter with the reported error, above, is the sharper test.

Lensing B: variance, not bias. Every simulated sky also exists without its lensing B, with the same CMB E and the same noise. Lensing B is a parity-even signal: on average it adds to C^{BB} and nothing to C^{EB}. In the cross-spectrum example above it is a small addition to σ_y^2, so it can only add variance, and next to the noise, which is more than a thousand times larger near $\ell = 1000$, the addition is tiny. The skies agree. With lensing the joint estimate averages 0.000°, without it −0.000°, both within the Monte Carlo error of zero; the scatters, 0.0496° and 0.0499°, are equal to within a per cent, and which of the two comes out larger depends on the particular skies drawn; and the sky-by-sky difference between the two estimates, which cancels the shared CMB and noise, scatters by only 0.0033°. So lensing B leaves the angle unbiased and adds variance too small for this noise level to show, as the parity argument says.

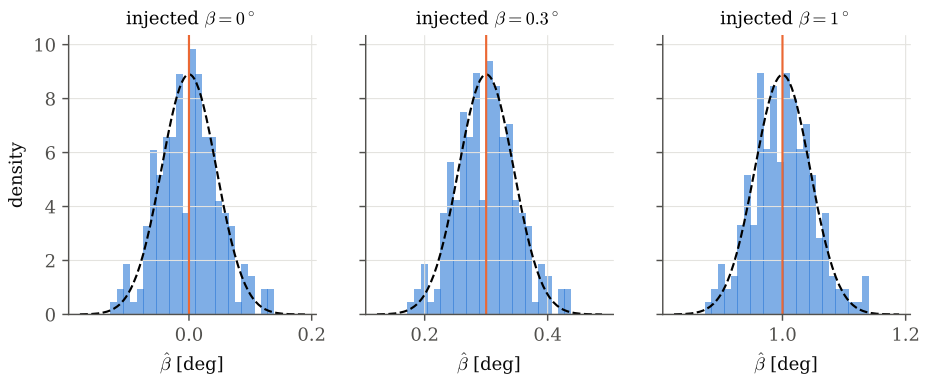


Figure 13.57: The joint estimator on the 200 held-out simulated skies, for three injected angles (orange lines). Each histogram shows the recovered angles. The dashed Gaussian is centred on the injected angle and has a width equal to the mean reported error bar. Rotating the same skies by three different angles shifts the histograms while preserving their shapes.

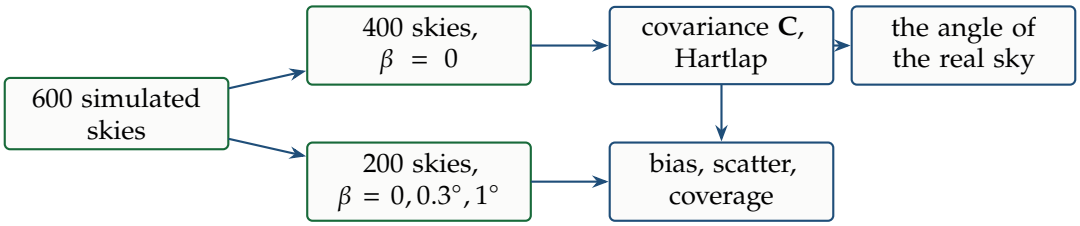


Figure 13.58: How the simulated skies are used. Two thirds build the covariance of the parity-odd spectra; the other third, never used for the covariance, tests the estimator with known angles. The same covariance is then used on the real sky.

13.32 What the SMICA sky says

The ruler has been checked on skies whose angle we chose. It is unbiased, and on four hundred skies' worth of covariance its honest error for the joint estimate is the scatter of the test skies, 0.05°. Now we lay it on the one sky whose angle nobody chose.

Listing 13.13: The estimator (13.69) and its formal error (13.70) (`code/ch13/lib_biref.py`)

```
def estimate_beta(s, Cinv, which="joint"):
    """Least squares for e = 2 beta D + noise:  beta = D C^-1 e / (2 D C^-1 D),  sigma = 1/(2 sqrt(D C^-1 D))."""
    e, D = vectors(s, which)
    F = D @ Cinv @ D
    return (D @ Cinv @ e) / (2 * F), 1.0 / (2 * np.sqrt(F))
```

`vectors` stacks the binned spectra into $\boldsymbol{e}$ and $\boldsymbol{D}$ as in (13.68), the two dot products are $\boldsymbol{D}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{e}$ and $\boldsymbol{D}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{D}$, and the function returns (13.69) and (13.70). The data spectra go in exactly as the simulated ones did.

The event. The SMICA half-mission cross-spectra give

$$\hat{\beta}_{EB} = 0.28^\circ \pm 0.048^\circ, \qquad \hat{\beta}_{TB} = 0.37^\circ \pm 0.092^\circ, \qquad \hat{\beta} = 0.31^\circ \pm 0.05^\circ \ \text{(EB and TB together)}, \tag{13.75}$$

where the first two errors are the formula's, which the test skies confirmed for EB and TB alone, and the last is the scatter of the test skies (Section 13.31.3). The two separate angles agree: their difference is about one TB error. Figure 13.59 shows what was fitted. Above $\ell \approx 500$ the

measured $\widehat{D}_b^{EB}$ wander around zero with error bars larger than the curve $2\hat{\beta}(\widehat{D}_b^{EE} - \widehat{D}_b^{BB})$, which follows the acoustic wiggles of the E spectrum. Below $\ell \approx 500$ the noise is smallest and the error bars shrink, and there the points follow the first rise of the curve. Those bins carry most of the weight, but the fit does not rest on any one of them: it adds up 48 bins of EB and as many of TB, which on balance lean in the direction the template predicts. The wallpaper again.

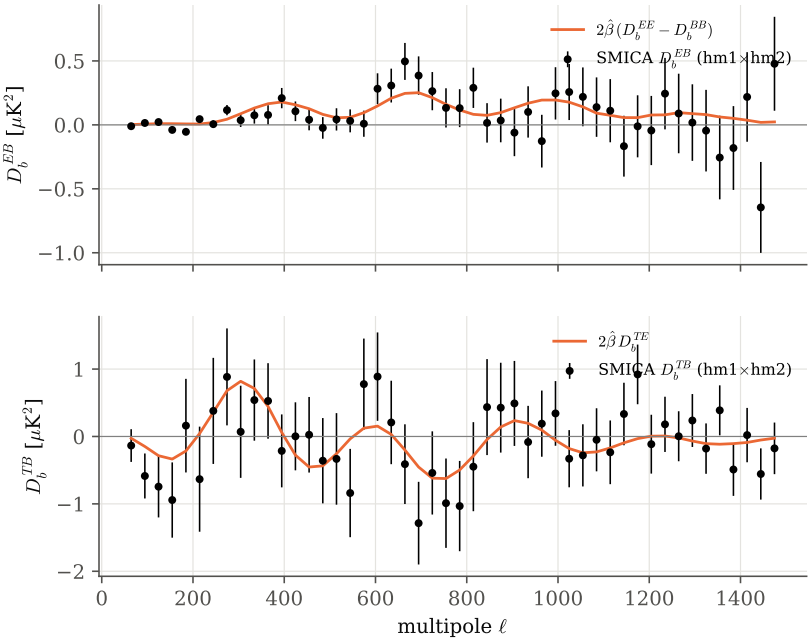


Figure 13.59: The SMICA half-mission cross-spectra that measure the angle, with error bars from the simulated skies. Top: $\widehat{D}_b^{EB}$ (points) and the rotation model $2\hat{\beta}(\widehat{D}_b^{EE} - \widehat{D}_b^{BB})$ for the joint angle (orange). Bottom: $\widehat{D}_b^{TB}$ and $2\hat{\beta}\,\widehat{D}_b^{TE}$. Above $\ell \approx 500$ the model is smaller than one error bar; the angle is measured by the coherent lean of all bins together, with most weight in the low-noise EB bins below $\ell \approx 500$.

What could have produced it? We return to the candidates of Section 13.29, now with a number to explain.

1. *Chance, with no rotation at all.* This is the one candidate whose sampling distribution we know, so it is our null hypothesis. If the true angle were zero, would an estimate of 0.31°, or one even farther from zero, be something we would reasonably expect? The angle is 6.3 times its error. A cleaner way to ask, which does not lean on the Gaussian shape of the estimate, is to compare how much better the rotated model fits than the unrotated one, $\Delta\chi^2 = \chi^2(0) - \chi^2(\hat{\beta})$, with the same quantity on the test skies with no rotation. On the data $\Delta\chi^2 = 48.2$. On the skies it follows the χ^2 distribution with one degree of freedom that Wilks' theorem predicts (Section 4.7), and none of the 200 skies comes anywhere near the data (Figure 13.60a). With 200 skies the simulations alone can only say that the probability is below about 1/200; the χ^2_1 tail, which the skies follow, puts it many orders of magnitude lower. Chance is ruled out.

2. *A leak of E into B through the mask.* Ruled out by (13.66) and by the test skies, whose mask leaks exactly as the data's does and whose angle averages to $0.000° \pm 0.004°$.

3. *Galactic dust with an EB of its own.* Dust residuals are strongest near the plane, so if they made the angle, removing more sky near the plane would change it. With the strict window the angle is $0.32° \pm 0.05°$. The two windows share most of the sky, so the right yardstick for

their difference is not either error bar but the scatter of the difference over the test skies, each analysed through both windows: 0.03°. The data's shift, −0.01°, is −0.2 times that scatter (Figure 13.60b). No sign of a dust-made angle; but dust whose twist is the same everywhere would not change with the mask, and this test cannot exclude it.

4. *A miscalibration α or a cosmic rotation β.* Both remain. Both produce exactly the pattern we fitted, and Section 13.33 explains why no test on this map can tell them apart.

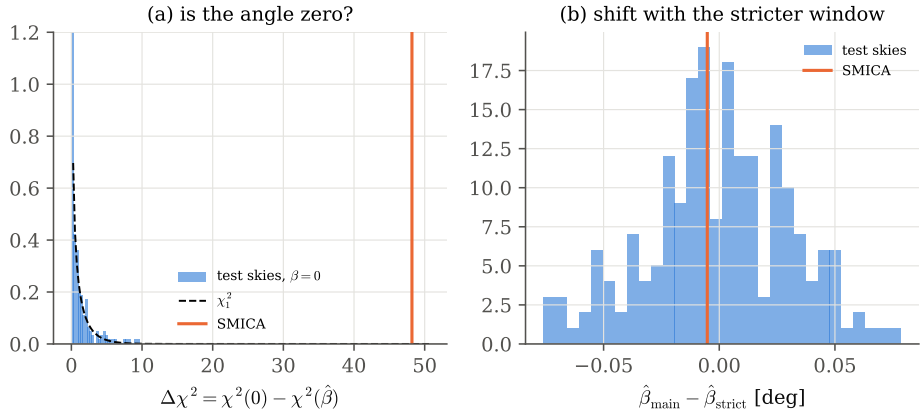


Figure 13.60: Two checks against the simulated skies. (a) How much the rotated model improves the fit, on the test skies with no rotation (histogram, following the χ^2_1 curve) and on the data (orange line). (b) The change of the angle between the main and the strict window, over the test skies and on the data: the data sit in the middle of the distribution.

13.32.1 Is the rotation model enough?

A model can win against no rotation and still be wrong. Three more questions test whether one angle explains the parity-odd spectra.

Does one angle fit all the bins? With the best angle, the remaining $\chi^2 = (\boldsymbol{e} - 2\hat{\beta}\boldsymbol{D})^{\mathsf{T}}\mathbf{C}^{-1}(\boldsymbol{e} - 2\hat{\beta}\boldsymbol{D})$ is 95.6 for 95 degrees of freedom (96 numbers, one parameter). The test skies, fitted the same way, give a χ^2 larger than this in a fraction 0.51 of cases. Nothing in the EB and TB spectra calls for more than one angle.

Do low and high multipoles agree? The angle from the bins with $\ell < 770$ is 0.32° ± 0.05°, and from $\ell \geq 770$ it is 0.25° ± 0.12° (formal errors). Their difference is less than one standard error of the difference. Most of the weight sits below $\ell \approx 800$, where the E spectrum is large compared with the noise. A rotation must be the same at every scale; an EB made by dust or by a beam effect would generally not be.

Does the template matter? The template $\boldsymbol{D}$ was made of the measured spectra. Replacing it by the mean of the simulated spectra, a noiseless version of the fiducial model seen through the same window, gives 0.31°. The two agree because the measured E spectrum is close to the fiducial one: in the first twenty bins their ratio is 1.008.

The answer. The SMICA polarization is turned, relative to the directions the instrument assumes, by

$$\hat{\beta}_{\rm SMICA} = 0.31^\circ \pm 0.05^\circ \quad (68\%,\ \text{statistical}), \tag{13.76}$$

an effect that chance does not explain, that the mask cannot produce, that does not move when we cut away more of the Galaxy and that has the same size at large and small scales. What it is an angle *of* is a separate question, and the published measurements, made with a different kind of map, answer it.

13.33 The published angle, and what one map can measure

Their question. By 2016 the *Planck* team had measured the rotation of the CMB polarization from the foreground-cleaned maps of its 2015 release and found $0.31^\circ \pm 0.05^\circ$ [214]. The statistical error was small, but next to it stood a second one, $\pm 0.28^\circ$: the uncertainty of the ground measurement of the detectors' polarization directions. That second error swallowed the first, and it could not be reduced by more data, because every CMB measurement of the angle measures the same sum $\alpha + \beta$. Minami and Komatsu asked whether the sky itself could calibrate the detectors, so that β could be separated from α [62, 76]. That is the door frame of Section 13.29: something in the picture that is known to hang straight.

13.33.1 Their setup, in plain terms

Instead of one cleaned map they used the four polarized frequency maps of the High Frequency Instrument (100, 143, 217 and 353 GHz), each in its two half-mission versions, and all their cross-spectra, so that no noise bias enters. Each frequency was allowed its own miscalibration α_ν, because different detectors can be misaligned differently, and the sky one common β. The frequency maps still contain the Galaxy, and that is the point: the dust emission, bright at 353 GHz, travels only a few hundred parsecs and is rotated by α_ν but not by β. They masked only bright CO regions, bright sources and unobserved pixels, keeping most of the sky, tapered the masks by 0.5°, computed the spectra with NaMaster [169] between $\ell = 51$ and 1500 in bins of $\Delta\ell = 20$, and sampled the posterior of the five angles with emcee. The baseline assumes that the dust itself has no EB correlation (Table 13.12).

13.33.2 Their key equation

Chapter T1 derives the relation behind their method from the rotated spectra. Its content, with the reason for each step in a sentence, is short. Take one frequency and write the sky in it as the CMB plus the Galactic foreground. On the way to the detector the CMB is rotated by β and then everything by α; the foreground is rotated only by α. Assume, as their baseline does, that neither component has an EB of its own and that the two are uncorrelated, so their cross terms vanish on average. Each component then contributes to the observed spectra as in (13.62), with its own angle and its own $\Delta_\ell = C_\ell^{EE} - C_\ell^{BB}$ before rotation:

$$\begin{aligned} C_\ell^{EB,\mathrm{o}} &= \tfrac{1}{2}\sin\big(4(\alpha+\beta)\big)\,\Delta_\ell^{\mathrm{cmb}} + \tfrac{1}{2}\sin(4\alpha)\,\Delta_\ell^{\mathrm{fg}}, \\ C_\ell^{EE,\mathrm{o}} - C_\ell^{BB,\mathrm{o}} &= \cos\big(4(\alpha+\beta)\big)\,\Delta_\ell^{\mathrm{cmb}} + \cos(4\alpha)\,\Delta_\ell^{\mathrm{fg}}. \end{aligned} \tag{13.77}$$

The second line follows by multiplying out as in (13.65): a rotation by θ multiplies the $EE - BB$ of a component with no EB by $c^2 - s^2 = \cos 4\theta$. We do not know $\Delta_\ell^{\mathrm{fg}}$, so we eliminate it: subtracting $\frac{1}{2}\tan 4\alpha$ times the second line from the first cancels the foreground, because $\tan 4\alpha \cos 4\alpha = \sin 4\alpha$, and the sine subtraction formula $\sin A \cos B - \cos A \sin B = \sin(A - B)$, with $A = 4(\alpha + \beta)$ and $B = 4\alpha$, collects what is left of the CMB:

$$C_\ell^{EB,\mathrm{o}} - \tfrac{1}{2}\tan(4\alpha)\big(C_\ell^{EE,\mathrm{o}} - C_\ell^{BB,\mathrm{o}}\big) = \frac{\sin 4\beta}{2\cos 4\alpha}\,\Delta_\ell^{\mathrm{cmb}}. \tag{13.78}$$

Figure 13.61 draws the two paths. This is the single-frequency form of their equation [62, (2)], derived in [76]. Read it as a fit with two templates. The *observed* $C^{EE,\mathrm{o}} - C^{BB,\mathrm{o}}$, CMB and foreground together, enters with the coefficient $\frac{1}{2}\tan 4\alpha$; the *theoretical* CMB $\Delta_\ell^{\mathrm{cmb}}$ enters with a coefficient set by β. Symbol map to ours: their $C_\ell^{XY,\mathrm{o}}$ are our binned cross-spectra $\widehat{D}_b^{XY}$; their Δ^{cmb} is the fiducial $C^{EE} - C^{BB}$ through beam and pixels; our single β in (13.68) is their $\alpha + \beta$.

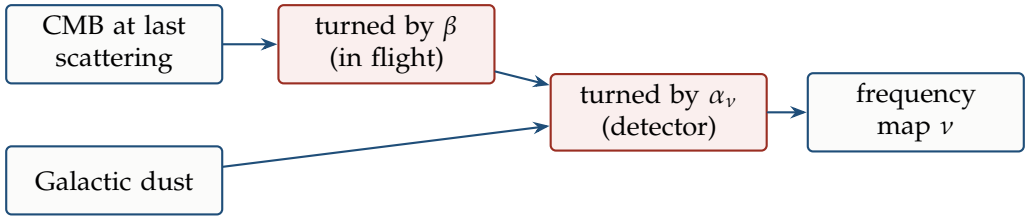


Figure 13.61: The two rotations of (13.77). The CMB is turned twice, once on its way across the Universe and once by the detector; the dust of our Galaxy is turned only by the detector. In a map that still contains the dust the two components carry different angles, and the fit (13.79) can tell α_ν from β; in a map from which the dust has been removed only $\alpha_\nu + \beta$ is left.

Why the two angles separate, and when they do not. For small angles the first line of (13.77) reads

$$C_\ell^{EB,\mathrm{o}} \approx 2(\alpha + \beta)\,\Delta_\ell^{\mathrm{cmb}} + 2\alpha\,\Delta_\ell^{\mathrm{fg}}, \tag{13.79}$$

a sum of two known shapes with two unknown amplitudes. Two amplitudes can be fitted separately only if the two shapes are different. They are: the dust spectrum is a smooth power law, large at low ℓ, while the CMB spectrum has acoustic peaks. So the fit finds α from the foreground shape and $\alpha + \beta$ from the CMB shape, and their difference is β. Now remove the foreground: $\Delta^{\mathrm{fg}} = 0$, and (13.79) contains α and β only through their sum. Any pair with the same sum fits equally well. This is exactly the situation of a component-separated map: SMICA was built to suppress the Galaxy, so it removed the door frame along with the dust. The degeneracy is not a weakness of our estimator. It is a property of the map.

13.33.3 What we compute

With the SMICA map, (13.78) collapses to our model (13.68), and our angle is

$$\hat{\beta}_{\mathrm{SMICA}} = \widehat{\alpha_{\mathrm{eff}} + \beta}, \tag{13.80}$$

where α_{eff} is the miscalibration that the SMICA combination inherits from its input channels. SMICA forms its B map from the channels' B maps with harmonic weights, and its E map from their E maps with other weights. If channel ν is turned by a small α_ν, its B map contains $2\alpha_\nu$ times its E map, so the B map of SMICA contains $2\sum_\nu w_{\ell\nu}^B \alpha_\nu$ times the CMB E map. The effective angle is the average of the channel angles weighted by the SMICA B-mode weights, which favour the cleanest CMB channels, 143 and 217 GHz [19].

13.33.4 Side by side

Our SMICA angle sits on top of the two other measurements of $\alpha + \beta$, within a fraction of their statistical errors, and inside the range of the β measurements. Given how different the data and the pipelines are, the agreement is the first thing to explain.

analysis	data	what it measures	angle [deg]
Planck XLIX [214]	2015 cleaned maps, EB+TB	$\alpha + \beta$	0.31 ± 0.05 (stat) ± 0.28 (syst)
Minami & Komatsu, $\alpha_\nu = 0$ [62]	PR3 HFI channels	$\alpha + \beta$	0.289 ± 0.048
Minami & Komatsu [62]	PR3 HFI channels	β	0.35 ± 0.14
Diego-Palazuelos et al. [79]	PR4 HFI, $f_{\rm sky} = 0.93$	β	0.30 ± 0.11
Eskilt & Komatsu [80]	PR4 LFI+HFI, WMAP	β	$0.342^{+0.094}_{-0.091}$
SMICA, here	PR3 SMICA halves, EB+TB	$\alpha_{\rm eff} + \beta$	0.31 ± 0.05 (stat)

Table 13.12: The rotation of the CMB polarization in the literature and here. Errors are 68% intervals. The first, second and last rows can only measure the sum of the cosmic angle and the miscalibration; the middle rows use the Galactic foreground to calibrate α_ν and quote β alone.

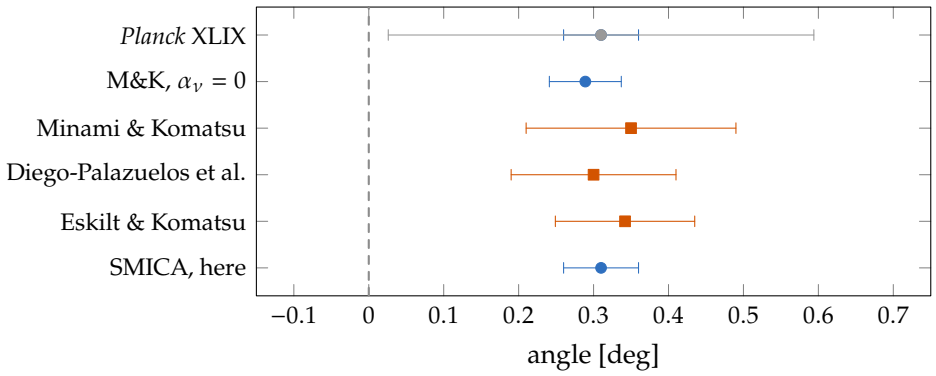


Figure 13.62: The angles of Table 13.12. Blue circles measure $\alpha + \beta$ with their statistical errors; the grey bar on the top row adds the 0.28° uncertainty of the ground calibration in quadrature. Orange squares are measurements of β alone, calibrated on the Galactic foreground.

13.33.5 Why they differ

Each row of Table 13.12 is a different estimate of a different combination from a different cut of the data, so agreement is expected only within the stated causes.

- *Different angles are measured.* Our number and the first two rows are $\alpha + \beta$, and the miscalibration in each is that of a different combination of detectors: the SMICA B-mode weights for us, the cleaning weights of 2015 for *Planck* XLIX, a single α forced on all four channels for the second row. Minami and Komatsu found channel angles between −0.28° and +0.07° [62, Table I], so the effective α of a combination can differ from one combination to the next by a tenth of a degree.
- *Different noise.* Our error is set by the half-mission noise of the SMICA map, 91 μK arcmin per half (Section 13.30.4). The frequency-channel analyses cross many more maps, but each channel is noisier than the combination and the foreground must be fitted at the same time; the end result is a statistical error on $\alpha + \beta$ of the same size as ours, and an error on β alone about three times larger, the price of separating the two angles.
- *The same sky.* All rows observe the same CMB, so their CMB fluctuations are shared and their differences are smaller than either error would suggest; the rows differ mostly through noise, sky fraction and foreground treatment.
- *What we left out.* We correct neither the temperature-to-polarization leakage of the beams nor the dust's own EB, and our noise model is Gaussian and approximate. Minami and Komatsu found the beam leakage to change β by less than 0.01° [62]; the dust EB is the leading open question, which Diego-Palazuelos et al. saw as an angle that drifts with the sky fraction [79], and which our strict-window check of Section 13.32.1 probes in the SMICA map.

The honest reading. Our $\hat{\beta}_{\rm SMICA} = 0.31^\circ \pm 0.05^\circ$ is a measurement of $\alpha_{\rm eff} + \beta$. To turn it into a statement about β we would subtract $\alpha_{\rm eff}$, and the only independent knowledge of it is the ground calibration, uncertain by 0.28° [214]. Added in quadrature to our statistical error, that gives about $\sqrt{0.05^2 + 0.28^2}$ degrees for β, and no claim about cosmic birefringence survives it. What the measurement does show is that the SMICA map contains a rotation of the size the frequency analyses attribute to β, and that a careful pipeline on a single public map finds it.

What the foreground method would need. Separating β from α takes, in order: maps that still contain the Galaxy, one per frequency (or per detector set), since each has its own α_ν; cross-spectra between all of them, half-missions crossed so that noise does not bias them; a foreground bright enough in at least one channel to fix α_ν, which for *Planck* is the dust at 353 GHz; a model of the dust's own EB correlation, since a twist of the dust filaments relative to the Galactic magnetic field makes $C^{EB,\rm dust} \neq 0$ and is degenerate with α [62, 79]; and end-to-end simulations of the instrument to show that no systematic effect makes EB. None of these is in a single component-separated map. The published value of β is therefore a measurement of the CMB *relative to the dust*: it is as good as the assumption about the dust's EB.

Lesson. Reproducing the number taught us what the number is. One CMB map, however carefully analysed, measures how much its polarization is turned relative to the instrument's idea of north; only a second, nearby source of polarization seen through the same instrument can tell the instrument's error from the Universe's.

13.34 The B modes that must be there: lensing

Everything so far asked about a B signal that might not exist. There is one B signal that must exist, and the same cross-spectra can look for it. Look at a tiled floor through a sheet of old, slightly wavy window glass. The grout lines were straight, but through the glass they wobble, and a pattern that was nothing but straight lines now has little bends and twists in it. Nothing was added to the floor; the light was bent on its way to us. The CMB is seen through such a glass: the matter between us and the last scattering surface deflects the photons by a few arcminutes, in directions that change from place to place (Chapter T1). The deflection rearranges the E pattern, and a rearranged mirror-symmetric pattern is no longer purely mirror-symmetric: some of it is B [47, 69]. The standard model predicts this lensing B spectrum precisely from the E spectrum and the matter distribution; it is the BB of our fiducial CAMB spectra, which are computed with lensing switched on, and near $\ell = 980$ it is $D_\ell^{BB} \approx 0.097\,\mu\mathrm{K}^2$, hundreds of times below the E spectrum.

Where else B comes from. The classification of B-mode sources in Chapter T1 sorts them by where they are made. In the early Universe: primordial gravitational waves, which at the current limits are far below lensing at these multipoles, and a rotation of the polarization, which turns a fraction $\sin^2 2\beta \approx (2\beta)^2$ of E into B, about 10^{-4} of the E power for a third of a degree, also far below lensing. On the way: lensing. In our Galaxy: dust and synchrotron emission, which have their own B, with roughly half as much B power as E power for dust [71]. In the instrument and the analysis: leakage of E into B by the mask (Section 13.30.2), leakage of temperature into polarization by the beams, and noise. Noise adds no B to a cross-spectrum of independent halves on average, but noise correlated between the halves would. Figure 13.63 sorts these by whether they can survive in the mean of our cross-spectrum.

The questions, in words. Do the SMICA cross-spectra contain B power, after the leakage is removed? Is its amount and shape that of lensing? And how sharp a test can this map make, at its noise level?

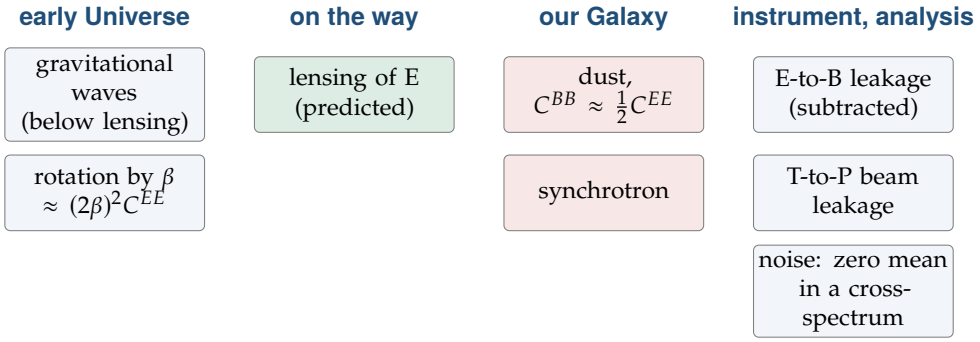


Figure 13.63: What can put B power into the mean of a half-mission cross-spectrum, sorted by where it is made. Lensing (green) is the signal we test for; the Galactic emission (red) is what SMICA tries to remove and what the confidence mask is meant to exclude; the E-to-B leakage of the mask is known from simulations and subtracted; independent noise averages out.

13.34.1 From a pseudo-BB spectrum to B power

What is in the mean of the data. In a simulated sky we kept the E part and the B part apart (Section 13.31.2), and their pseudo-coefficients add: $\widetilde{B} = \widetilde{B}[\text{E part}] + \widetilde{B}[\text{B part}] + \widetilde{B}[\text{noise}]$. The first term is the leakage. Taking the mean of the binned cross-spectrum of the halves,

$$\big\langle \widehat{D}_b^{BB} \big\rangle \overset{(1)}{=} \big\langle |\widetilde{B}[\text{E}]|^2 \big\rangle_b + \big\langle |\widetilde{B}[\text{B}]|^2 \big\rangle_b \overset{(2)}{=} L_b + F_b\, D_b^{BB}. \tag{13.81}$$

(1) Multiply out the cross-spectrum. The noise of one half has nothing to multiply in the other, the E and B parts of the sky are independent, so all cross terms have zero mean. Each half sees the same sky, so each sky term appears as its own power.

(2) Name the two pieces. L_b is the leaked E power, the mean pseudo-BB of an E-only sky. The B part through window, beam and pixels gives some multiple F_b of its true bandpower D_b^{BB}; F_b is a *transfer function*, measured as the mean pseudo-BB of the B-only skies divided by the fiducial lensing bandpower, the Monte Carlo transfer function of the MASTER method [168] in its simplest form.

So the estimator of the B bandpower, applied identically to the data and to every simulated sky, is

$$\widehat{\mathcal{B}}_b = \frac{\widehat{D}_b^{BB} - L_b}{F_b}. \tag{13.82}$$

It is unbiased if the leakage of the real sky equals that of the fiducial skies, which holds if the real E spectrum is the fiducial one; the comparison of Section 13.32.1 found them equal to about a per cent at the multipoles where the leakage matters. A spin-2 MASTER correction would do the same job with the coupling matrices of the window instead of simulated means [169]; for a single amplitude the Monte Carlo version is enough.

One number for the comparison. The lensing prediction is a known shape t_b, the binned fiducial BB. Write $\widehat{\mathcal{B}}_b = A\, t_b + \text{noise}$ and fit the one amplitude A by least squares: the algebra of (13.69) with $2\boldsymbol{D}$ replaced by $\boldsymbol{t}$ gives $\hat{A} = \boldsymbol{t}^{\mathsf{T}}\mathbf{C}_B^{-1}\widehat{\boldsymbol{\mathcal{B}}}/(\boldsymbol{t}^{\mathsf{T}}\mathbf{C}_B^{-1}\boldsymbol{t})$, with the covariance $\mathbf{C}_B$ of the 48 bandpowers from the 400 covariance skies. On the test skies with lensing, $\hat{A}$ averages 0.98 and scatters by 0.54; on the same skies with the lensing B removed it averages -0.02 and scatters by

0.54 (Figure 13.64b). That second distribution is the null: what this pipeline reports when there is no B in the sky.

What the comparison can and cannot show. The two distributions overlap. A sky with exactly the predicted lensing B gives an amplitude about 1/0.54 standard deviations from zero, a two-sigma effect at best. The reason is the noise: near $\ell = 980$ one bandpower has an error of $0.63\,\mu\mathrm{K}^2$ against a lensing signal of $0.097\,\mu\mathrm{K}^2$, and the noise power of one half is 1471 times the lensing power there. So SMICA can tell us whether the B power is far above lensing, and roughly how far; it cannot measure the shape of the lensing spectrum, and it cannot by itself detect lensing B with any confidence. (The simulated lensing B is a Gaussian field with the lensed BB spectrum; the real one is slightly non-Gaussian, which adds a little variance to $\hat{A}$, negligible beside this noise.) The experiments that measure lensing B, from the ground at a few μK arcmin, have noise twenty to thirty times lower in amplitude.

13.34.2 What the data say

The event. Through the main window the data give $\hat{A} = 7.01$, with an error of 0.54 from the test skies. None of the 200 null skies reaches it, and none of the skies *with* lensing does either (Figure 13.64). There is B power in the SMICA cross-spectra, and there is far too much of it to be lensing: the χ^2 of the bandpowers against the lensing prediction is 288.2 for 48 bins. Figure 13.64a shows where: above $\ell \approx 400$ almost every bin sits above the prediction.

What could have produced the excess? We take the candidates of Figure 13.63 one at a time.

1. *Lensing alone, larger than predicted.* An amplitude of 7.01 times the standard model's would contradict every lensing measurement made with lower noise; and lensing cannot depend on the mask, which the excess does (item 4).
2. *Noise correlated between the halves.* It would raise the cross-spectrum roughly uniformly over the sky. The half-difference map is our test of the noise model: its BB spectrum agrees with the simulated half-differences to 3% in every bin below $\ell = 1100$, and is up to 6% below them at the highest multipoles, where our smoothed noise model is slightly too high. That does not touch the cross-spectrum's mean, and item 4 argues against a sky-wide cause.
3. *The rotation itself.* It moves only $(2\beta)^2 \approx 10^{-4}$ of the E power into B, far less than what is seen.
4. *Galactic emission.* Compare the cross-spectra of the data with the mean of the simulated skies between $\ell = 800$ and 1400. Through the main window the data have an excess of $1.14 \pm 0.08\,\mu\mathrm{K}^2$ in pseudo-EE and $0.60 \pm 0.07\,\mu\mathrm{K}^2$ in pseudo-BB: about twice as much E as B, the ratio of polarized dust. Through the strict window, which also removes the region the temperature mask removes, the excesses drop to 0.21 ± 0.07 and $-0.01 \pm 0.06\,\mu\mathrm{K}^2$. The amplitude drops from 7.01 to 2.50, a change of 13.3 times the scatter of that change over the test skies, 0.34. An excess that lives in the sky near the Galactic plane and in the regions around bright sources, with twice as much E as B, is a residual of Galactic emission that the SMICA confidence mask lets through.

The strict window does not remove everything: $\hat{A} = 2.50$, with a scatter of 0.63 over the test skies, is still 2.4 of those errors above one. A stricter cut removes most of the residual, not all of it. Through the main window the excess grows with multipole, roughly as $D_\ell \propto \ell^{0.8}$ between $\ell \approx 650$ and 1250. That shape is set as much by the SMICA weights, which decide how much of each frequency channel, and so of the dust in it, survives at each multipole, as by the dust itself, and one map cannot tell the two apart.

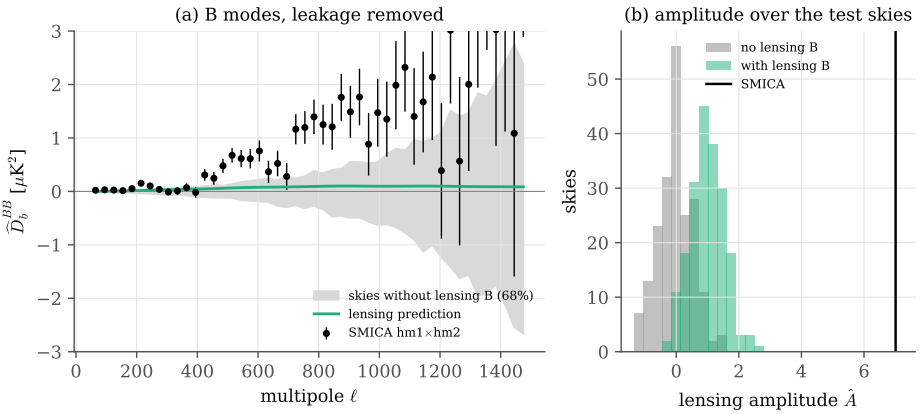


Figure 13.64: (a) The B bandpowers (13.82) of the SMICA half-missions through the main window (points, errors from the test skies), the lensing prediction (green), and the range in which 68% of the test skies without lensing B fall (grey). (b) The fitted lensing amplitude on the test skies without lensing (grey) and with lensing (green), and on the data (black line), far outside both.

What this means for the angle. The excess B is parity-even and sits in EE and BB. If it carried no EB of its own, it could not create an angle, and the angle did not move when the strict window removed most of the excess (Section 13.32.1). It enters the template $\widehat{D}^{EE} - \widehat{D}^{BB}$ as the difference of its E and B parts, about half a μK^2 against an E spectrum tens of times larger.

Lensing B: variance, not bias. The same holds for lensing, and for a stronger reason. Lensing remaps the polarization pattern by a deflection that is the gradient of a potential, a mirror-symmetric operation, so a sky with no EB before lensing has none after it on average. Lensing therefore adds BB but no EB: it adds to the scatter of $\widehat{C}_\ell^{EB}$ through the B variance in the cross-spectrum example of Section 13.31.1, and nothing to its mean. The simulated skies with and without lensing confirmed it in Section 13.31.3: the mean angle is $0.000°$ with lensing and $-0.000°$ without, and the scatter is $0.0496°$ with lensing and $0.0499°$ without: equal to within a per cent, with the sign of the tiny difference set by the particular skies drawn, because the B variance is dominated by the noise. For a future experiment with noise below the lensing B, the lensing variance would be the floor of the angle's error, and removing it (*delensing*) would sharpen the angle.

Lesson. A B-mode spectrum from a component-separated map is a test of the component separation before it is a test of cosmology: the one B signal guaranteed by the standard model is, at SMICA's noise level, smaller than what the cleaning leaves behind near the Galaxy.

13.35 The full analysis, and what the laptop version gives up

Every step above was scaled down so that it runs in about an hour of computing, and each reduction has a price. Table 13.13 lists them side by side with what the published analyses do.

How good is our error bar itself? The variance of the angle is estimated from the scatter of 200 test skies, so its square root, the error bar, carries a relative Monte Carlo error of $1/\sqrt{2(n-1)}$ with $n = 200$ (Section 8.13.1), a few per cent. The covariance built from 400 skies is noisy too; the Hartlap factor removes the bias of its inverse on average, and the coverage test of Section 13.31.3

step	research version	here	what it costs
maps	frequency maps per detector set, $N_{\text{side}} = 2048$	SMICA halves, $N_{\text{side}} = 512$	only $\alpha + \beta$; $\ell \leq 1489$
mask	CO, sources, Galactic cuts of several sizes	SMICA confidence mask, one strict variant	a coarse dust check
spectra	NaMaster, beam leakage matrices	pseudo-spectra of the halves, no beam leakage	small, by the published tests
covariance	analytic, or from end-to-end simulations	400 Gaussian skies, Hartlap	error bar uncertain by a few per cent
noise	simulated time streams through the pipeline	Gaussian, modulated, from the half-difference	no stripes, no non-Gaussian tails
foreground	per-channel α_ν, dust EB models	none	no separation of α and β

Table 13.13: The angle measurement at research scale and at laptop scale.

checks the combined effect directly on skies that did not build it. A research analysis would add skies until both effects were well below a per cent.

Verification and research. Two jobs were done by the same simulations. The tests at known angles *verified* two statements we had derived: that the rotation passes through the mask unchanged (Section 13.30.3), and that least squares with the right covariance gives an unbiased angle with the stated error (Section 13.31.1). The covariance of the parity-odd spectra, the distribution of the $\Delta\chi^2$ under no rotation, the scatter of the shift between two windows on one sky, and the distribution of the lensing amplitude without lensing are *research* products: no formula gives them for this mask and this noise, and the simulations are the measurement.

Recap

- A uniform rotation of the polarization turns E into B: $C^{EB} = \frac{1}{2}\sin 4\beta\,(C^{EE} - C^{BB})$ and $C^{TB} = \sin 2\beta\, C^{TE}$; for a small angle, the parity-odd spectra are 2β times a known template.
- The rotation commutes with the mask, the beam and the pixels, so the same relation holds for pseudo-spectra; E-to-B leakage adds scatter but no EB on average, and no spin-2 MASTER is needed.
- Least squares gives $\hat{\beta} = \boldsymbol{D}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{e}/(2\boldsymbol{D}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{D})$ and $\sigma_\beta = 1/(2\sqrt{\boldsymbol{D}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{D}})$, with $\mathbf{C}$ from simulated skies; on held-out skies with known angles it is unbiased. For EB alone the formula's error is right; for EB and TB together a covariance from 400 skies makes the true scatter larger by about $\sqrt{1 + p/n}$, so the scatter of the held-out skies is quoted instead.
- The SMICA half-missions give $\hat{\beta} = 0.31^\circ \pm 0.05^\circ$ from EB and TB together. It measures $\alpha + \beta$: a miscalibrated instrument and a rotating Universe make the same pattern in a CMB map. Only a foreground rotated by α alone, in per-frequency maps, separates them.
- The same spectra contain B power far above lensing ($\hat{A} = 7.01 \pm 0.54$ times the prediction through the main window, 2.50 through the strict one), with twice as much E as B in the excess and a strong dependence on the mask: a Galactic residual. At SMICA's noise, lensing alone would be a two-sigma signal. Lensing adds BB but no EB, so it adds variance, not bias, to the angle.

Reading the literature

- *The CMB power spectrum on the real sky.* The *Planck* 2018 overview, power-spectrum and parameter papers [25, 75, 186]; the MASTER method [168]; the pseudo-C_ℓ and likelihood approximations of [16, 134]; a modern pseudo-C_ℓ code and its conventions [169]; HEALPix [7]; CAMB [8].
- *Topology of the real CMB sky. Planck* 2018 isotropy and statistics [114]; Minkowski functionals on the sphere [110]; Betti numbers of the CMB [5]; the bias of an inverted simulated covariance [6].
- *Galaxy clustering on a real survey.* The BOSS DR12 catalogues and weights [276]; the systematics behind the weights [277]; the first CMASS detection [262]; the DR12 correlation-function BAO [278] and the final BOSS distances [224]; the estimator [221] and weights [223]; the first detection [42]; DESI's analysis with blinding [106]; jackknife errors [279, 280].
- *A Higgs search on collider open data.* The discovery papers [118, 124], of which the ATLAS one documents the diphoton categories, background functions and spurious-signal test in detail; the 13 TeV open-data release and its samples [121, 123]; the asymptotic formulas for q_0, $\tilde{q}_\mu$ and the Asimov data set [27] and the structure of the likelihood with constraint terms [29]; the look-elsewhere effect and upcrossings [15]; the CL_s method [30, 31].
- *Cosmic birefringence in Planck polarization.* The SMICA maps and their confidence masks [19]; the first *Planck* angle and its ground-calibration error [214]; the use of the Galactic foreground to calibrate the detector angles [76] and its application to the 2018, PR4 and combined WMAP and *Planck* data [62, 79, 80]; the dust EB caveat [72] and the polarized foregrounds [71]; spin-2 pseudo-C_ℓ [169] and pure-B estimators [215]; lensing B modes [47, 69]; the cost of a noisy simulated covariance [281].

Chapter summary

Each capstone analysis runs the full pipeline on real data: a physical model, a field, catalogue or list of events, a summary, its sampling distribution, and an inference. The CMB power spectrum, the topology of the CMB sky, the clustering of CMASS galaxies, the diphoton search and the twist of the CMB polarization differ in what is observed and what is undone, and share the last three stages (Figure 13.65). In the first three the summary is a curve or a set of numbers and the question is a parameter or a consistency test; in the Higgs search the summary is a histogram and the question is first a test (is there a peak?), then a measurement (how strong, where?), then an exclusion (where is there none?), all from one likelihood. In the birefringence analysis the summary is a pair of spectra that must vanish on average, EB and TB, and the question is how much of a known template they contain; the answer is an angle, and what that angle is an angle *of* ($\alpha + \beta$, not β) is decided by the data used, not by the estimator.

concept	operational meaning	used later in
The CMB power spectrum of the real sky		
masked, apodised map	multiply the map by a smooth mask that hides the Galaxy and point sources (Section 8.10.4)	Chapter 14
pseudo-C_ℓ and MASTER	the masked sky's spectrum is $\sum_{\ell'} M_{\ell\ell'} C_{\ell'}$; invert the binned coupling matrix (Section 8.11)	Section 13.20.1, Chapter 14
beam and pixel window	divide the spectrum by $b_\ell^2 p_\ell^2$ (Section 8.2.1)	Section 13.10.1

concept	operational meaning	used later in
simulated covariance	push Gaussian skies of the fiducial model through the same mask and noise (Chapter 8)	Section 13.12.2
reduced parameter fit	sample (A_s, n_s) with the Metropolis–Hastings sampler of Chapter 9	Chapter 14
The topology of the real CMB sky		
one sky as one draw	test the sky against the simulated distribution of the Gaussian null (Section 13.9)	Chapter 14
degrading a map	multiply $a_{\ell m}$ by $b_\ell^{\text{new}} p_\ell^{\text{new}} / (b_\ell^{\text{old}} p_\ell^{\text{old}})$ (Section 13.10.1)	
functionals on a cut sphere	area, boundary length and $V - E + F$ inside the mask (Section 13.11.1)	Chapter 14
Mahalanobis χ^2 and Hartlap	quadratic form of the statistics with a simulated covariance, its inverse rescaled (Sections 13.12.1 and 13.12.2)	Section 13.18.2
rank p-value	fraction of simulations at least as extreme, with error $\sqrt{p(1-p)/n}$ (Section 13.12.3)	Section 13.24.2, Chapter 14
Betti numbers and persistence on the sphere	$\beta_1 = \beta_0 - \chi$ on a cut sphere; H_0 diagrams compared by W_1 (Section 13.13)	Section 13.20.2
Galaxy clustering of CMASS South		
positions from (RA, Dec, z)	comoving distance of the fiducial model, then Cartesian coordinates (Section 13.15.3)	Section 13.19.3
catalogue weights	$w_{\text{tot}} = w_{\text{systot}}(w_{\text{cp}} + w_{\text{noz}} - 1)$ for galaxies, times w_{FKP}; randoms carry w_{FKP} (Section 13.16)	Chapter 14
window in pair counts	$\langle DD \rangle / \langle RR \rangle = 1 + \xi(s)$ in each thin shell (Section 13.20.1)	Chapter 14
leave-one-region-out counts	$C_{-k} = C - C(k,\cdot) - C(\cdot,k) + C(k,k)$ (Section 13.17.2)	Section 13.18
jackknife covariance	$\frac{N-1}{N} \sum_k (\hat{\boldsymbol{\xi}}_{-k} - \bar{\boldsymbol{\xi}})(\cdots)^{\mathsf{T}}$; exact for a mean, biased upward otherwise (Section 13.18)	Chapter 14
α to a distance	$D_V / r_d = \alpha\, D_V^{\text{fid}} / r_d^{\text{tmpl}}$, with h cancelling (Section 13.19.3)	Chapter 14
detection and location	$\sqrt{\Delta\chi^2}$ for existence, rise of $\chi^2(\alpha)$ for position; $\sigma_\alpha \sqrt{\Delta\chi^2}$ fixed by the bump's shape (Sections 13.19.5 and 13.19.6)	Section 13.25.2, Chapter 14
subset against parent sample	$\text{Var}(\alpha_S - \alpha_{NS}) = \sigma_S^2 - \sigma_{NS}^2$ (Section 13.19.6)	Chapter 14
A Higgs search on collider open data		
invariant mass	$m_{\gamma\gamma}^2 = 2 p_{T1} p_{T2} [\cosh \Delta\eta - \cos \Delta\phi]$ from the stored photon momenta (Section 13.21.1)	Chapter 14
weighted simulation	each selected simulated event stands for $L \sigma w_j \text{SF}_j / \sum_k w_k$ real events (Section 13.21.1)	Section 13.22.1
Crystal Ball line shape	Gaussian core joined with equal value and slope to a power-law tail; tail area $\frac{n}{\alpha(n-1)} e^{-\alpha^2/2}$ (Section 13.22.1)	Section 13.23.1
spurious signal	fit a background-only truth with candidate background plus signal; keep the families inventing less than 20% of σ_S (Section 13.22.2)	Section 13.23.1
price of flexibility	fitted background parameters keep $1 - \rho^2 = 1/(F_{\mu\mu}(\mathbf{F}^{-1})_{\mu\mu})$ of $q_{0,\text{A}}$ (Section 13.22.2)	Sections 13.23.3 and 13.27

concept	operational meaning	used later in
expected significance of a peak	$q_{0,\mathrm{A}} \approx \sum_i s_i^2/b_i \approx S^2/(2\sqrt{\pi}\sigma\beta)$ (Section 13.23.3)	Section 13.27, Chapter 14
local and global significance	$Z_{\mathrm{loc}} = \sqrt{q_0}$; $p_{\mathrm{glob}} \approx p_{\mathrm{loc}} + \mathbb{E}[N(c_0)]e^{-(c-c_0)/2}$ from 100 toys (Sections 13.23.2 and 13.24.2)	Chapter 14
profile-likelihood measurement	crossings of $-2\ln\lambda = 1$; systematic part $\sqrt{\sigma_{\mathrm{tot}}^2 - \sigma_{\mathrm{stat}}^2}$ (Section 13.25.1)	Chapter 14
mass precision	$\sigma_m = \sqrt{2}\,\sigma/Z$ for a Gaussian peak (Section 13.25.2)	Section 13.27
categories	q_0 adds over independent spectra; splitting gains $(S_1B_2 - S_2B_1)^2/(B_1B_2(B_1 + B_2))$ (Section 13.25.3)	Section 13.27
CL_s limits and the Brazil band	$\mathrm{CL}_{s+b}/\mathrm{CL}_b < 0.05$; band $\sigma_{\mathrm{A}}[N + \Phi^{-1}(1 - \alpha\Phi(N))]$ from the background-only Asimov set (Section 13.26)	Chapter 14
Cosmic birefringence in Planck polarization		
spin-2 degrade	zero the masked Q, U, transform to E, B, swap the polarization pixel windows, synthesise (Section 13.29.1)	Chapter 14
E-to-B leakage	a cut E pattern is read partly as B along every edge (Section 10.17); measured on E-only skies and subtracted (Section 13.30.2)	Section 13.34.1
rotation commutes with the window	$We^{2i\beta}(Q + iU) = e^{2i\beta}W(Q + iU)$, so pseudo-spectra obey the rotation formulas; no spin-2 MASTER for the angle (Section 13.30.3)	Section 13.31.2
template fit for one angle	$\hat{\beta} = \boldsymbol{D}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{e}/(2\boldsymbol{D}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{D})$ with measured $EE - BB$ and TE as template (Section 13.31.1)	Chapter 14
noisy covariance, inflated error	a simulated $\mathbf{C}$ from n skies raises the scatter by $\sqrt{1 + p/n}$; quote the scatter of held-out skies (Section 13.31.3)	Chapter 14
α–β degeneracy	a cleaned map measures $\alpha + \beta$; only a foreground rotated by α alone separates them (Section 13.33.2)	Chapter 14
lensing-B amplitude	subtract leakage, divide by a Monte Carlo transfer function, fit A to the lensing template; null from skies without lensing B (Section 13.34.1)	Chapter 14

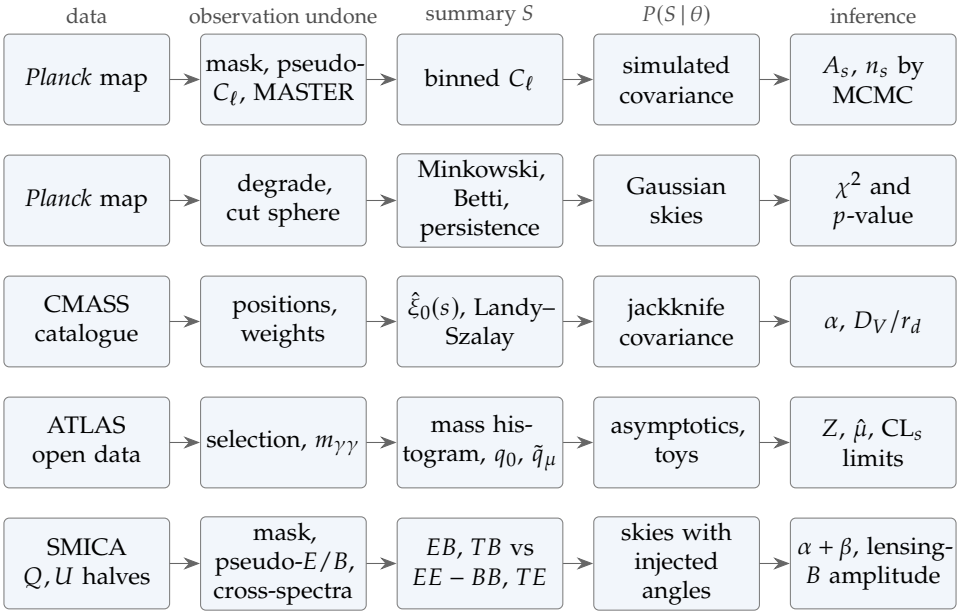


Figure 13.65: The five capstone analyses as rows of one pipeline: the CMB power spectrum (top), the topology of the CMB sky, the clustering of CMASS galaxies, the diphoton Higgs search, and the rotation of the CMB polarization (bottom). Each column is one stage of the pipeline.

Chapter 14

Synthesis: Which Idea, When

Where this chapter is going

A bump in a mass histogram raises two distinct questions. How often would background alone produce a bump this large, and what signal strength is consistent with the observed counts? Both start from the same data, yet the first needs the sampling distribution of a test statistic under the background model, while the second needs the dependence of the likelihood on signal strength. The same distinction recurs from a coin toss to the *Planck* sky. Stating the question first helps us decide whether to test a fluctuation, estimate an effect, compare models or forecast a future measurement. We can then trace the calculation from the physical model through the noise and the chosen summary to the inference. Comparing a collider, a CMB satellite, a gravitational-wave detector and a galaxy survey makes clear which steps they share and where using a quantity that answers a different question can lead us astray.

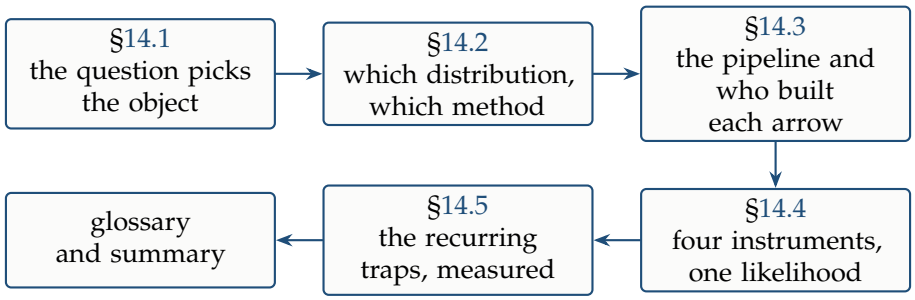


Figure 14.1: From question to inference. The top row is about choosing: the question, then the distribution and the method, then their place in the pipeline. The bottom row tests the choices: the same pipeline in four instruments, and the traps that appear when a quantity built for one question is used to answer another.

14.1 The question picks the object

A colleague at a collider stops us in the corridor. In a narrow window of invariant mass the background model predicts $b = 4$ events on average, and the detector recorded $n = 11$. "Is that something?" The honest reply is another question: *what exactly do you want to know?* Whether a fluctuation of the background could easily give eleven events? How large the signal is, and how uncertain? Which signal strengths are still believable? How much more data a discovery would need? Each of these is a different question about the same number, each has its own answer, and the answers are different numbers. Every one of them was built somewhere in the book. The danger is never a lack of tools. It is answering one question with the object built for another.

Two directions, once more. All the objects fall into two families, and the family tells us what an object can claim. In the forward direction, prediction, we fix a hypothesis or a parameter value and ask what data it produces and how often (Section 0.1). The probability mass function of a count, the sampling distribution of an estimator, the p-value and the Fisher forecast all live here: each one starts from an assumed truth. In the backward direction, inference, the data are already in hand and we ask which hypotheses could have produced them, and how probable each one is in the light of what was seen (Section 5.1). The likelihood read as a function of the parameters, the posterior and the evidence live here. The habit that keeps the two apart is to say, before any symbol, which possible hypotheses could have produced the observation, and then how probable the observation would have been under each one (Section 5.6). Cowan puts the same split in terms of what a probability means [Cowan §1.2, pp. 4–6]. A forward probability can always be read as a long-run frequency over repeated experiments, which suits colliders, where collisions really are repeated. A statement such as "the electron mass lies in this interval with probability 95%" cannot: the mass is a constant, always inside or always outside, and the probability describes our state of knowledge. Every object of the backward direction needs that second reading, and with it a prior.

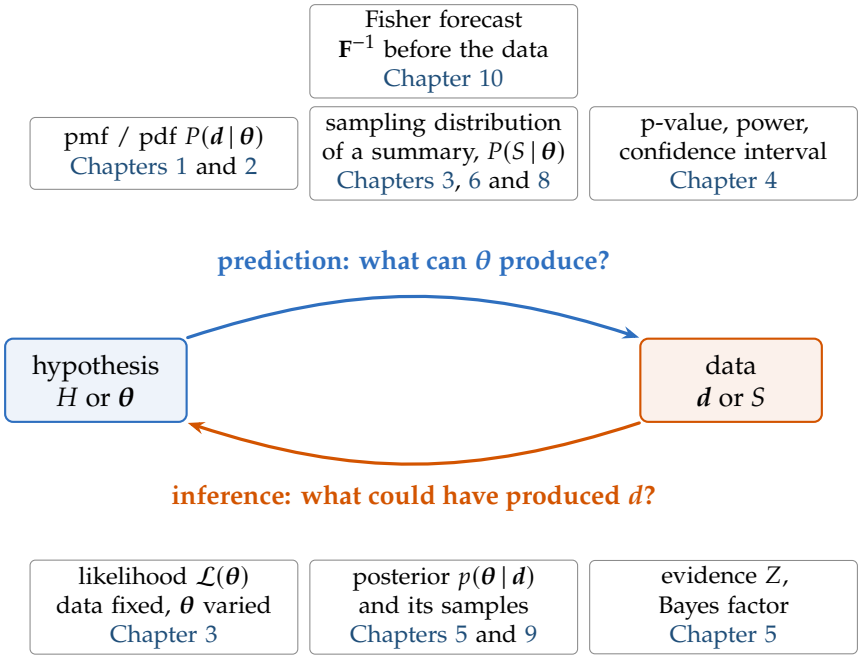


Figure 14.2: The objects of the book sorted by direction. Above, the forward direction: each object starts from an assumed $\boldsymbol{\theta}$ and describes the data it would produce; the confidence interval belongs here because it is built from such predictions and then inverted. Below, the backward direction: each object starts from the data actually recorded and weighs the possible $\boldsymbol{\theta}$; all of them except the likelihood need a prior.

The questions in words. Nine questions cover everything the book computed. Eight of them take a summary of the data as given; the ninth goes upstream and asks which summary to compute in the first place. Each comes with the object that answers it and the place where that object was built.

The ninth question is different in kind. A complicated field holds more information than any single summary, and which summary is right depends on what we want to know. If the question is how much the field fluctuates on each scale, the power spectrum is the natural answer. If it is how likely the data are under a model, the likelihood is. If it is which parameter values survive, the posterior is. If it is whether a feature survives a whole class of smooth

Taxonomy: which object answers which question

the question in words	the object that answers it	direction	built in
1. If this hypothesis holds, how often does this event happen?	$\mathrm{P}(E \mid H)$ from the rules of probability, trees, counting	forward	Chapter 1
2. How is this number distributed, and why?	a distribution derived from the machine that made the number	forward	Chapter 2
3. What value do the data point to, and how far off might it be?	an estimator, its sampling distribution, the likelihood and its curvature	both	Chapter 3
4. If only chance were at work, would a result like this, or one even more extreme in the same direction, be something we would reasonably expect?	a test statistic, its null distribution, the p-value; inverted, a confidence interval, or after a search that finds nothing an upper limit (CL_s)	forward	Chapter 4
5. Which parameter values could have produced the data, and how probable is each?	the posterior, summarised by its samples and credible intervals	backward	Chapters 5 and 9
6. Which of two models better predicted the data we got?	the evidence of each, their ratio (Bayes factor), posterior odds	backward	Chapter 5
7. The algebra is too hard: can we get any of these by drawing random numbers?	Monte Carlo: simulated data sets, simulated sampling distributions, chains	either	Chapters 6, 8 and 9
8. Before the experiment exists, how well could it measure $\boldsymbol{\theta}$?	the Fisher matrix at a fiducial model, checked by a chain on mock data	forward	Chapter 10
9. Which property of the field carries the physics and survives the nuisances of observation?	a chosen summary: C_ℓ, $\xi(r)$, peaks, Minkowski functionals, Betti numbers, persistence	upstream	Chapters 7 and 12

distortions, a topological invariant may be the answer, and if it is how features appear and disappear as a threshold is lowered, persistence is (Sections 12.1 and 12.9). Topology is therefore not an extra layer added after probability and inference. It is one answer to the first question of the whole pipeline: what should we measure from the raw data at all?

Example: eleven events where four were expected. Back to the corridor. The background in the window is known, $b = 4$, and a possible signal adds a mean $s \geq 0$, so the count is Poisson with mean $s + b$ (Section 2.4). The data are one number, $n = 11$. We put eight of the nine questions to it in turn; Figure 14.3 shows four of the answers.

Could the background alone do this? This is question 4. The hypothesis whose predictions we know is "background only", $s = 0$, and the result "in the same direction or further" means eleven events or more:

$$p \overset{(1)}{=} \mathrm{P}(N \geq n \mid s = 0) \overset{(2)}{=} 1 - \sum_{k=0}^{n-1} \frac{b^k e^{-b}}{k!} \overset{(3)}{=} 0.00284, \qquad Z = \Phi^{-1}(1 - p) = 2.77. \tag{14.1}$$

(1) The p-value of Section 4.1.3: the probability, under the null, of a count at least as large as the one observed, since an excess is what a signal would produce.

(2) The complement of the probability of $0, 1, \ldots, 10$ events, with the Poisson probabilities of mean b.
(3) The sum evaluated by `scipy.stats.poisson.sf` for $b = 4$, $n = 11$.

The significance Z is the number of Gaussian standard deviations with the same one-sided tail (Section 4.9). Question 7, simulation, checks the sum: of 2×10^6 background-only experiments drawn by the script, a fraction 0.00282 ± 0.00004 gave eleven or more.

How big is the signal? This is question 3. Read the Poisson probability of the observed count as a function of s:

$$\ln \mathcal{L}(s) \overset{(1)}{=} n \ln(s+b) - (s+b) - \ln n!, \qquad \frac{\mathrm{d} \ln \mathcal{L}}{\mathrm{d}s} \overset{(2)}{=} \frac{n}{s+b} - 1, \qquad \frac{\mathrm{d}^2 \ln \mathcal{L}}{\mathrm{d}s^2} \overset{(3)}{=} -\frac{n}{(s+b)^2}. \tag{14.2}$$

(1) The logarithm of $(s+b)^n e^{-(s+b)}/n!$.
(2) Differentiate term by term; the constant $\ln n!$ drops out.
(3) Differentiate once more.

The first derivative vanishes at $\hat{s} = n - b = 7$. At that point $s + b = n$, so the curvature is $-1/n$ and the curvature error (Section 3.12) is $\hat{\sigma} = \sqrt{n} = 3.32$. The likelihood itself is not symmetric: the two values of s where $\ln \mathcal{L}$ has fallen by $\frac{1}{2}$ from its peak are 4.01 and 10.66, a little closer below $\hat{s}$ than above it (panel b of Figure 14.3).

Which signal strengths could have produced eleven events? This is question 5, and it needs a prior. With a flat prior on $s \geq 0$ the posterior is the likelihood cut at $s = 0$ and normalised (Section 5.7). Its median is 7.67, its mean 8.01, and 68% of its probability lies between 4.61 and 11.40. These numbers are statements about s given the one count we have. The likelihood interval of the previous paragraph is a statement about a procedure over repeated counts, and the two need not coincide (Section 14.5).

Is there a signal at all? This is question 6, and it asks the data to choose between two models. Model H_0 says $s = 0$. Model H_1 says the signal lies somewhere in $[0, s_{\max}]$ with $s_{\max} = 20$, every value equally likely beforehand. Each model's evidence is the probability it gave to the count actually seen (Section 5.15):

$$B_{10} \overset{(1)}{=} \frac{Z_1}{Z_0} \overset{(2)}{=} \frac{\frac{1}{s_{\max}} \int_0^{s_{\max}} \mathrm{P}(n \mid s+b)\, \mathrm{d}s}{\mathrm{P}(n \mid b)} \overset{(3)}{=} \frac{0.04983}{0.00192} = 25.9, \tag{14.3}$$

(1) The Bayes factor is the ratio of the two evidences.
(2) Z_0 is the Poisson probability of n with mean b; Z_1 averages the Poisson probability over the uniform prior of height $1/s_{\max}$.
(3) The integral done numerically by the script.

How large could B_{10} have been with a cleverer prior? An average never exceeds the maximum of what is averaged, so $Z_1 \leq \mathcal{L}(\hat{s})$ for every prior, and

$$B_{10} \overset{(1)}{\leq} \frac{\mathcal{L}(\hat{s})}{\mathcal{L}(0)} \overset{(2)}{=} \left(\frac{n}{b}\right)^n e^{-(n-b)} \overset{(3)}{=} 62. \tag{14.4}$$

(1) $Z_1 = \int \mathcal{L}(s)\pi(s)\,\mathrm{d}s \leq \mathcal{L}(\hat{s}) \int \pi(s)\,\mathrm{d}s = \mathcal{L}(\hat{s})$, and $Z_0 = \mathcal{L}(0)$.
(2) $\mathcal{L}(s) \propto (s+b)^n e^{-(s+b)}$ evaluated at $\hat{s} = n - b$ and at 0; the factor $1/n!$ cancels.
(3) $(11/4)^{11} e^{-7}$.

The bound is reached only by a prior that is a spike at $\hat{s}$, chosen after looking (Section 5.17.4). A wider honest prior lowers Z_1: doubling $s_{\max}$ roughly halves B_{10}, because the extra prior range is spent on signals the data rule out (Section 5.17.2).

How much more data would a discovery need? This is question 8, asked before new data exist. The forecast uses the *median experiment* (Section 10.13.1): the data are replaced by their expected value under the signal hypothesis, $n \to s+b$ with $s = \hat{s}$. The likelihood-ratio statistic for testing $s = 0$ is

$$q_0 \overset{(1)}{=} -2\ln\frac{\mathcal{L}(0)}{\mathcal{L}(\hat{s})} \overset{(2)}{=} 2\Big[n\ln\frac{n}{b} - (n-b)\Big] \overset{(3)}{=} 2\Big[(s+b)\ln\Big(1+\frac{s}{b}\Big) - s\Big], \qquad Z_A = \sqrt{q_0}. \tag{14.5}$$

(1) The statistic of Section 4.7.3, whose square root is the significance.
(2) $\ln\mathcal{L}(\hat{s}) - \ln\mathcal{L}(0) = n\ln(n/b) - (n-b)$, from the logarithm in (14.2) at $s+b=n$ and at $s+b=b$.
(3) Replace the observed n by its expected value $s+b$.

This Z_A is called the Asimov significance [27]. With $s = 7$ and $b = 4$ it is 2.87. Collecting k times more data multiplies both s and b by k, and the bracket in (14.5) by k, so $Z_A \propto \sqrt{k}$. Reaching 5σ needs $k = (5/2.87)^2 = 3.0$ times the present exposure, if the signal is real and has the size we estimated (panel d).

And if nothing had been seen? Had the detector recorded four events or fewer, the useful question would have been question 4 turned around: which signal strengths does the count exclude? The answer is an upper limit on s. Collider experiments compute it with the CL_s rule, which divides the p-value of the signal hypothesis by that of the background alone, so that a downward fluctuation of the background cannot exclude a signal the experiment was never sensitive to (Section 4.16).

What shape does it have? The ninth question has no meaning for a single number. It needs a field, and it was asked of simulated maps in Chapter 12 and of the real sky in Section 13.9.

What the example shows. None of the eight numbers is wrong, and none can replace another. The p-value of 0.00284 is not the probability that the background-only hypothesis is true; the Bayes factor of 25.9 depends on a prior range that the p-value never needed; the curvature error ±3.32 is a Gaussian summary of a likelihood that is lopsided; the forecast of 3.0 times the data assumes a signal that the data have only hinted at. Every trap collected in Section 14.5 is one of these objects asked to answer a question that belongs to another.

Key idea: the question first

Say the question in words before computing anything. The question fixes the direction (prediction or inference), and the direction fixes the object: a distribution, a p-value, a likelihood, a posterior, an evidence or a forecast. A number is only as meaningful as the match between the question asked and the object computed.

14.2 Which distribution, which method

Two choices precede an inference. First we use the measurement process to specify the distribution of the observable at fixed parameters. Evaluating that distribution at the observed data gives the likelihood as a function of the parameters. We then choose how to use the model according to the question we want to answer, whether it calls for a forecast, a test, an interval, a posterior or a model comparison. The two decision charts below make those choices explicit, so that the shape of a histogram or familiarity with a particular method does not decide the analysis for us.

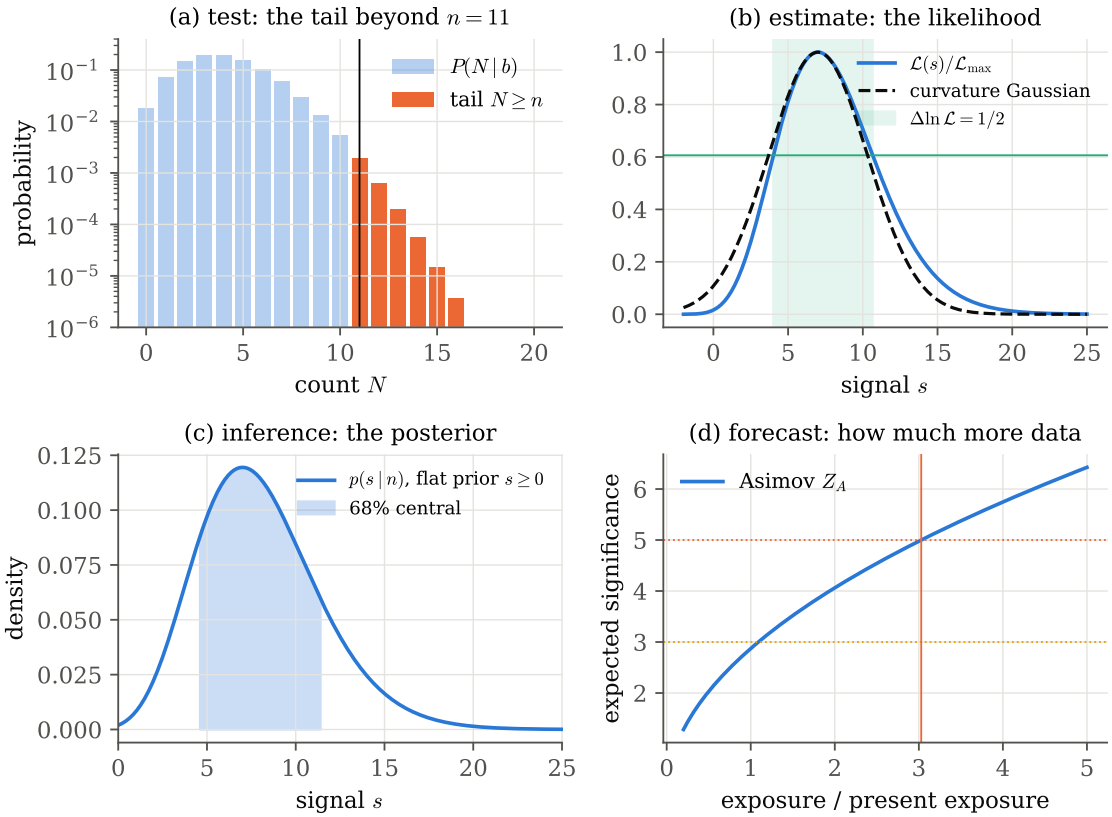


Figure 14.3: One count, four objects. (a) The prediction of the background alone; the orange bars are the outcomes at least as large as the eleven observed, and their total is the p-value. (b) The same Poisson probability read as a function of the signal s; the dashed Gaussian has the curvature at the peak, and the green band is where $\ln \mathcal{L}$ is within $\frac{1}{2}$ of its maximum. (c) The posterior for s with a flat prior on $s \geq 0$, with its central 68%. (d) The expected significance as the exposure grows; the vertical line marks the exposure that would give 5σ.

14.2.1 Which distribution: ask what made the number

Section 2.16 gave the rule: the question "which distribution?" is really "which machine made these numbers?", and the histogram comes afterwards, to check the answer in the tails, where the conclusion of a search rests. The machines met since then add a few branches to the family trees drawn there: estimated spectra, quadratic forms with simulated covariances, and summaries so elaborate that no machine can be written down. Figure 14.4 arranges all of them as a chart. Start at the left with the question *how was the number made?*, follow the branch that describes the operation, and read off the distribution.

Three recognitions show the chart at work. Each one is a number the book met, read off the chart and then checked with a short calculation.

Example: the matched filter with an unknown phase. A gravitational-wave search correlates the detector output with a template whose phase is unknown, so it uses two templates a quarter cycle apart and combines their outputs z_c and z_s (Section 10.8.3). In pure noise each output is a standard normal, and the two are independent. *How was the number made?* The detection statistic $\rho^2 = z_c^2 + z_s^2$ is the power of a complex Gaussian, two squared Gaussians added. The chart gives χ^2_2, whose density is $\frac{1}{2}e^{-x/2}$ (Section 2.13). The false-alarm probability above a

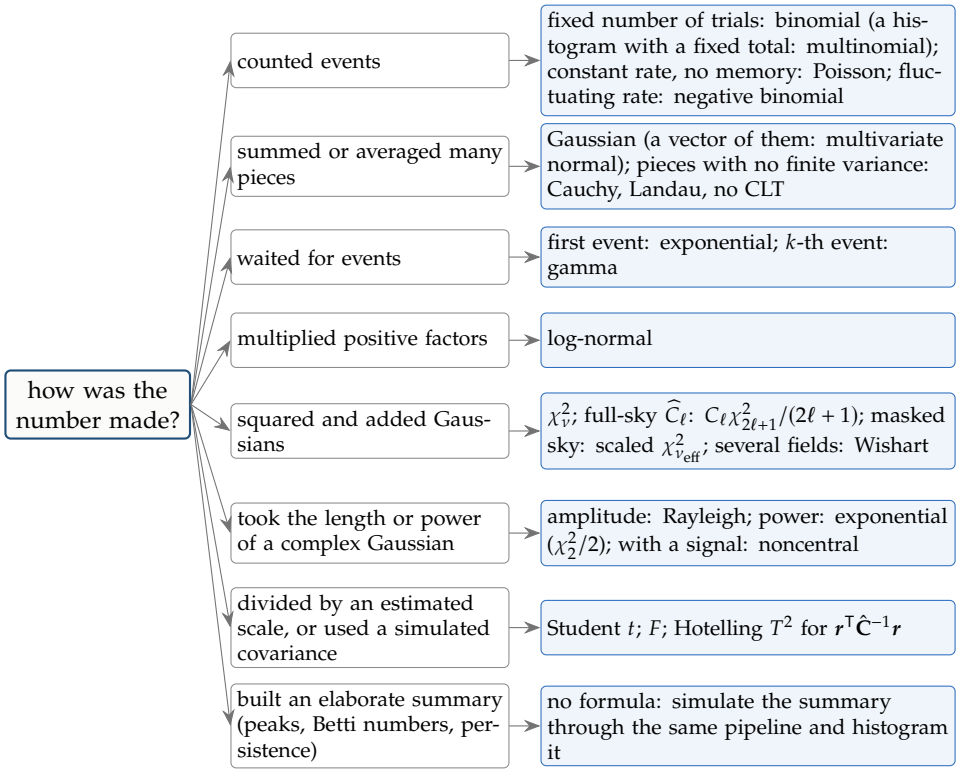


Figure 14.4: Which distribution, chosen by the operation that produced the number. The middle column names the operation, the right column the distribution it produces, with the variants that apply when the conditions change. The last row is where the book ended up most often: summaries for which no machine can be written, whose distribution is measured by simulation. Whatever the branch, the tails are checked separately, because that is where tests and discoveries live.

threshold x_* is then

$$\mathrm{P}(\rho^2 > x_* \,|\, \text{noise}) \overset{(1)}{=} \int_{x_*}^{\infty} \tfrac{1}{2} e^{-x/2}\, \mathrm{d}x \overset{(2)}{=} \left[-e^{-x/2}\right]_{x_*}^{\infty} \overset{(3)}{=} e^{-x_*/2}. \tag{14.6}$$

(1) The tail of the χ^2_2 density.
(2) The antiderivative of $\frac{1}{2}e^{-x/2}$ is $-e^{-x/2}$.
(3) At the upper limit the exponential vanishes.

A false-alarm probability of 10^{-3} per template therefore needs $\rho_* = \sqrt{2\ln 10^3} = 3.72$, not the 3.09 of a one-sided Gaussian: the unknown phase is a second chance for the noise to look like a signal.

Example: a χ^2 whose covariance came from simulations. In Section 13.12 thirty-six Minkowski-functional values of the sky were combined into one number $\boldsymbol{r}^{\mathsf{T}}\hat{\mathbf{C}}^{-1}\boldsymbol{r}$, with $\boldsymbol{r}$ the difference from the simulated mean and $\hat{\mathbf{C}}$ estimated from simulated skies. The chart sends this number not to χ^2_{36} but to the branch "used a simulated covariance". Its mean is $p(n-1)/(n-p-2)$ times too large for p bins and n simulations (Section 8.13.2), and even after that is corrected its tail is the Hotelling law, a scaled F distribution, rather than χ^2_p (Section 12.24.2). Section 14.5.5 measures the size of the error.

Example: a bandpower on a cut sky. On the full sky the estimate $\widehat{C}_\ell$ at one multipole is $C_\ell \chi^2_{2\ell+1}/(2\ell+1)$ (Section 7.10). With a mask the multipoles are mixed and no exact law is known,

but the chart's branch "squared and added Gaussians" still suggests a scaled χ^2, $X = \langle X\rangle \chi^2_\nu/\nu$, with the number of degrees of freedom ν left free. It is fixed by matching the variance measured on simulations:

$$\operatorname{Var} X \overset{(1)}{=} \frac{\langle X\rangle^2}{\nu^2} \operatorname{Var} \chi^2_\nu \overset{(2)}{=} \frac{\langle X\rangle^2}{\nu^2} 2\nu \overset{(3)}{=} \frac{2\langle X\rangle^2}{\nu} \qquad \Longrightarrow \qquad \nu_{\text{eff}} = \frac{2\langle X\rangle^2}{\operatorname{Var} X}. \tag{14.7}$$

(1) A constant factor c multiplies a variance by c^2, here $c = \langle X\rangle/\nu$.
(2) The variance of χ^2_ν is 2ν (Section 2.13).
(3) Cancel one factor of ν, then solve for ν.

On the full sky this returns $\nu = 2\ell + 1$; on a cut sky ν_{eff} comes out close to $(2\ell+1)f_{\text{sky}}$ for smooth masks, and is measured, not assumed, for sharp ones (Section 8.12). The machine fixes the family, and the simulations fix the one number the machine could not.

14.2.2 Which method: ask what is known and what is wanted

With the distribution of the summary in hand, the method follows from three questions asked in order. *Do we have data yet?* If not, the only question that makes sense is how well a planned experiment could do, and the answer is a forecast. *Is the sampling distribution $P(S \mid \boldsymbol{\theta})$ known in closed form?* If not, it is built by simulation before anything else. *What do we want to know?* Whether the result could be chance, which parameter values are allowed, which of two models is better, or whether the model fits at all: four questions, four methods. Figure 14.5 draws the order.

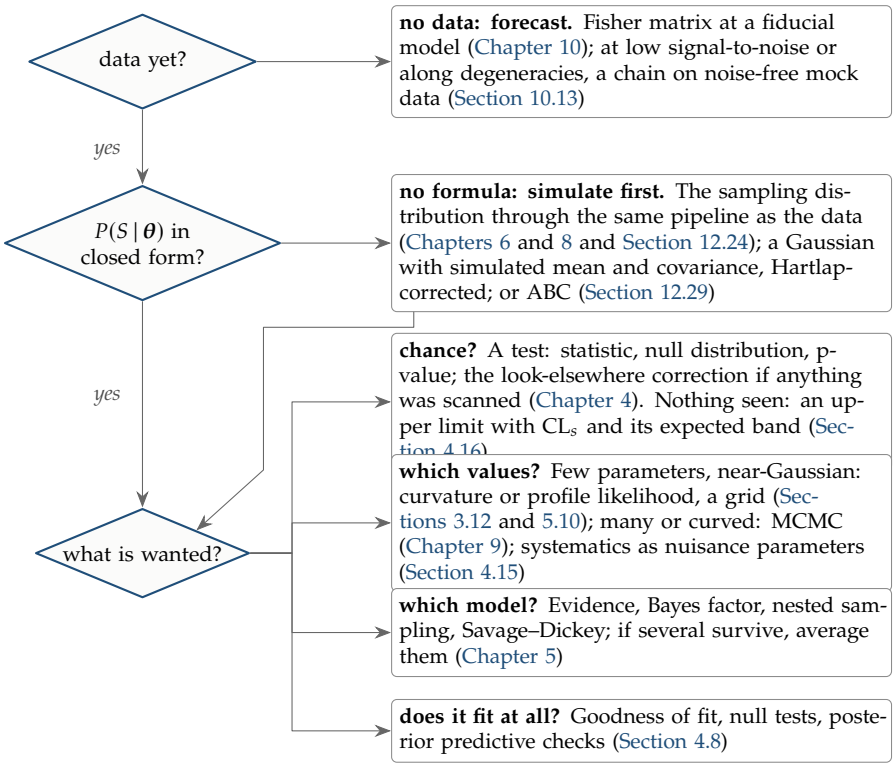


Figure 14.5: Which method, chosen by what is already known and what is wanted. A forecast needs no data; a simulated sampling distribution rejoins the main line as soon as it exists; the last decision splits by the question asked of the data, and each of the four branches has its own object.

Estimation and comparison ask different questions. Two of the four branches look alike and are not. "Which values?" and "which model?" can both be put to the question "is the

effect zero?". The estimation branch looks at the whole posterior of the effect and asks whether the values near zero are among the credible ones. The comparison branch sets up a model in which the effect is exactly zero and asks whether that model predicted the data better than a model in which it is free. Kruschke's point is that both are legitimate and that they pose the question at different levels, so they need not agree [Kruschke §12.3, pp. 352–354]. The comparison is meaningful only if an exactly zero effect is physically plausible, and only if the alternative's prior is a real statement about the effect [Kruschke §12.4, p. 354]. In physics both conditions often hold: a massless photon, $f_{\mathrm{NL}} = 0$ from single-field inflation, a dark-matter environment that either exists or does not. Where they do not hold, the estimate is the more informative answer.

Example: a Bayes factor for the null with a wide posterior. A coin gives seven heads in fourteen tosses. Model H_0 says $\theta = \frac{1}{2}$; model H_1 says θ is uniform on $[0, 1]$. The binomial coefficient $\binom{14}{7}$ is common to both evidences and cancels, so

$$B_{01} \overset{(1)}{=} \frac{(\frac{1}{2})^{14}}{\int_0^1 \theta^7 (1-\theta)^7 \, \mathrm{d}\theta} \overset{(2)}{=} \frac{2^{-14}}{7!\,7!/15!} \overset{(3)}{=} 3.14. \tag{14.8}$$

(1) Each evidence is the probability of the observed sequence under the model, averaged over the model's prior (Section 5.15).

(2) The beta integral $\int_0^1 \theta^{a-1}(1-\theta)^{b-1}\mathrm{d}\theta = (a-1)!(b-1)!/(a+b-1)!$ with $a = b = 8$ (Section 2.15).

(3) $2^{-14} = 6.10 \times 10^{-5}$ and $7!\,7!/15! = 1.94 \times 10^{-5}$.

The data favour "exactly fair" by about three to one, yet the posterior of θ under H_1 is a Beta$(8, 8)$ whose 95% highest-density interval runs from 0.27 to 0.73 [Kruschke §12.2.1.1, p. 347]. Both statements are true. The first compares two predictions of the data. The second says how precisely θ is known, and it is not known well at all. Reporting only the Bayes factor would suggest a precision the data do not have. When several models survive a comparison, the parameter posteriors can be averaged with the posterior model probabilities as weights, so that no single model is pretended to be true (Section 5.18).

Sampling distributions keep two jobs. Within the inference branches a Bayesian never uses the cloud of data sets that were not observed. Kruschke stresses that this cloud is still the right tool for two jobs [Kruschke §11.5, pp. 329–330]: planning, where the experiment has not been run and its possible outcomes are exactly what we want to know about, and checking, where data simulated from the fitted model are compared with the real data to see whether the model is any good. The forecast and "does it fit?" boxes of Figure 14.5 are those two jobs.

Four walks through the chart. *Can LISA see a dark-matter spike around an intermediate-mass black hole?* No data exist, so the chart says forecast: the Fisher matrix of the waveform phase (Section 10.11). The environment's effect is weak and nearly degenerate with the chirp mass, which is exactly the situation in which the forecast must be checked by a chain on noise-free mock data; at the signal-to-noise ratios LISA expects, the chain's widths exceeded the Fisher widths by factors of up to two to four (Section 10.13.2). *Is the topology of the Planck temperature map that of a Gaussian field?* The data exist; no formula gives the sampling distribution of Betti numbers on a masked, beamed, noisy sphere, so the chart says simulate first; the question is "could it be chance?", so the method is a test with a simulated null and a p-value counted from the simulations (Section 13.9). *What are the amplitude and tilt of the primordial spectrum?* The data exist; the bandpowers are close to Gaussian with a simulated covariance, so the likelihood is available once the covariance is simulated; the question is "which values?" with two curved,

correlated parameters, so the method is a Markov chain (Section 13.7). *Is there a new particle decaying to two photons near* 125 *GeV?* The data exist: a histogram of the invariant mass of photon pairs (Chapter T4). Each bin is a Poisson count, so the sampling distribution is known in closed form, but the shape of the background is not predicted by theory: it is a smooth function fitted to the data, and its coefficients are nuisance parameters (Section 4.15). The question is "could it be chance?", so the method is a test with the profile likelihood ratio q_0, whose distribution under the background alone is known asymptotically; because the mass was scanned, the local p-value needs the look-elsewhere correction (Section 4.14.3). At masses where nothing is seen, the same likelihood answers the other question, which signal strengths are excluded, with CL_s and the band of limits the background alone would give (Section 4.16). The search was run on public ATLAS data in Sections 13.23 and 13.26.

Key idea: two charts, two questions

The distribution is fixed by the operation that made the number, not by the look of its histogram. The method is fixed by three questions in order: are there data yet, is the sampling distribution known in closed form, and what exactly do we want to know.

14.3 The pipeline, and who built each arrow

The very first chapter drew one chain, (0.5): a physical model makes a field or a data set; we choose which information in it matters; we compress the data into a summary; we derive how the summary scatters for each value of the parameters; and only then do we infer physics. At that point the chain was a promise. Every later chapter built one or more of its arrows, and many arrows were built twice, once where a formula exists and once where only simulation can do the job. Seeing the finished chain with each chapter in its place answers a practical question: faced with a new experiment, which parts of the book does each step need?

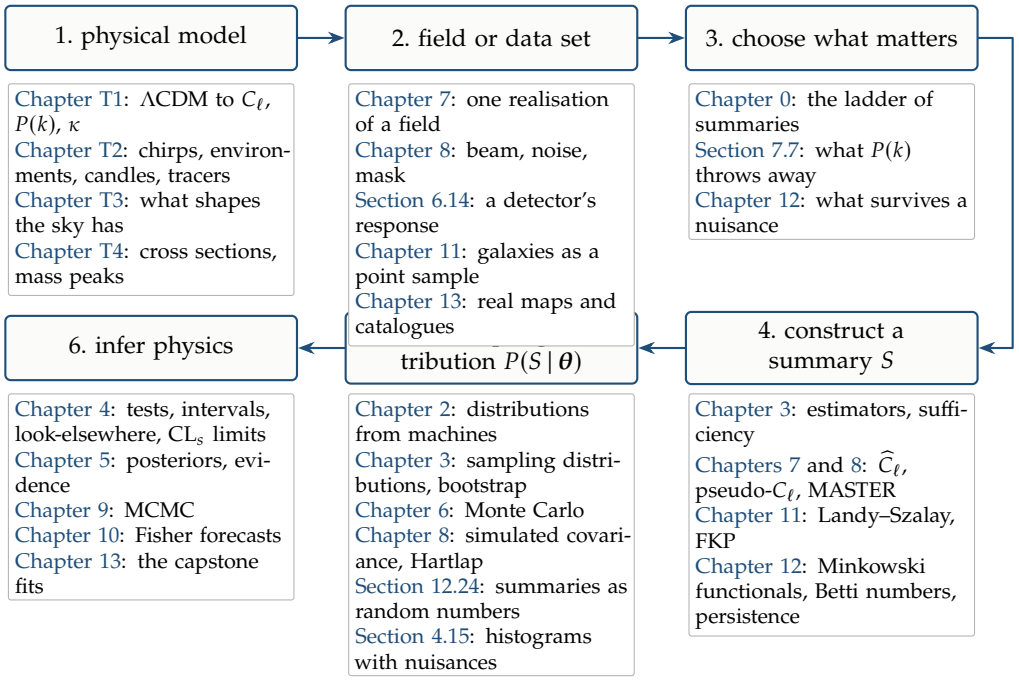


Figure 14.6: The pipeline of the book with the chapters that built each stage. The probability rules of Chapter 1 and the Python tools of Chapter T0 enter throughout. The first three stages establish which quantity to measure, and the last three use its behaviour under repetition to interpret the observed value.

Reading the chain from left to right. The first stage is physics, and in the book it was borrowed: the cosmological model turns six numbers into the expected C_ℓ and $P(k)$ (Section T1.1), and a compact binary turns masses and an environment into a phase $\Psi(f)$ (Section T2.5), and a collision model turns a particle's mass and couplings into a cross section, a number of expected events and a peak in the invariant mass (Chapter T4). A model never predicts the data themselves, only their statistics, because the data are one draw from a random process (Section 7.1). The second stage is the instrument and the survey: what the beam smooths, where the noise is added, which part of the sky is hidden, which galaxies made it into the catalogue (Sections 8.2 and 13.16), and how a particle detector smears and loses the energies it records. That last effect is described by a response matrix measured on simulated events, and it can either be folded into the prediction or, with care, undone in the data (Section 6.14). The third stage is a decision rather than a calculation: which property of the field should carry the physics. The power spectrum is the natural choice for a Gaussian field, because it is then a sufficient summary (Section 3.16); for a field made non-Gaussian by gravity, or for a question about robustness, shapes and topology become candidates (Section 12.1).

The fourth stage turns the decision into an estimator. The fifth is where the probability of the book does most of its work: the sampling distribution $P(S \mid \boldsymbol{\theta})$. For a few summaries it has a closed form, the $\chi^2_{2\ell+1}$ of a full-sky $\widehat{C}_\ell$, the Gaussian of a matched filter in Gaussian noise, or the product of Poisson probabilities of a histogram of collision events. In the last case the predicted signal and background shapes are themselves uncertain, and each uncertainty enters as a nuisance parameter that is profiled or integrated away (Section 4.15). For most of the summaries a research analysis actually uses, the masked pseudo-C_ℓ, the jackknifed correlation function, the Betti numbers of a noisy map, it is measured by sending simulated data through exactly the same code as the real data. The sixth stage reads that distribution backwards. Tests and intervals do it in the frequentist way (Chapter 4), and a search that finds nothing turns the same test into an upper limit on the signal, protected by the CL_s construction against excluding a signal to which the experiment was not sensitive (Section 4.16); posteriors in the Bayesian way (Chapter 5), with chains when the parameters are many (Chapter 9); and forecasts do it before the data exist, by replacing the data with their expected value (Chapter 10).

Where formulas end and simulations begin. The same summary can be on either side of that line depending on what the instrument did to the data. On the full sky the mean, the variance and the shape of $\widehat{C}_\ell$ all have formulas, and simulations only verify them. With a mask, the mean can still be undone with a coupling matrix computed from the mask alone, but the covariance and the shape are measured on simulations (Section 8.12). That is the line between verification and research drawn in Section 8.8, and Figure 14.7 draws it for the summaries of the book.

Example: the whole chain for one CMB amplitude. How precisely could a perfect CMB experiment measure the overall amplitude A of the temperature spectrum, $C_\ell = A\, C_\ell^{\mathrm{fid}}$? Follow the chain. The model gives C_ℓ^{fid} (stage 1). The sky is one realisation, seen on a fraction f_{sky} of the sphere up to ℓ_{max} without noise (stage 2). For a Gaussian sky the spectrum holds everything (stage 3), estimated by $\widehat{C}_\ell$ (stage 4), whose scatter is cosmic variance (stage 5). The forecast of stage 6 is the CMB Fisher matrix of Section 10.6.1, here with one parameter and no noise:

$$F_{AA} \overset{(1)}{=} \sum_{\ell=2}^{\ell_{\mathrm{max}}} \frac{(2\ell+1) f_{\mathrm{sky}}}{2} \left(\frac{\partial C_\ell / \partial A}{C_\ell} \right)^2 \overset{(2)}{=} \frac{f_{\mathrm{sky}}}{2A^2} \sum_{\ell=2}^{\ell_{\mathrm{max}}} (2\ell+1) \overset{(3)}{=} \frac{f_{\mathrm{sky}} \left[(\ell_{\mathrm{max}}+1)^2 - 4 \right]}{2A^2} \equiv \frac{N_{\mathrm{modes}}}{2A^2}, \quad (14.9)$$

(1) The CMB Fisher matrix with $N_\ell = 0$, for the single parameter A.
(2) $\partial C_\ell / \partial A = C_\ell^{\mathrm{fid}} = C_\ell / A$, so each term of the bracket is $1/A^2$.

	mean	covariance	shape
counts in a bin (Poisson)	formula	formula	formula
full-sky $\widehat{C}_\ell$	formula	formula	formula
matched filter in Gaussian noise	formula	formula	formula
masked pseudo-C_ℓ (MASTER)	formula	simulated	simulated
galaxy $\hat{\xi}(s)$ with a survey window	formula	simulated	simulated
χ of a Gaussian field, no mask	formula	simulated	simulated
Betti numbers of a masked, noisy sky	simulated	simulated	simulated
peaks of a lognormal mass map	simulated	simulated	simulated

Figure 14.7: For each summary of the book, whether the mean, the covariance and the shape of its sampling distribution came from a formula (blue) or had to be measured on simulations (orange). Going down the list, the observation and the summary become more realistic, and the simulations take over, first for the scatter and then for everything.

(3) $\sum_{\ell=0}^{L}(2\ell+1) = (L+1)^2$, the number of $a_{\ell m}$ up to L, minus the $1+3=4$ modes of $\ell = 0, 1$.

so $\sigma_A/A = \sqrt{2/N_{\text{modes}}}$: each independent mode measures the variance once, and a variance measured from N Gaussian numbers has a relative error $\sqrt{2/N}$ (Section 3.4.3). For $\ell_{\text{max}} = 2000$ and $f_{\text{sky}} = 0.6$, $N_{\text{modes}} = 0.6 \times (2001^2 - 4) \approx 2.4 \times 10^6$ and $\sigma_A/A \approx 9 \times 10^{-4}$. *Planck* measures $\ln(10^{10}A_s)$ to 0.014, an error of 1.4% on A_s [25], fifteen times larger. The difference is not noise: the temperature spectrum measures only the combination $A_s e^{-2\tau}$, and the error on the optical depth τ is passed on to A_s (Sections T1.3 and 9.22). Stage 6 is where degeneracies between parameters, not the count of modes, set the final error.

Example: the chain for a gravitational-wave amplitude. The detector records strain plus Gaussian noise with power spectral density $S_n(f)$ (stages 1 and 2, Section T2.7). Because the noise is Gaussian and its spectrum known, the data themselves are a sufficient summary and the likelihood is exact: $\mathcal{L} \propto e^{-(d-h|d-h)/2}$ with the noise-weighted inner product of Section 10.8 (stages 3 to 5). For the overall amplitude, $h = \mathcal{A}\, g$ with g fixed, the Fisher matrix of Section 10.9 gives

$$\Gamma_{\ln\mathcal{A}\,\ln\mathcal{A}} \overset{(1)}{=} \left(\frac{\partial h}{\partial \ln\mathcal{A}}\middle|\frac{\partial h}{\partial \ln\mathcal{A}}\right) \overset{(2)}{=} (h|h) \overset{(3)}{=} \rho^2, \qquad \sigma_{\ln\mathcal{A}} = \frac{1}{\rho}, \tag{14.10}$$

(1) The Gaussian-noise Fisher matrix, $\Gamma_{ij} = (\partial_i h|\partial_j h)$.
(2) $\partial h/\partial \ln\mathcal{A} = \mathcal{A}\,\partial h/\partial\mathcal{A} = \mathcal{A}g = h$.
(3) The squared signal-to-noise ratio is $\rho^2 = (h|h)$.

A source at $\rho = 20$ has its amplitude measured to 5%, independent of everything else about the waveform. Here the CMB example's lesson is reversed: the amplitude is the easy parameter, and the hard ones are those whose effect on the phase other parameters can imitate (Section 10.9.3).

Example: the chain for the topology of the sky. The model is a Gaussian ΛCDM sky, possibly with a small f_{NL} (stage 1, Section T3.2); the data are a masked, beamed, noisy temperature map (stage 2). The choice of stage 3 is to look at the shape of the excursion sets at thresholds set by the area fraction, which removes any monotone distortion of the temperature values (Section 12.4). Stage 4 counts areas, boundary lengths, Euler characteristics and Betti numbers inside the mask. For an unmasked Gaussian field the mean of the Euler characteristic has a

formula (Section 12.5); for everything else in stage 5 there are only simulated skies sent through the same mask and code. Stage 6 is a test against the Gaussian null with a simulated covariance (Section 13.12). The chain is the same as for the amplitude; what moved is the line between formula and simulation in Figure 14.7.

What the chain buys us. A new problem can be placed on the chain before any calculation. Its stage 5 decides the cost: if the summary has a known sampling distribution, the analysis is a few lines; if not, the number of simulations sets the precision of everything after it (Section 8.13). Its stage 6 decides the dangers: the traps of Section 14.5 all live there.

How this is used in real research: the chain in a research collaboration

A large analysis divides the chain among teams: a theory team owns the Boltzmann or waveform code (stage 1), an instrument team the beam, noise and mask models and the end-to-end simulations (stages 2 and 5), and an analysis team the estimators and the likelihood (stages 4 and 6). The *Planck* likelihood, the LIGO–Virgo parameter estimation and the BOSS clustering analysis all have this shape. The weak points are the hand-overs: a covariance simulated with a different mask from the one in the data, or a template built at a different fiducial cosmology from the one used for the distances, each passes silently from one stage to the next.

14.4 Four instruments, one likelihood

A particle physicist scanning a diphoton spectrum, a cosmologist fitting the *Planck* temperature spectrum, a gravitational-wave astronomer estimating the masses of a binary and a survey astronomer measuring the clustering of galaxies read different journals and use different words. In Section 0.5 we saw that underneath all of them lies one question: for a trial value of the parameters, how probable is the noise realisation that would turn the model's prediction into the data actually recorded? The likelihood is the density of that required noise (0.9). At that point the claim rested on three short formulas. Now every stage of the pipeline has been built for each instrument, and the comparison can be made stage by stage. The interesting part is no longer that the idea is shared; it is *where* the four implementations differ, because each difference is a place where a method that works for one instrument fails for another.

The question in words. For each instrument: what is the noise that the data require, what is its distribution, and how do the parameters change it? The answers fill Table 14.1. Two of its rows deserve a derivation, because they are where the columns part company: the form of the likelihood for counts and for a Gaussian field with several components, and the place where the parameters enter.

Example: counts in many bins. A spectrum of k independent bins with counts n_i and expected counts $\mu_i(\boldsymbol{\theta}) = s_i(\boldsymbol{\theta}) + b_i$ has the likelihood of Section 3.11:

$$-2\ln\mathcal{L}(\boldsymbol{\theta}) \overset{(1)}{=} -2\sum_{i=1}^{k}\ln\frac{\mu_i^{n_i}e^{-\mu_i}}{n_i!} \overset{(2)}{=} 2\sum_{i=1}^{k}\big[\mu_i - n_i\ln\mu_i\big] + 2\sum_{i=1}^{k}\ln n_i!$$
$$\overset{(3)}{=} 2\sum_{i=1}^{k}\big[\mu_i(\boldsymbol{\theta}) - n_i\ln\mu_i(\boldsymbol{\theta})\big] + \text{const.} \tag{14.11}$$

(1) Independent bins multiply, so their logarithms add.

	collider counts	**CMB**	**gravitational waves**	**LSS and topology**
data	events in bins of mass	a temperature (and polarization) map	a strain time series	a galaxy catalogue, a mass map
noise	Poisson counting	cosmic variance of a Gaussian field, plus detector noise	stationary Gaussian noise with spectrum $S_n(f)$	cosmic variance, shot noise, shape noise; non-Gaussian
summary	the counts themselves	$\widehat{C}_\ell$, pseudo-C_ℓ bandpowers	the data themselves, or the matched-filter output	$\hat{\xi}(s)$, $\hat{P}(k)$; peaks, Minkowski functionals, Betti numbers
its distribution	Poisson, exact	$\chi^2_{2\ell+1}$ (Wishart for several fields); simulated with a mask	Gaussian, exact	simulated (mocks, simulated maps)
$-2\ln\mathcal{L}$	$2\sum_i[\mu_i - n_i\ln\mu_i]$	sum over ℓ of $(2\ell+1)\big[\mathrm{Tr}(\mathbf{C}^{-1}\widehat{\mathbf{C}}) + \ln\det\mathbf{C}\big]$	$(d-h \mid d-h)$	$\boldsymbol{r}^{\mathsf{T}}\hat{\mathbf{C}}^{-1}\boldsymbol{r}$, $\boldsymbol{r} = \boldsymbol{S} - \hat{\boldsymbol{\mu}}$, Hartlap-corrected
$\boldsymbol{\theta}$ enters	the mean $\mu_i(\boldsymbol{\theta})$, with nuisance parameters for the shapes	the covariance $\mathbf{C}_\ell(\boldsymbol{\theta})$	the mean $h(\boldsymbol{\theta})$	mostly the mean; covariance fixed at a fiducial model
forecast	Asimov significance, $\sigma_s^2 = (s+b)$	Knox, CMB Fisher matrix	$\Gamma_{ij} = (\partial_i h \mid \partial_j h)$	Fisher with simulated derivatives and its noise bias
many tests	the look-elsewhere effect in a mass scan	anomalies chosen after looking	a bank of templates and arrival times	many thresholds and summaries: combine, do not pick
built in	Sections 3.11, 4.1, 4.14.3, 4.15, 4.16, 13.23 and 13.26 and Chapter T4	Chapters 7, 8 and 13 and Section 9.19	Chapter T2 and Sections 10.8 and 10.9	Chapters 11 and 12 and Section 12.24

Table 14.1: The same pipeline in four instruments. Each column is one kind of experiment and each row one stage of the analysis. The rows that differ, where the parameters enter and whether the distribution is known or simulated, decide which methods transfer from one column to another.

(2) $\ln(\mu^n e^{-\mu}/n!) = n\ln\mu - \mu - \ln n!$, multiplied by -2.
(3) The data's own term $\sum \ln n_i!$ does not depend on $\boldsymbol{\theta}$.

This is Cash's statistic [20]. The parameters enter only through the mean. Nothing in it assumes many counts, which is why it, and not a χ^2 with $\sqrt{n_i}$ errors, is right for the sparse tails where new particles are sought. Its price is that the null distribution of the best bump, scanned over all masses, has no Wilks theorem behind it (Section 4.14.3). Panel (a) of Figure 14.8 shows one background-only spectrum of 40 bins scanned bin by bin. Its largest local excess, $Z = 2.95$, would be matched or beaten somewhere in 0.06 of all background-only spectra; a 3σ local excess, with local probability 0.00135, appears somewhere in 0.050 of them.

Example: a CMB multipole with temperature and polarization. At one multipole ℓ the temperature and the E-mode polarization give, for each m, a pair $\boldsymbol{a}_{\ell m} = (a^T_{\ell m}, a^E_{\ell m})$. Isotropy makes different m uncorrelated and gives every m the same 2×2 covariance

$$\mathbf{C}_\ell = \begin{pmatrix} C_\ell^{TT} & C_\ell^{TE} \\ C_\ell^{TE} & C_\ell^{EE} \end{pmatrix}$$

(Section 7.9). As in the one-field count of Section 7.10, the independent real numbers are the vector $\boldsymbol{x}_0 = \boldsymbol{a}_{\ell 0}$, which is real with covariance $\mathbf{C}_\ell$, and for each $m = 1, \ldots, \ell$ the real and

imaginary parts x_m and y_m of $a_{\ell m} = x_m + i y_m$, each with covariance $\mathbf{C}_\ell/2$; the negative m are fixed by the reality condition $a_{\ell,-m} = (-1)^m a^*_{\ell m}$. *What is the probability of the observed $a_{\ell m}$ if the spectra are $\mathbf{C}_\ell$?* It is the product of the Gaussian densities of these independent vectors, and each contributes $x^\mathsf{T}\Sigma^{-1}x + \ln\det\Sigma$ (plus a constant) to $-2\ln\mathcal{L}$:

$$
\begin{aligned}
-2\ln\mathcal{L}(\mathbf{C}_\ell) &\overset{(1)}{=} x_0^\mathsf{T}\mathbf{C}_\ell^{-1}x_0 + \ln\det\mathbf{C}_\ell + \sum_{m=1}^{\ell}\Big[x_m^\mathsf{T}\big(\tfrac{1}{2}\mathbf{C}_\ell\big)^{-1}x_m + y_m^\mathsf{T}\big(\tfrac{1}{2}\mathbf{C}_\ell\big)^{-1}y_m + 2\ln\det\big(\tfrac{1}{2}\mathbf{C}_\ell\big)\Big] + c_1 \\
&\overset{(2)}{=} x_0^\mathsf{T}\mathbf{C}_\ell^{-1}x_0 + 2\sum_{m=1}^{\ell}\big[x_m^\mathsf{T}\mathbf{C}_\ell^{-1}x_m + y_m^\mathsf{T}\mathbf{C}_\ell^{-1}y_m\big] + (2\ell+1)\ln\det\mathbf{C}_\ell + c_2 \\
&\overset{(3)}{=} \sum_{m=-\ell}^{\ell} a_{\ell m}^\dagger\mathbf{C}_\ell^{-1}a_{\ell m} + (2\ell+1)\ln\det\mathbf{C}_\ell + c_2 \\
&\overset{(4)}{=} \mathrm{Tr}\Big[\mathbf{C}_\ell^{-1}\sum_{m=-\ell}^{\ell} a_{\ell m}a_{\ell m}^\dagger\Big] + (2\ell+1)\ln\det\mathbf{C}_\ell + c_2 \\
&\overset{(5)}{=} (2\ell+1)\Big[\mathrm{Tr}\big(\mathbf{C}_\ell^{-1}\widehat{\mathbf{C}}_\ell\big) + \ln\det\mathbf{C}_\ell\Big] + c_2 .
\end{aligned}
\tag{14.12}
$$

(1) The Gaussian density of each independent real vector; c_1 collects the factors of 2π, which do not depend on $\mathbf{C}_\ell$.

(2) $(\frac{1}{2}\mathbf{C}_\ell)^{-1} = 2\mathbf{C}_\ell^{-1}$, and for a 2×2 matrix $\det(\frac{1}{2}\mathbf{C}_\ell) = \frac{1}{4}\det\mathbf{C}_\ell$, so each $m \geq 1$ gives $2\ln\det\mathbf{C}_\ell$ plus a constant; with $m = 0$ that is $1 + 2\ell$ copies of $\ln\det\mathbf{C}_\ell$.

(3) For a real symmetric matrix M, $a^\dagger M a = x^\mathsf{T}Mx + y^\mathsf{T}My$: the cross terms $ix^\mathsf{T}My - iy^\mathsf{T}Mx$ cancel because $x^\mathsf{T}My = y^\mathsf{T}Mx$. By the reality condition $a_{\ell,-m}$ gives the same value as $a_{\ell m}$, so the factor 2 is the sum over $\pm m$.

(4) A number is its own trace, and the trace is cyclic: $a^\dagger M a = \mathrm{Tr}(M a a^\dagger)$.

(5) Define the estimated spectrum matrix $\widehat{\mathbf{C}}_\ell = \frac{1}{2\ell+1}\sum_m a_{\ell m}a_{\ell m}^\dagger$, the two-field version of $\widehat{C}_\ell$; it is real because the $\pm m$ terms are complex conjugates.

For one field, $\mathrm{Tr}(\mathbf{C}^{-1}\widehat{\mathbf{C}}) + \ln\det\mathbf{C}$ becomes $\widehat{C}_\ell/C_\ell + \ln C_\ell$ and (14.12) is the exact full-sky likelihood of Section 9.19. Read the other way, as a distribution of the data $(2\ell+1)\widehat{\mathbf{C}}_\ell$ for fixed $\mathbf{C}_\ell$, it is a Wishart distribution with $2\ell+1$ degrees of freedom, the matrix version of $\chi^2_{2\ell+1}$ and the same law that governs a covariance matrix estimated from simulations (Section 8.13.2). Here the parameters enter only through the *covariance*, and at low ℓ the likelihood is far from Gaussian in C_ℓ (panel b of Figure 14.8): at $\ell = 2$ there are only 5 modes, and the likelihood has a long tail to large C_ℓ. Detector noise adds $\boldsymbol{N}_\ell$ to $\mathbf{C}_\ell$, and a mask replaces $2\ell+1$ by an effective number of modes close to $(2\ell+1)f_{\rm sky}$ (Section 9.19).

Example: a gravitational-wave signal in Gaussian noise. For a detector with stationary Gaussian noise, the noise in different frequency bins is independent, with variance set by $S_n(f)$, and the likelihood is $\mathcal{L} \propto \exp[-\frac{1}{2}(d-h|d-h)]$ with the noise-weighted inner product (Section 10.8). The parameters enter only through the *mean*, the waveform $h(\boldsymbol{\theta})$. When the phase is unknown the detection statistic is the χ^2_2 of (14.6) in noise and a noncentral χ^2_2 with a signal; at $\rho = 4$ a threshold set for a false-alarm probability of 10^{-3} ($\rho^2 > 13.82$) catches the signal with probability 0.66 (panel c).

Example: summaries with no formula. For a galaxy correlation function measured through a survey window, or for peak counts and Betti numbers of a lognormal mass map, nothing is known in closed form. The standard choice is a Gaussian likelihood whose mean $\hat{\boldsymbol{\mu}}(\boldsymbol{\theta})$ and covariance $\hat{\mathbf{C}}$ are both measured on simulations, the covariance at one fiducial model

(Sections 11.6 and 12.24). The Gaussian form is an assumption, checked by histogramming the simulated summaries, and the estimated covariance changes the distribution of the χ^2 even when the data are exactly Gaussian. Panel (d) shows 10 summaries with a covariance from 40 simulations. After the Hartlap factor the mean of the χ^2 is right, 10.02 against 10, but the tail is heavier: the 5% critical value of χ^2_{10} is exceeded in 0.075 of trials.

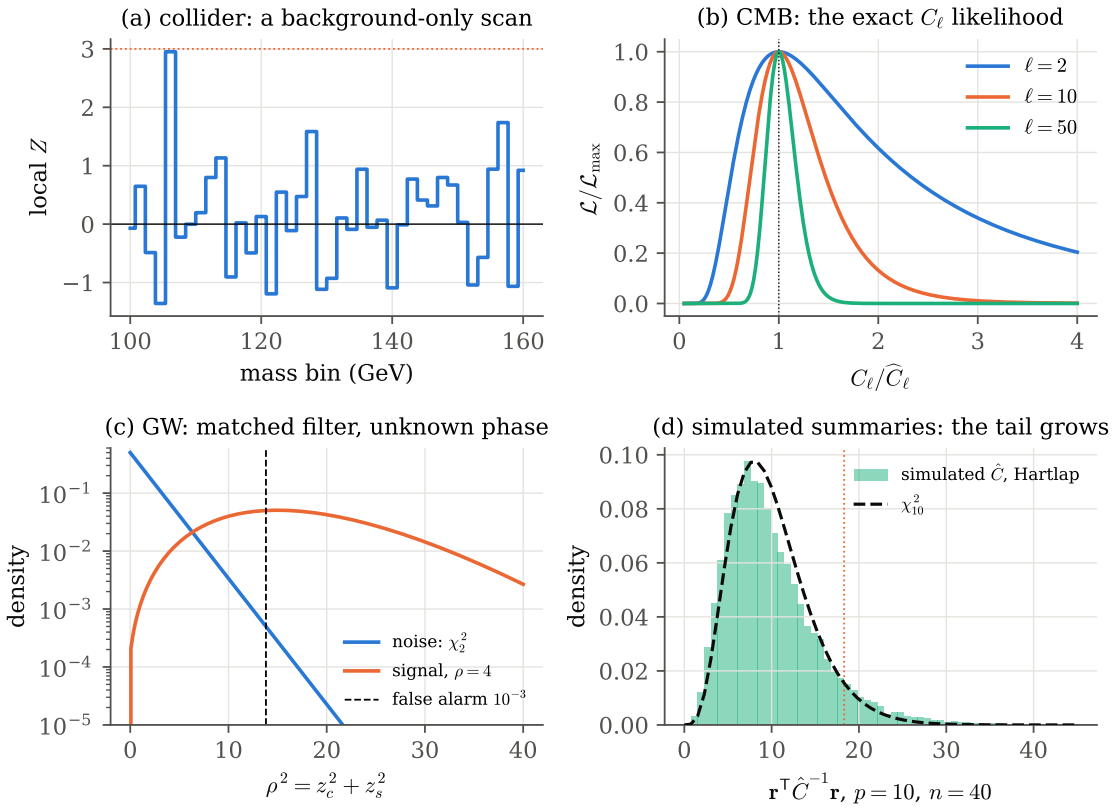


Figure 14.8: The distribution each instrument's inference rests on. (a) A background-only spectrum scanned bin by bin. When forty masses are tried, a local excess of two standard deviations somewhere is common; the largest in this spectrum is $Z = 2.95$. (b) The exact likelihood of a CMB spectrum at one multipole, when the measured $\widehat{C}_\ell$ equals the fiducial value: lopsided at $\ell = 2$, nearly Gaussian at $\ell = 50$. (c) The matched-filter statistic with an unknown phase, in noise and with a signal of $\rho = 4$; the dashed line is the threshold for a false-alarm probability of 10^{-3}. (d) A χ^2 built with a covariance estimated from forty simulations: after the Hartlap factor its mean is right but its tail is heavier than χ^2_{10}.

Where the parameters enter. All four likelihoods are, or are approximated by, a Gaussian in some data vector, with

$$-2\ln\mathcal{L}(\boldsymbol{\theta}) = \big[\boldsymbol{d} - \boldsymbol{\mu}(\boldsymbol{\theta})\big]^\mathsf{T}\mathbf{C}(\boldsymbol{\theta})^{-1}\big[\boldsymbol{d} - \boldsymbol{\mu}(\boldsymbol{\theta})\big] + \ln\det\mathbf{C}(\boldsymbol{\theta}) + \text{const},$$

the Poisson case joining through its Gaussian limit at large counts. The parameters can sit in the mean, in the covariance, or in both, and the Fisher matrix inherits the split: a mean part $\boldsymbol{\mu}_{,i}^\mathsf{T}\mathbf{C}^{-1}\boldsymbol{\mu}_{,j}$ and a covariance part $\frac{1}{2}\,\mathrm{Tr}[\mathbf{C}^{-1}\mathbf{C}_{,i}\mathbf{C}^{-1}\mathbf{C}_{,j}]$ (Section 10.3). The consequences differ in practice. When the parameters sit in the mean, the likelihood is Gaussian in the data even at low signal-to-noise, and the danger is a curved, nonlinear model; this is the GW case, where the Fisher ellipse fails because the waveform is not linear in its parameters (Section 10.13). When they sit in the covariance, the likelihood is not Gaussian in the data until many modes are averaged; this is the CMB at low ℓ, where a Gaussian approximation in C_ℓ biases the

answer (Section 9.19). When neither is known, simulations supply both, and their finite number becomes part of the error budget. Figure 14.9 collects the three cases.

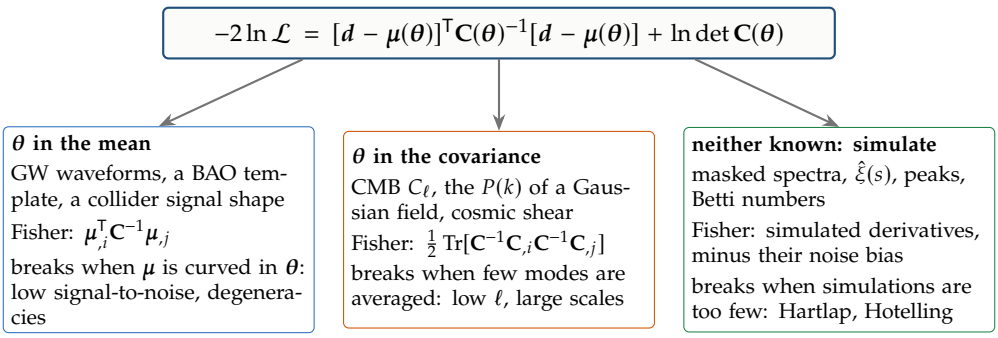


Figure 14.9: Where the parameters enter the likelihood decides how the analysis behaves. In the mean (left), the likelihood stays Gaussian in the data and the trouble is a nonlinear model. In the covariance (middle), the trouble is a non-Gaussian likelihood when few modes are available. Measured from simulations (right), the trouble is the finite number of simulations.

Which methods transfer, and which do not. The table and the three cases explain why some recipes travel and others do not. Inverting the Fisher matrix for a forecast transfers everywhere, but it is checked against a chain in the left and right columns of Figure 14.9 and against the exact likelihood at low ℓ in the middle one. A χ^2 with a known covariance is exact for the gravitational-wave matched filter and for Gaussian bandpowers at high ℓ, approximate for counts, and has the Hotelling tail for simulated summaries. The look-elsewhere correction is unavoidable in a mass scan, and has analogues wherever many places, templates or thresholds are tried. The safest habit for simulated summaries is to combine every threshold into one pre-declared statistic rather than to quote the most significant one (Section 12.25).

Key idea: same question, different noise

The likelihood of every instrument is the probability of the noise that the data require. What changes from one instrument to the next is where the parameters enter, in the mean or in the covariance, and whether the distribution of the summary is known or simulated. Those two facts decide which approximations hold and which checks are needed.

14.5 The traps that kept coming back

The same few mistakes appeared in chapter after chapter, in coins, in collider bumps, in CMB spectra and in gravitational waves. None of them is a mistake of arithmetic. Each is a number built to answer one question being read as the answer to another, exactly the mismatch the corridor example of Section 14.1 warned about. Each trap has the same anatomy: the statement as people make it, the everyday question it seems to answer, the assumption hidden inside that question, a small simulated example that breaks the assumption (the numbers come from `code/ch14/02_pitfalls.py`), and the lesson in one line.

14.5.1 A p-value is not the probability that the null is true

The statement. "We found $p = 0.05$, so there is only a 5% chance that the effect is a fluke."

The question it seems to answer. Given what we saw, how probable is it that nothing is there?

The hidden assumption. The p-value runs in the forward direction: *if* the null were true, how often would a result this extreme or more appear (Section 4.1.3)? The question runs backwards: given the result, how probable is the null? Turning one into the other is Bayes' theorem, and Bayes' theorem needs two things the p-value never used: how readily the *alternative* would have produced the same result, and how often nulls are true to begin with (Section 4.4.4). The statement silently sets the first to one and ignores the second.

An example where the assumption fails. Picture a field in which half the hypotheses tested are true nulls and the other half have real effects of varied size: the true effect μ, in units of the measurement error, is drawn from $\mathcal{N}(0, \tau^2)$ with $\tau = 2$, and each experiment reports $z = \mu + \text{noise}$ with unit Gaussian noise and a two-sided p-value. Among the experiments that landed near $p = 0.05$, what fraction were true nulls? Under the null, $z \sim \mathcal{N}(0, 1)$. Under the alternative, z is the sum of two independent Gaussians, the effect and the noise, so it is Gaussian with the variances added, $z \sim \mathcal{N}(0, 1 + \tau^2)$ (Section 2.12). The Bayes factor for the null at the observed z is the ratio of these two densities:

$$B_{01} \overset{(1)}{=} \frac{e^{-z^2/2}/\sqrt{2\pi}}{e^{-z^2/2(1+\tau^2)}/\sqrt{2\pi(1+\tau^2)}} \overset{(2)}{=} \sqrt{1+\tau^2}\,\exp\!\left[-\frac{z^2}{2}\left(1-\frac{1}{1+\tau^2}\right)\right]$$
$$\overset{(3)}{=} \sqrt{1+\tau^2}\,\exp\!\left[-\frac{z^2}{2}\,\frac{\tau^2}{1+\tau^2}\right] \overset{(4)}{=} 0.481. \tag{14.13}$$

(1) The two Gaussian densities at the observed z, variance 1 under H_0 and $1 + \tau^2$ under H_1.
(2) The $\sqrt{2\pi}$ cancels; the two exponentials combine into one.
(3) $1 - 1/(1 + \tau^2) = \tau^2/(1 + \tau^2)$.
(4) $z = 1.96$ (the two-sided $p = 0.05$) and $\tau = 2$: $\sqrt{5}\,e^{-1.92\times0.8}$.

With equal prior probabilities the posterior probability of the null is $B_{01}/(1 + B_{01}) = 0.325$. The simulation of 4×10^6 experiments agrees: of the 61737 that gave $0.045 < p < 0.055$, a fraction 0.322 ± 0.002 were nulls (Figure 14.10a). A different spread of effects gives a different number, but there is a floor. Over every alternative that is symmetric about zero and falls away from it, the most a p-value can support is the odds $\bar{B}_{10} = -1/(e\,p \ln p)$ (Section 5.17.4), which at $p = 0.05$ is 2.46, so the null keeps at least 0.289 of its probability.

The lesson. A p-value is a prediction made from the null; the probability of the null is an inference that needs the alternative and the prior, and at $p = 0.05$ it is nearer one third than one twentieth.

14.5.2 A confidence interval is not a credible interval

The statement. "The 90% interval is $[a, b]$, so the true value lies in $[a, b]$ with probability 90%."

The question it seems to answer. How sure can we be that the parameter is in this range, given the data we have?

The hidden assumption. A confidence interval is a recipe whose output covers the fixed true value in 90% of repeated experiments (Section 4.12). The probability belongs to the recipe, not to the one interval in hand. "The value lies in $[a, b]$ with probability 90%" is a statement about the parameter given these data, which is a credible interval and needs a prior (Section 4.17). The two coincide for a Gaussian mean far from any boundary with a flat prior, which is where the habit is learned.

An example where the assumption fails. A mass $\mu \geq 0$ is measured once with Gaussian error $\sigma = 1$, and the reading is $x = -1.5$, below zero by a downward fluctuation. The central 90% confidence interval is $x \pm 1.645 = [-3.145, 0.145]$. The recipe is valid: it covers the truth in 90% of repetitions, whatever μ is. But how much posterior probability does this particular

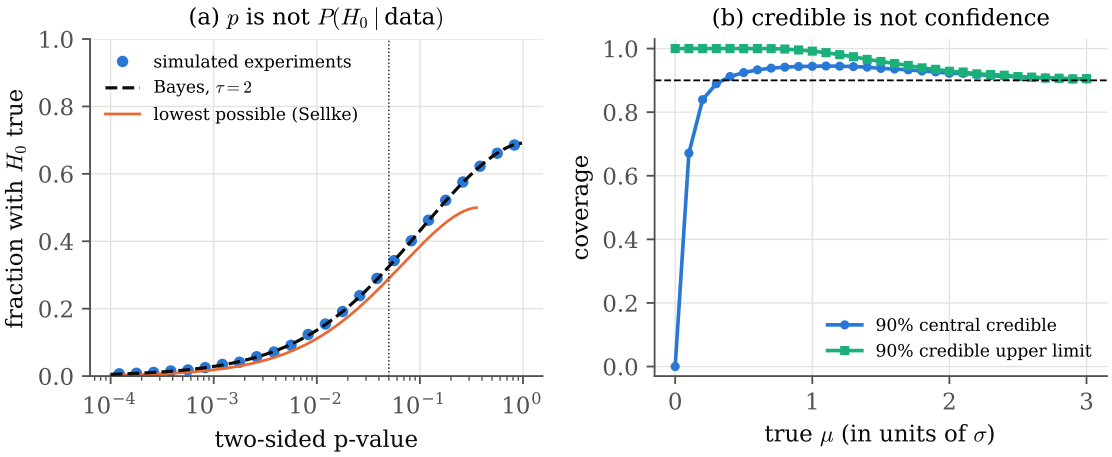


Figure 14.10: (a) Among simulated experiments, half of them true nulls, the fraction of nulls at each p-value. The dots are the simulation, the dashed curve the Bayes factor of (14.13), and the orange curve the lowest value any symmetric, decreasing alternative can produce. At $p = 0.05$ (dotted line) about a third are nulls. (b) How often a 90% credible interval for a non-negative mean contains the truth, as a function of the truth. The central interval never contains $\mu = 0$; the upper limit always does; neither has 90% coverage near the boundary.

interval hold? With a flat prior on $\mu \geq 0$ the posterior is the Gaussian $\mathcal{N}(x, 1)$ cut at zero and renormalised, and the interval's physical part is $[0, 0.145]$:

$$\mathrm{P}(0 \leq \mu \leq x + 1.645 \mid x) \overset{(1)}{=} \frac{\Phi(1.645) - \Phi(-x)}{1 - \Phi(-x)} \overset{(2)}{=} \frac{0.950 - 0.933}{1 - 0.933} \overset{(3)}{=} 0.25. \tag{14.14}$$

(1) The probability of the Gaussian $\mathcal{N}(x, 1)$ between 0 and $x + 1.645$, in units of standard deviations from x, divided by its total probability above zero.
(2) $-x = 1.5$, and $\Phi(1.5) = 0.933$.
(3) The script's value.

So this "90%" interval holds a quarter of the posterior probability. The reverse also fails. The central 90% credible interval, with 5% of the posterior on each side, never reaches down to zero, so if the true mass is zero it covers it in 0.000 of experiments; at $\mu = 0.5$ it covers in 0.924, and only far from the boundary, at $\mu = 3$, does it settle at 0.905. The credible upper limit covers $\mu = 0$ every time (1.000, Figure 14.10b). A further difference is what each depends on. For the same seven heads in twenty-four tosses, Kruschke's one-tailed p-value, and with it the confidence interval, is 0.032 if the plan was to toss twenty-four times and 0.017 if the plan was to stop at the seventh head [Kruschke §§11.1.2–11.1.3, pp. 302–308]; the posterior is the same in both cases, because the likelihoods are proportional (Section 4.2.2).

The lesson. Coverage is a property of the recipe over repeated experiments, credibility a property of the parameter given these data and a prior; check which one the question asks for.

14.5.3 A Fisher forecast is not the posterior

The statement. "The Fisher matrix gives $\sigma = 0.45$, so the experiment will measure the parameter to ± 0.45."

The question it seems to answer. When the data come in, how wide will the posterior be?

The hidden assumption. The Fisher matrix is the average curvature of the log-likelihood at the true parameters (Section 10.1.1). Its inverse is the covariance of the posterior only if the log-likelihood is a quadratic in the parameters, which is exact for a model linear in the

parameters with Gaussian noise and approximately true when the data are many (Section 10.2.3). Away from that limit the posterior can be skewed, wider, or bounded, and the Gaussian of the Fisher matrix cannot show it.

An example where the assumption fails. N unstable particles decay at times t_i drawn from $e^{-t/\tau}/\tau$, and we want the lifetime τ. Write $S = \sum_i t_i$. The log-likelihood and the Fisher information are

$$\ln \mathcal{L}(\tau) \overset{(1)}{=} -N \ln \tau - \frac{S}{\tau}, \qquad I(\tau) \overset{(2)}{=} -\mathbb{E}\Big[\frac{N}{\tau^2} - \frac{2S}{\tau^3}\Big] \overset{(3)}{=} -\frac{N}{\tau^2} + \frac{2N\tau}{\tau^3} \overset{(4)}{=} \frac{N}{\tau^2}, \tag{14.15}$$

(1) The product of N exponential densities, $\tau^{-N} e^{-S/\tau}$, in logarithms.
(2) The Fisher information is minus the expected second derivative; differentiate $-N \ln \tau - S/\tau$ twice.
(3) Each decay time has mean τ, so $\mathbb{E}\, S = N\tau$.
(4) Collect the two terms.

so the forecast is $\sigma_\tau = \tau/\sqrt{N}$. With a flat prior in τ the posterior is the likelihood itself, $p(\tau \mid S) \propto \tau^{-N} e^{-S/\tau}$. That is the inverse-gamma shape $\tau^{-\alpha-1} e^{-\beta/\tau}$ with $\alpha = N - 1$ and $\beta = S$, whose mean is $\beta/(\alpha - 1) = S/(N - 2)$. Take the median experiment, $S = N\tau$ with $\tau = 1$. For $N = 5$ the Fisher interval is $[0.553, 1.447]$; the posterior's central 68% is $[0.847, 2.389]$, with median 1.362 and mean 1.667. The posterior is shifted up, more than twice as wide above the peak as below it, and excludes the short lifetimes that the Fisher interval allows. For $N = 50$ the two are $[0.859, 1.141]$ and $[0.894, 1.189]$: the posterior is approaching its Gaussian limit (Figure 14.11a).

The lesson. The Fisher matrix is the large-data limit of the posterior; whenever the signal is weak, the model nonlinear or a parameter near a boundary, check the forecast with a chain on mock data (Section 10.13).

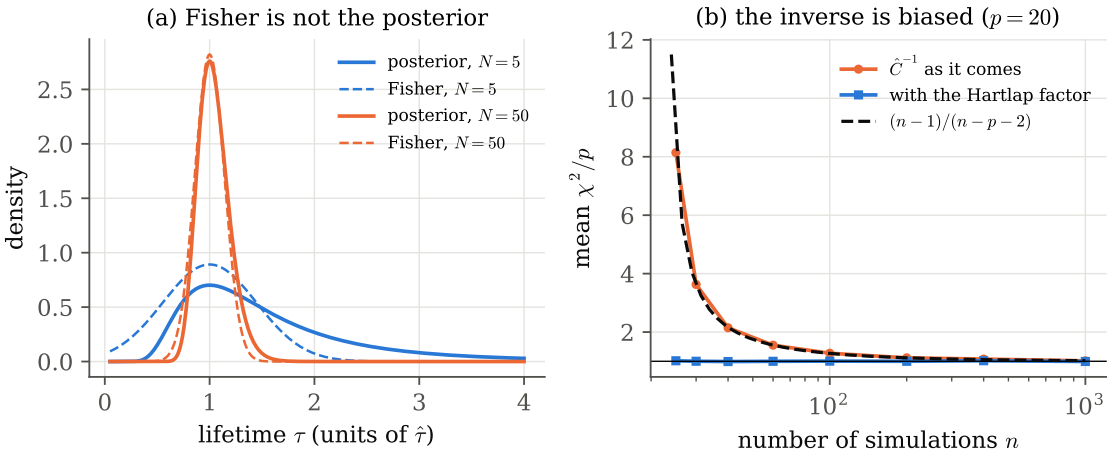


Figure 14.11: (a) Posterior for a lifetime from $N = 5$ and $N = 50$ decays (solid), with the Gaussian of the Fisher forecast (dashed), both for the median experiment. With five decays the posterior is skewed towards long lifetimes and the forecast misplaces it; with fifty the two nearly agree. (b) The mean of $\boldsymbol{r}^{\mathsf{T}} \hat{\mathbf{C}}^{-1} \boldsymbol{r}$ divided by the number of bins $p = 20$, for covariances estimated from n simulations: the raw inverse inflates it by $(n - 1)/(n - p - 2)$ (dashed), and the Hartlap factor removes the inflation.

14.5.4 A chain forgets where it started, not what it was told

The statement. "We ran the chain long enough for the initial choices to wash out."

The question it seems to answer. Does the answer still depend on anything we chose before seeing the data?

The hidden assumption. Two choices are made before the first step: the point where the walker starts and the prior. Only the first is part of the walk. An ergodic chain forgets its start geometrically fast (Section 9.5), but the prior is part of the target density $\mathcal{L} \times p$ that the chain samples (Section 9.14). A longer run only samples the posterior of that prior more precisely.

An example where the assumption fails. N Gaussian readings of unit error have mean $\bar{x} = 1$, the true value. The prior, wrong and confident, is $\mathcal{N}(-2, 0.3^2)$. The posterior is Gaussian (Section 5.8), with mean

$$\mathbb{E}(\mu \mid \boldsymbol{d}) \overset{(1)}{=} \frac{\mu_p/s_p^2 + N\bar{x}/\sigma^2}{1/s_p^2 + N/\sigma^2} \overset{(2)}{=} \frac{-2/0.09 + 10}{1/0.09 + 10} \overset{(3)}{=} -0.579 \quad (N = 10), \tag{14.16}$$

(1) The inverse-variance weighted mean of the prior centre μ_p and the data mean $\bar{x}$ (the Normal–Normal update).
(2) $\mu_p = -2$, $s_p = 0.3$, $\bar{x} = 1$, $\sigma = 1$, $N = 10$.
(3) $-12.2/21.1$.

Three Metropolis chains of 20000 steps, started at -20, 0 and $+20$, give after a burn-in of 1000 steps the means -0.582, -0.577 and -0.579: the start is gone, and all three agree on the wrong answer (Figure 14.12a). More data do what more steps cannot. The data's weight N/σ^2 grows while the prior's $1/s_p^2$ stays fixed, and the posterior mean moves from -1.75 at $N = 1$ through 0.70 at $N = 100$ to 0.967 at $N = 1000$ (Figure 14.12b, and Section 5.13).

The lesson. Steps remove the start, only data remove the prior; diagnose the first with $\hat{R}$ and burn-in, and the second by rerunning with another prior.

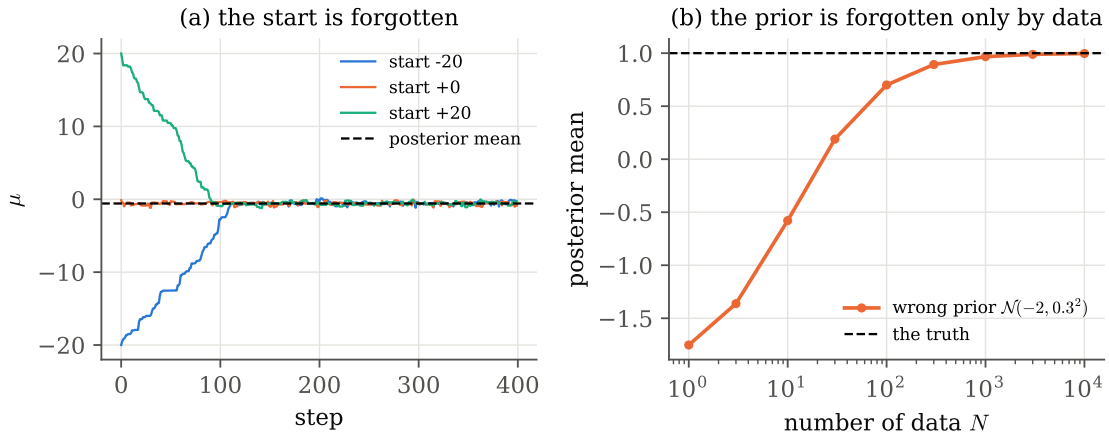


Figure 14.12: (a) The first four hundred steps of three Metropolis chains started far apart, for ten data points and a confident wrong prior: within about a hundred steps all three wander around the same posterior mean (dashed), which is far from the truth $\mu = 1$. (b) The posterior mean with the same prior as the number of data grows: only data pull it to the truth.

14.5.5 The inverse of an estimated covariance is biased

The statement. "The sample covariance from simulations is unbiased, so its inverse gives the right χ^2."

The question it seems to answer. Is the scatter of my data vector properly weighted?

The hidden assumption. That the inverse of an unbiased estimate is an unbiased estimate of the inverse. Inversion is a convex operation, so a fluctuation of a variance towards zero raises its inverse more than an equal fluctuation away from zero lowers it, and the average inverse is too large. For one bin and n simulations the factor is $(n-1)/(n-3)$ (8.76); for p bins it is $(n-1)/(n-p-2)$ (Section 8.13.2).

An example where the assumption fails. Take $p = 20$ Gaussian summaries with unit covariance, estimate the covariance from $n = 30$ simulations, and compute the χ^2 of an independent data vector, 4000 times. The mean should be 20; it is 72.6, against the prediction $20 \times 29/8 = 72.5$. At a nominal 5% threshold of χ^2_{20} the test rejects 0.88 of perfectly good data sets. The Hartlap factor $(n-p-2)/(n-1)$ brings the mean to 20.0, but the false-alarm rate is still 0.134, because the tail of the corrected statistic is the Hotelling law, not χ^2_p (Section 12.24.2). With $n = 100$ the raw rate is 0.225 and the corrected one 0.077 (Figure 14.11b).

The lesson. Correct the inverse, use the Hotelling (or a Student-type) likelihood for the tail, and plan many more simulations than bins (Section 8.13).

14.5.6 Where we looked changes how surprising it is

The statement. "A 3σ excess: a chance of one in seven hundred and forty."

The question it seems to answer. If there is nothing new, how often would we see something this striking?

The hidden assumption. That the excess was the only place we looked. If a bump at any of M masses would have been reported, the null distribution that answers the question is that of the *largest* excess among all M (Section 4.14).

An example where the assumption fails. Let the M search windows be independent, each with local significance $Z_j \sim \mathcal{N}(0,1)$ under the background. The largest is below a threshold c only if every one of them is:

$$p_{\text{glob}} \overset{(1)}{=} 1 - \mathrm{P}\big(\max_j Z_j < c\big) \overset{(2)}{=} 1 - \prod_{j=1}^{M} \mathrm{P}(Z_j < c) \overset{(3)}{=} 1 - (1 - p_{\text{loc}})^M \overset{(4)}{\approx} M p_{\text{loc}}. \tag{14.17}$$

(1) The complement of "no window exceeds c".
(2) Independent windows: the probability that all stay below c is a product.
(3) Each factor is $1 - p_{\text{loc}}$ with $p_{\text{loc}} = \mathrm{P}(Z > c)$.
(4) The binomial expansion to first order, valid while $M p_{\text{loc}} \ll 1$.

With $M = 20$ and $c = 3$, $p_{\text{loc}} = 0.00135$ becomes $p_{\text{glob}} = 0.0267$ (simulated: 0.0266), a global significance of 1.93σ and a trials factor of 19.7. Real windows overlap and are correlated, which lowers the effective number of trials; the scan of Section 4.14.3 counted them with the Gross–Vitells formula. Where the number of places looked at cannot be counted, as for CMB anomalies noticed after the map was made (Section T3.5), no global p-value can be computed at all.

The lesson. The null distribution must describe the whole procedure, including every place we would have been happy to find something.

14.5.7 A shared calibration does not average down

The statement. "With a hundred independent measurements the error is ten times smaller."

The question it seems to answer. How precise is the average?

The hidden assumption. That the errors are independent. When every measurement is multiplied or shifted by the same borrowed number, a zero point, the absolute magnitude of a supernova, the sound horizon of a BAO analysis, that part of the error is the same in all of them (Section 11.20).

An example where the assumption fails. Each of K measurements is $x_i = \theta + c + e_i$, with independent errors e_i of size σ and a common offset c of size σ_c. The average is $\bar{x} = \theta + c + \frac{1}{K}\sum_i e_i$, so

$$\operatorname{Var}\bar{x} \overset{(1)}{=} \operatorname{Var} c + \operatorname{Var}\Big(\frac{1}{K}\sum_{i=1}^{K} e_i\Big) \overset{(2)}{=} \sigma_c^2 + \frac{\sigma^2}{K}. \tag{14.18}$$

(1) The offset and the individual errors are independent, so their variances add.
(2) The variance of a mean of K independent errors is σ^2/K (Section 3.3.1).

With $\sigma = 2\%$ and $\sigma_c = 1\%$, one measurement has an error of 2.24% and a hundred have 1.02% (simulated: 1.02%), not 0.2%. For the Hubble constant from supernovae, an offset δM in the absolute magnitude moves every distance together, $\delta H_0/H_0 = (\ln 10/5)\,\delta M$, and the whole Hubble tension is an offset of 0.18 mag, about 8% in H_0 (Sections 3.17 and 11.20).

The lesson. Find the factor that every measurement shares before averaging; its error is a floor that no amount of data removes.

What the machinery assumes. The seven traps are special cases of a short list of assumptions that run through the whole book. Each holds in the examples where a method was first derived, and each fails somewhere in research. The table names the assumption, the place it breaks, and what the data show when it does.

Taxonomy: assumptions, where they break, and the symptom

assumption	where it breaks	symptom, and the check
the likelihood is Gaussian in the parameters	few counts, low ℓ, weak signals, nonlinear models	skewed chains; Fisher errors too small (Sections 9.19 and 10.13)
the data points are independent	a shared calibration, correlated bins, overlapping windows	errors that stop shrinking; χ^2 per degree of freedom far from one (Section 3.7)
the covariance is known	covariances from simulations or jackknife	inflated χ^2, too many rejections (Sections 8.13.2 and 13.18.2)
one test was made	mass scans, many thresholds, anomalies found by looking	excesses that fade with more data (Section 4.14)
the chain has converged	multiple modes, long correlation times	$\hat{R} > 1$, different answers from different starts (Section 9.13)
the prior is harmless	few data, wide prior ranges in model comparison	answers that move when the prior moves (Sections 5.13 and 5.17.2)
the sample is the population	flux limits, survey masks, fibre collisions	biased distances and clustering (Sections T2.13 and 13.16)
the model is right	any real data	poor goodness of fit; posterior predictive checks fail (Section 4.8)

14.6 The words, and the routes through them

Statistics reuses ordinary words for precise things, and two words that sound alike can belong to opposite directions of reasoning. "Likely" and "probable", "error" and "uncertainty", "confidence" and "credibility" are interchangeable in conversation and not interchangeable in a paper. Every trap measured in Section 14.5 began with one of these swaps. Pinning each word to a plain sentence and to the place where it was built removes the ambiguity. The words fall into the groups the pipeline of Section 14.3 uses: the language of probability, the estimate and its likelihood, tests and intervals, the Bayesian objects, simulation and chains, and the fields, forecasts and summaries of research analyses.

One word needs its two meanings stated before any table, because the others inherit them. A **probability** can be a long-run frequency: the fraction of repeated experiments in which something happens. That is the only reading the frequentist school allows, and with it a parameter, being a fixed constant, has no probability at all [Wasserman p. 175]. Or it can be a degree of belief: how strongly the information we hold supports a statement, including a statement about a constant such as the electron mass [Cowan §1.2, pp. 5–6]. Both obey the same rules of Section 1.2. The forward objects of Figure 14.2 can be read either way; the backward ones need the second reading. In large samples with a well-behaved model the two schools give nearly the same numbers, and in general they need not [Wasserman p. 185], which is why every entry below says which question its object answers.

The language of probability

word	what it means here	built in
conditional probability	the probability after keeping only the cases in which B happened, $\mathrm{P}(A \mid B) = \mathrm{P}(A \cap B)/\mathrm{P}(B)$	Section 1.4
independence	knowing that B happened does not change the probability of A	Section 1.6
random variable	a number attached to each outcome of an experiment	Section 1.8
pmf, pdf	probability of each value of a count; probability per unit length of a continuous number	Section 1.9
marginal distribution	the distribution of one variable with the nuisance variables averaged out of the joint distribution	Section 1.10.2
expectation, variance	the probability-weighted average; the average squared distance from it	Section 1.12
covariance, correlation	how two numbers move together; the same scaled to lie between -1 and 1	Section 1.13
Bernoulli, binomial	a single yes–no trial; the number of successes in many successive trials	Sections 2.1 and 2.2
Poisson	the count of rare, independent events at a constant rate in a fixed window	Section 2.5.1
central limit theorem	a sum of many independent pieces with finite variance is close to Gaussian	Section 2.10
χ^2_ν	the sum of ν squared independent standard Gaussians	Section 2.13
multivariate normal	Gaussian numbers with a covariance matrix; every marginal and conditional is again Gaussian	Section 2.12

The estimate and its likelihood

word	what it means here	built in
estimator	a recipe that turns data into a guess of a parameter; itself a random number	Section 3.1.2
sampling distribution	how the estimator scatters over repeated experiments at a fixed true value	Section 3.1.3
bias, MSE	the average offset of the estimator from the truth; bias squared plus variance	Section 3.2.1
standard deviation, standard error	the scatter of one measurement; the scatter of their mean, smaller by $\sqrt{N}$	Section 3.3.1
likelihood	for a trial $\boldsymbol{\theta}$, the probability of the noise needed to turn the prediction into the data seen; a function of $\boldsymbol{\theta}$, not a density in $\boldsymbol{\theta}$	Sections 3.8 and 3.8.3
maximum-likelihood estimate	the $\boldsymbol{\theta}$ under which the observed data were most probable	Section 3.9
Fisher information	the expected curvature of $-\ln \mathcal{L}$; large curvature means a sharp estimate	Section 3.12.3
Cramér–Rao bound	no unbiased estimator scatters less than the inverse Fisher information	Section 3.13.3
nuisance parameter	a parameter we must model but do not want, removed by profiling or integrating	Sections 3.14 and 4.15.2
profile likelihood	the likelihood of the wanted parameter with the nuisance parameters at their best value for it	Section 3.14
sufficient statistic	a summary that keeps every piece of information the data carry about $\boldsymbol{\theta}$	Section 3.16.3
least squares	maximum likelihood when the noise is Gaussian with known widths	Section 3.10.1
bootstrap	the sampling distribution estimated by resampling the data themselves	Section 3.15

Tests and intervals

word	what it means here	built in
null hypothesis	the hypothesis whose sampling distribution we know, usually "it happened by chance"	Section 4.1.1
p-value	if the null were true, the probability of a result like this one or further in the same direction	Section 4.1.3
significance Z	the p-value restated as the number of Gaussian standard deviations with the same one-sided tail	Section 4.9.1
size, power	the rate of false alarms we accept; the probability of detecting a real effect of given size	Sections 4.3 and 4.3.1
likelihood ratio, Wilks	best fit under the null over best fit overall; $-2\ln$ of it is χ^2 for large samples	Sections 4.7 and 4.7.1
confidence interval	a procedure whose intervals contain the true value in a stated fraction of repetitions	Section 4.10.1
coverage	that fraction, measured by simulation at each true value	Section 4.12
Feldman–Cousins	a confidence belt ordered by likelihood ratio, which moves smoothly from limits to two-sided intervals	Section 4.13.3
look-elsewhere effect	the extra chance of a large fluctuation somewhere when many places were searched	Section 4.14
CL_s	the p-value of signal plus background divided by that of background alone; excludes only signals the experiment could see	Section 4.16.4
Asimov data, Brazil band	the data replaced by their expected values; the band of limits expected from background alone	Sections 4.16.3 and 4.16.5
goodness of fit	whether the best model is an acceptable description at all	Section 4.8

The Bayesian objects

word	what it means here	built in
prior	how probable each parameter value was before these data	Section 5.12
posterior	the prior reweighted by how readily each value could have produced the data, then normalised	Section 5.7
conjugate prior	a prior whose posterior has the same family, so the update is a change of two numbers	Section 5.8
credible interval, HDI	a region holding a stated posterior probability; the narrowest such region	Sections 4.17.1 and 4.17.3
Bernstein–von Mises	with enough data the posterior forgets a smooth prior and becomes the Gaussian of the likelihood	Section 5.13
evidence	the probability a model gave to the data actually seen, averaged over its prior	Section 5.15.1
Bayes factor	the ratio of two evidences: which model predicted the data better	Section 5.15.2
Occam factor	the fraction of a model's prior range that the data allow; the price of flexibility	Section 5.16
Savage–Dickey ratio	for nested models, posterior over prior density at the null value	Section 5.18.1
model averaging	posteriors of several models combined with their posterior probabilities as weights	Section 5.18.2

Simulation and chains

word	what it means here	built in
pseudo-random numbers, seed	a deterministic sequence that passes tests of randomness; the number that fixes it	Section 6.5
Monte Carlo error	the scatter of a simulated average, falling as $1/\sqrt{N}$	Section 6.11.2
importance sampling	drawing from a convenient density and reweighting, with thicker tails than the target	Section 6.12
response matrix, unfolding	the probability that a true bin lands in each measured bin, learned by simulation; its regularised inversion	Section 6.14
stationary distribution	the distribution a Markov chain keeps once it has reached it	Section 9.4
detailed balance	equal probability flow each way between any two states; enough for stationarity	Section 9.4.2
Metropolis–Hastings	a random walk biased by an accept–reject step to spend more time where the probability is higher	Section 9.9.1
burn-in	the early steps that still remember the starting point	Section 9.12.1
autocorrelation time	how many steps the chain needs to produce one independent draw	Section 9.12.2
$\hat{R}$	chains from different starts compared; values above one mean they have not met	Section 9.13

Fields, forecasts and summaries

word	what it means here	built in
random field	a field whose values are random, described by their joint distribution at all points	Section 7.1
homogeneity, isotropy	the statistics do not change from place to place, or with direction	Section 7.1.2
C_ℓ, $\widehat{C}_\ell$	the variance of the harmonic coefficients at multipole ℓ; its estimate from one sky	Section 7.10
cosmic variance	the scatter of $\widehat{C}_\ell$ because only $2\ell+1$ modes exist at each ℓ	Section 7.10.2
pseudo-C_ℓ, MASTER	the spectrum of a masked sky; its mean undone with a mode-coupling matrix	Sections 8.9.3 and 8.11
Hartlap factor	the correction for the bias of the inverse of a covariance estimated from simulations	Section 8.13.2
jackknife	a covariance from the data by leaving out one region at a time	Section 13.18
Fisher forecast	the curvature of the expected log-likelihood at a fiducial model, before any data	Section 10.1.1
noise spectrum, matched filter	the variance of the noise per frequency; the correlation of the data with a template, weighted by it	Sections 10.8.3 and 10.8.6
excursion set, Minkowski functionals	the region above a threshold; its area, boundary length and Euler characteristic	Sections 12.2 and 12.4
Betti numbers, persistence	the numbers of pieces and holes; the thresholds at which each is born and dies	Sections T3.6.2 and 12.9
ABC	inference that keeps the parameter draws whose simulated summary lands near the observed one	Section 12.29

Routes through the book. A glossary answers "what does this word mean?". A new analysis asks a different question: which chapters does it need, and in what order? The pipeline of Figure 14.6 answers by stage. A shorter answer comes from following the four instruments

of Section 14.4 through the book, one route each (Figure 14.13). The routes share their first stops, the rules of probability and the distributions that come from machines, and part ways at the stage where the instrument decides what the noise is. Each route starts from the Python of Chapter T0, which every chapter uses. Its section on numerical trust (Section T0.10) asks a question that every number in this book had to answer before it was printed: is the result a property of the data, or of the rounding, the step size and the random seed that produced it?

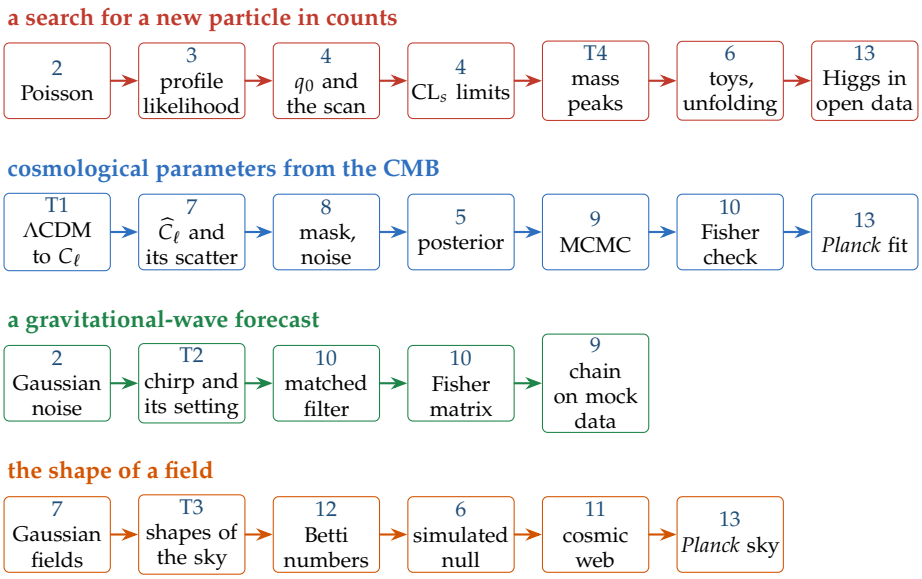


Figure 14.13: Four routes through the book, one for each instrument of Table 14.1. Each box names a chapter and what the route takes from it. All four start after the probability rules of Chapter 1 and the code of Chapter T0; they differ where the instrument fixes the noise, and they meet again in the machinery of tests, posteriors and simulations.

Recap

- The question, said in words before any symbol, fixes the direction: prediction (what can this hypothesis produce?) or inference (which hypotheses could have produced what we saw?). The direction fixes the object: a distribution, a p-value, an interval, a likelihood, a posterior, an evidence or a forecast (Section 14.1). One count of eleven events on a background of four gave eight different, correct numbers.
- The distribution of a number is chosen by the operation that made it; the method by three questions in order: are there data yet, is the sampling distribution known in closed form, and what do we want to know (Section 14.2).
- Every analysis runs the same pipeline: model, data, choice of what matters, summary, sampling distribution, inference. Formulas give the sampling distribution in the simplest cases; simulations through the same code as the data give it in research (Section 14.3).
- The likelihood is always the probability of the noise the data require. For counts it is Poisson and the parameters move the mean; for a Gaussian field the parameters move the covariance and the likelihood is Wishart; for detector noise they move the mean against a known spectrum; for elaborate summaries the distribution is simulated (Section 14.4).
- Seven traps recur: a p-value read as the probability of the null, a confidence interval read as a credible one, a Fisher forecast read as a posterior, a prior confused with a starting point, an inverted simulated covariance, a scan without a look-elsewhere correction, and

a shared calibration averaged as if it were independent (Section 14.5).

Reading the literature

- **The two schools side by side.** Kruschke, *Doing Bayesian Data Analysis*, Chapter 11 (how the stopping and testing intentions change a p-value, confidence against highest-density intervals, what sampling distributions remain good for) and Chapter 12 (estimation against model comparison for a null value). Wasserman, *All of Statistics*, §11.1 and §11.9 [Wasserman pp. 175–176, 185–189], for the postulates of each school and examples where they part.

- **The physicist's view.** Cowan, *Statistical Data Analysis*, Chapter 1 for the two meanings of probability, and the whole book as a compact reference for the frequentist toolkit of colliders.

- **Searches at colliders.** Cowan, Cranmer, Gross and Vitells on the asymptotic formulae for q_0, $\tilde{q}_\mu$ and the Asimov data set [27]; Read and Junk on CL_s [30, 31]; Gross and Vitells on the look-elsewhere effect [15]; Feldman and Cousins on unified intervals [28]; Cranmer's lectures on practical statistics for the LHC [29]; the statistics review of the Particle Data Group [119].

- **Calibrating evidence.** Sellke, Bayarri and Berger on how much a p-value can support an alternative [35]; Kass and Raftery on Bayes factors [33]; Trotta's review of Bayesian methods in cosmology [32].

- **Simulated covariances.** Hartlap, Simon and Schneider on the bias of the inverse [6]; Taylor, Joachimi and Kitching on how many simulations a covariance needs [165]; Sellentin and Heavens on the likelihood that replaces the Gaussian when the covariance is estimated [272].

- **The likelihoods of the four instruments.** Cash on Poisson fits [20]; Hamimeche and Lewis on the CMB likelihood beyond the Gaussian approximation [16]; Cutler and Flanagan on the matched-filter likelihood and Fisher matrix for inspirals [87]; Hivon et al. on pseudo-C_ℓ [168].

Chapter summary

The machinery of the book was assembled one piece at a time. Laid side by side, the pieces answer one practical question, when is each one the right one, and the answer comes in a fixed order. Say the question in words; read off the direction and the object; pick the distribution from the machine that made the number and the method from what is known and what is wanted; place the step in the pipeline; and check that the object was not built to answer a different question. The same ideas ran four very different instruments, and the same seven traps waited in all of them.

concept	operational meaning	built in
the question first	say in words what is wanted; it fixes prediction or inference	Sections 0.2 and 14.1
two directions	forward objects start from an assumed $\boldsymbol{\theta}$; backward objects start from the data and need a prior, except the likelihood	Sections 0.1 and 5.1
one count, eight answers	$n = 11$ on $b = 4$: p-value, Z, $\hat{s}$ and its error, likelihood interval, posterior, B_{10} and its bound, Asimov exposure	Section 14.1
upper limit, CL_s	when nothing is seen, the signal strengths the count excludes, protected against downward fluctuations	Section 4.16
which distribution	the operation that made the number: counted, summed, waited, multiplied, squared, divided by an estimate, or simulated	Sections 2.16 and 14.2.1
ν_{eff}	$2\langle X\rangle^2/\mathrm{Var}\,X$: a scaled χ^2 matched to simulated variance	Sections 8.12 and 14.2.1
which method	no data: forecast; no formula: simulate; then test, estimate, compare or check	Section 14.2.2
estimation vs comparison	a posterior for the effect against a model with the effect exactly zero; they need not agree	Sections 5.18.5 and 14.2.2
the pipeline	model, data, what matters, summary, $P(S \mid \boldsymbol{\theta})$, physics; each chapter builds an arrow	Sections 0.3 and 14.3
formula vs simulation	mean, covariance and shape of a summary: each either derived or measured on simulations	Sections 8.8 and 14.3
the likelihood across instruments	the probability of the noise the data require: Poisson, Wishart, Gaussian with a spectrum, simulated	Sections 0.5 and 14.4
where $\boldsymbol{\theta}$ enters	the mean (counts, waveforms, most summaries) or the covariance (a Gaussian field)	Section 14.4
p vs $\mathrm{P}(H_0 \mid d)$	at $p = 0.05$ about a third of the nulls survive; at least 0.289 for any symmetric alternative	Section 14.5.1
confidence vs credible	a procedure's coverage against a statement about this data set; they differ near boundaries	Section 14.5.2
Fisher vs posterior	Fisher is the Gaussian at the peak; the posterior is wider and skewed with few data	Section 14.5.3
start vs prior	a chain forgets its start; only data overwhelm a wrong prior	Section 14.5.4
inverse covariance	$\hat{\mathbf{C}}^{-1}$ is biased high; Hartlap fixes the mean, Hotelling gives the tail	Section 14.5.5
look-elsewhere	$1-(1-p)^M$ for M independent places; the null must describe the whole search	Section 14.5.6
shared calibration	$\sigma_c^2 + \sigma^2/K$: the common offset is a floor	Section 14.5.7
assumptions	Gaussian likelihood, independence, known covariance, one test, convergence, a harmless prior, a fair sample, a right model: each with its symptom	Section 14.5
routes	the chapters each kind of analysis needs, in order	Section 14.6

Bibliography

[1] T. Vachaspati and A. Lue. "Topological properties of the cosmic microwave background polarization map". In: *Phys. Rev. D* 67 (2003), p. 121302. DOI: 10.1103/PhysRevD.67.121302. arXiv: astro-ph/0304394.

[2] D. Huterer and T. Vachaspati. "Distribution of singularities in the cosmic microwave background polarization". In: *Phys. Rev. D* 72 (2005), p. 043004. DOI: 10.1103/PhysRevD.72.043004. arXiv: astro-ph/0405474.

[3] A. Challinor and G. Chon. "Geometry of weak lensing of CMB polarization". In: *Phys. Rev. D* 66 (2002), p. 127301. DOI: 10.1103/PhysRevD.66.127301. arXiv: astro-ph/0301064.

[4] G. Wilding et al. "Persistent homology of the cosmic web – I. Hierarchical topology in ΛCDM cosmologies". In: *Mon. Not. Roy. Astron. Soc.* 507.2 (2021), pp. 2968–2990. DOI: 10.1093/mnras/stab2326. arXiv: 2011.12851 [astro-ph.CO].

[5] P. Pranav et al. "The Topology of the Cosmic Web in Terms of Persistent Betti Numbers". In: *Mon. Not. Roy. Astron. Soc.* 465.4 (2017), pp. 4281–4310. DOI: 10.1093/mnras/stw2862. arXiv: 1608.04519 [astro-ph.CO].

[6] J. Hartlap, P. Simon, and P. Schneider. "Why your model parameter confidences might be too optimistic: Unbiased estimation of the inverse covariance matrix". In: *Astron. Astrophys.* 464 (2007), p. 399. DOI: 10.1051/0004-6361:20066170. arXiv: astro-ph/0608064.

[7] K. M. Górski et al. "HEALPix - A Framework for high resolution discretization, and fast analysis of data distributed on the sphere". In: *Astrophys. J.* 622 (2005), pp. 759–771. DOI: 10.1086/427976. arXiv: astro-ph/0409513.

[8] A. Lewis, A. Challinor, and A. Lasenby. "Efficient computation of CMB anisotropies in closed FRW models". In: *Astrophys. J.* 538 (2000), pp. 473–476. DOI: 10.1086/309179. arXiv: astro-ph/9911177.

[9] D. Goldberg. "What every computer scientist should know about floating-point arithmetic". In: *ACM Computing Surveys* 23.1 (1991), pp. 5–48. DOI: 10.1145/103162.103163.

[10] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, et al. "Array programming with NumPy". In: *Nature* 585 (2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2. arXiv: 2006.10256.

[11] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, et al. "SciPy 1.0: fundamental algorithms for scientific computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2. arXiv: 1907.10121.

[12] J. D. Hunter. "Matplotlib: A 2D graphics environment". In: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95. DOI: 10.1109/MCSE.2007.55.

[13] A. Zonca et al. "healpy: equal area pixelization and spherical harmonics transforms for data on the sphere in Python". In: *Journal of Open Source Software* 4.35 (2019), p. 1298. DOI: 10.21105/joss.01298.

[14] S. K. Lam, A. Pitrou, and S. Seibert. "Numba: a LLVM-based Python JIT compiler". In: *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*. 2015, pp. 1–6. DOI: 10.1145/2833157.2833162.

[15] E. Gross and O. Vitells. "Trial factors for the look elsewhere effect in high energy physics". In: *Eur. Phys. J. C* 70 (2010), pp. 525–530. DOI: 10.1140/epjc/s10052-010-1470-8. arXiv: 1005.1891 [physics.data-an].

[16] S. Hamimeche and A. Lewis. "Likelihood Analysis of CMB Temperature and Polarization Power Spectra". In: *Phys. Rev. D* 77 (2008), p. 103013. DOI: 10.1103/PhysRevD.77.103013. arXiv: 0801.0554 [astro-ph].

[17] M. Tegmark, A. de Oliveira-Costa, and A. Hamilton. "A high resolution foreground cleaned CMB map from WMAP". In: *Phys. Rev. D* 68 (2003), p. 123523. DOI: 10.1103/PhysRevD.68.123523. arXiv: astro-ph/0302496.

[18] J. Delabrouille et al. "A full sky, low foreground, high resolution CMB map from WMAP". In: *Astron. Astrophys.* 493 (2009), p. 835. DOI: 10.1051/0004-6361:200810514. arXiv: 0807.0773 [astro-ph].

[19] Y. Akrami et al. "Planck 2018 results. IV. Diffuse component separation". In: *Astron. Astrophys.* 641 (2020), A4. DOI: 10.1051/0004-6361/201833881. arXiv: 1807.06208 [astro-ph.CO].

[20] W. Cash. "Parameter estimation in astronomy through application of the likelihood ratio". In: *The Astrophysical Journal* 228 (1979), p. 939. DOI: 10.1086/156922.

[21] S. Baker and R. D. Cousins. "Clarification of the use of CHI-square and likelihood functions in fits to histograms". In: *Nuclear Instruments and Methods in Physics Research* 221 (1984), pp. 437–442. DOI: 10.1016/0167-5087(84)90016-4.

[22] S. S. Wilks. "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses". In: *The Annals of Mathematical Statistics* 9 (1938), pp. 60–62. DOI: 10.1214/aoms/1177732360.

[23] B. Efron. "Bootstrap Methods: Another Look at the Jackknife". In: *The Annals of Statistics* 7.1 (1979), pp. 1–26. DOI: 10.1214/aos/1176344552.

[24] A. G. Riess, W. Yuan, L. M. Macri, et al. "A Comprehensive Measurement of the Local Value of the Hubble Constant with 1 km s^{-1} Mpc^{-1} Uncertainty from the Hubble Space Telescope and the SH0ES Team". In: *The Astrophysical Journal Letters* 934 (2022), p. L7. DOI: 10.3847/2041-8213/ac5c5b. arXiv: 2112.04510.

[25] N. Aghanim et al. "Planck 2018 results. VI. Cosmological parameters". In: *Astron. Astrophys.* 641 (2020). [Erratum: Astron.Astrophys. 652, C4 (2021)], A6. DOI: 10.1051/0004-6361/201833910. arXiv: 1807.06209 [astro-ph.CO].

[26] D. Brout, D. Scolnic, B. Popovic, et al. "The Pantheon+ Analysis: Cosmological Constraints". In: *The Astrophysical Journal* 938 (2022), p. 110. DOI: 10.3847/1538-4357/ac8e04. arXiv: 2202.04077.

[27] G. Cowan, K. Cranmer, E. Gross, and O. Vitells. "Asymptotic formulae for likelihood-based tests of new physics". In: *Eur. Phys. J. C* 71 (2011). [Erratum: Eur.Phys.J.C 73, 2501 (2013)], p. 1554. DOI: 10.1140/epjc/s10052-011-1554-0. arXiv: 1007.1727 [physics.data-an].

[28] G. J. Feldman and R. D. Cousins. "A Unified approach to the classical statistical analysis of small signals". In: *Phys. Rev. D* 57 (1998), pp. 3873–3889. DOI: 10.1103/PhysRevD.57.3873. arXiv: physics/9711021.

[29] K. Cranmer. "Practical Statistics for the LHC". In: *2011 European School of High-Energy Physics*. 2014, pp. 267–308. DOI: 10.5170/CERN-2014-003.267. arXiv: 1503.07622 [physics.data-an].

[30] T. Junk. "Confidence level computation for combining searches with small statistics". In: *Nucl. Instrum. Meth. A* 434 (1999), pp. 435–443. DOI: 10.1016/S0168-9002(99)00498-2. arXiv: hep-ex/9902006.

[31] A. L. Read. "Presentation of search results: the CL_s technique". In: *J. Phys. G* 28.10 (2002), pp. 2693–2704. DOI: 10.1088/0954-3899/28/10/313.

[32] R. Trotta. "Bayes in the sky: Bayesian inference and model selection in cosmology". In: *Contemp. Phys.* 49 (2008), pp. 71–104. DOI: 10.1080/00107510802066753. arXiv: 0803.4089 [astro-ph].

[33] R. E. Kass and A. E. Raftery. "Bayes Factors". In: *Journal of the American Statistical Association* 90.430 (1995), pp. 773–795. DOI: 10.1080/01621459.1995.10476572.

[34] D. V. Lindley. "A Statistical Paradox". In: *Biometrika* 44.1-2 (1957), pp. 187–192. DOI: 10.1093/biomet/44.1-2.187.

[35] T. Sellke, M. J. Bayarri, and J. O. Berger. "Calibration of p Values for Testing Precise Null Hypotheses". In: *The American Statistician* 55.1 (2001), pp. 62–71. DOI: 10.1198/000313001300339950.

[36] N. Metropolis et al. "Equation of State Calculations by Fast Computing Machines". In: *The Journal of Chemical Physics* 21.6 (June 1953), 1087–1092. ISSN: 1089-7690. DOI: 10.1063/1.1699114.

[37] A. Lewis and S. Bridle. "Cosmological parameters from CMB and other data: A Monte Carlo approach". In: *Phys. Rev. D* 66 (2002), p. 103511. DOI: 10.1103/PhysRevD.66.103511. arXiv: astro-ph/0205436.

[38] A. Gelman and D. B. Rubin. "Inference from Iterative Simulation Using Multiple Sequences". In: *Statistical Science* 7.4 (Nov. 1992). ISSN: 0883-4237. DOI: 10.1214/ss/1177011136.

[39] J. Skilling. "Nested sampling for general Bayesian computation". In: *Bayesian Analysis* 1.4 (2006), pp. 833–859. DOI: 10.1214/06-BA127.

[40] L. E. Padilla, L. O. Tellez, L. A. Escamilla, and J. A. Vazquez. "Cosmological Parameter Inference with Bayesian Statistics". In: *Universe* 7.7 (2021), p. 213. DOI: 10.3390/universe7070213. arXiv: 1903.11127 [astro-ph.CO].

[41] R. Takahashi et al. "Revising the Halofit Model for the Nonlinear Matter Power Spectrum". In: *Astrophys. J.* 761 (2012), p. 152. DOI: 10.1088/0004-637X/761/2/152. arXiv: 1208.2701 [astro-ph.CO].

[42] D. J. Eisenstein et al. "Detection of the Baryon Acoustic Peak in the Large-Scale Correlation Function of SDSS Luminous Red Galaxies". In: *Astrophys. J.* 633 (2005), pp. 560–574. DOI: 10.1086/466512. arXiv: astro-ph/0501171.

[43] M. Bartelmann and P. Schneider. "Weak gravitational lensing". In: *Phys. Rept.* 340 (2001), pp. 291–472. DOI: 10.1016/S0370-1573(00)00082-X. arXiv: astro-ph/9912508.

[44] M. Kilbinger. "Cosmology with cosmic shear observations: a review". In: *Rept. Prog. Phys.* 78 (2015), p. 086901. DOI: 10.1088/0034-4885/78/8/086901. arXiv: 1411.0115 [astro-ph.CO].

[45] M. Asgari et al. "KiDS-1000 Cosmology: Cosmic shear constraints and comparison between two point statistics". In: *Astron. Astrophys.* 645 (2021), A104. DOI: 10.1051/0004-6361/202039070. arXiv: 2007.15633 [astro-ph.CO].

[46] N. Kaiser and G. Squires. "Mapping the dark matter with weak gravitational lensing". In: *Astrophys. J.* 404 (1993), pp. 441–450. DOI: 10.1086/172297.

[47] A. Lewis and A. Challinor. "Weak gravitational lensing of the CMB". In: *Phys. Rept.* 429 (2006), pp. 1–65. DOI: 10.1016/j.physrep.2006.03.002. arXiv: astro-ph/0601594.

[48] N. Aghanim et al. "Planck 2018 results. VIII. Gravitational lensing". In: *Astron. Astrophys.* 641 (2020), A8. DOI: 10.1051/0004-6361/201833886. arXiv: 1807.06210 [astro-ph.CO].

[49] C. Hikage et al. "Cosmology from cosmic shear power spectra with Subaru Hyper Suprime-Cam first-year data". In: *Publ. Astron. Soc. Jap.* 71.2 (2019), p. 43. DOI: 10.1093/pasj/psz010. arXiv: 1809.09148 [astro-ph.CO].

[50] A. Amon et al. "Dark Energy Survey Year 3 results: Cosmology from cosmic shear and robustness to data calibration". In: *Phys. Rev. D* 105.2 (2022), p. 023514. DOI: 10.1103/PhysRevD.105.023514. arXiv: 2105.13543 [astro-ph.CO].

[51] R. Takahashi et al. "Full-sky Gravitational Lensing Simulation for Large-area Galaxy Surveys and Cosmic Microwave Background Experiments". In: *Astrophys. J.* 850.1 (2017), p. 24. DOI: 10.3847/1538-4357/aa943d. arXiv: 1706.01472 [astro-ph.CO].

[52] J. M. Kratochvil, Z. Haiman, and M. May. "Probing Cosmology with Weak Lensing Peak Counts". In: *Phys. Rev. D* 81 (2010), p. 043519. DOI: 10.1103/PhysRevD.81.043519. arXiv: 0907.0486 [astro-ph.CO].

[53] J. P. Dietrich and J. Hartlap. "Cosmology with the shear-peak statistics". In: *Mon. Not. Roy. Astron. Soc.* 402 (2010), p. 1049. DOI: 10.1111/j.1365-2966.2009.15948.x. arXiv: 0906.3512 [astro-ph.CO].

[54] N. Jeffrey et al. "Dark Energy Survey Year 3 results: Curved-sky weak lensing mass map reconstruction". In: *Mon. Not. Roy. Astron. Soc.* 505.3 (2021), pp. 4626–4645. DOI: 10.1093/mnras/stab1495. arXiv: 2105.13539 [astro-ph.CO].

[55] H. S. Xavier, F. B. Abdalla, and B. Joachimi. "Improving lognormal models for cosmological fields". In: *Mon. Not. Roy. Astron. Soc.* 459.4 (2016), pp. 3693–3710. DOI: 10.1093/mnras/stw874. arXiv: 1602.08503 [astro-ph.CO].

[56] S. Heydenreich, B. Brück, and J. Harnois-Déraps. "Persistent homology in cosmic shear: constraining parameters with topological data analysis". In: *Astron. Astrophys.* 648 (2021), A74. DOI: 10.1051/0004-6361/202039048. arXiv: 2007.13724 [astro-ph.CO].

[57] W. Hu and T. Okamoto. "Mass reconstruction with CMB polarization". In: *Astrophys. J.* 574 (2002), pp. 566–574. DOI: 10.1086/341110. arXiv: astro-ph/0111606.

[58] M. Kesden, A. Cooray, and M. Kamionkowski. "Lensing reconstruction with CMB temperature and polarization". In: *Phys. Rev. D* 67 (2003), p. 123507. DOI: 10.1103/PhysRevD.67.123507. arXiv: astro-ph/0302536.

[59] D. Hanson, A. Challinor, G. Efstathiou, and P. Bielewicz. "CMB temperature lensing power reconstruction". In: *Phys. Rev. D* 83 (2011), p. 043005. DOI: 10.1103/PhysRevD.83.043005. arXiv: 1008.4403.

[60] M. S. Madhavacheril, K. M. Smith, B. D. Sherwin, and S. Naess. "CMB lensing power spectrum estimation without instrument noise bias". In: *JCAP* 05 (2021), p. 028. DOI: 10.1088/1475-7516/2021/05/028. arXiv: 2011.02475.

[61] T. Namikawa, D. Hanson, and R. Takahashi. "Bias-hardened CMB lensing". In: *Mon. Not. Roy. Astron. Soc.* 431 (2013), pp. 609–620. DOI: 10.1093/mnras/stt195. arXiv: 1209.0091.

[62] Y. Minami and E. Komatsu. "New extraction of the cosmic birefringence from the Planck 2018 polarization data". In: *Phys. Rev. Lett.* 125 (2020), p. 221301. DOI: 10.1103/PhysRevLett.125.221301. arXiv: 2011.11254.

[63] M. Zaldarriaga and U. Seljak. "An all-sky analysis of polarization in the microwave background". In: *Phys. Rev. D* 55 (1997), pp. 1830–1840. DOI: 10.1103/PhysRevD.55.1830. arXiv: astro-ph/9609170.

[64] M. Kamionkowski, A. Kosowsky, and A. Stebbins. "Statistics of cosmic microwave background polarization". In: *Phys. Rev. D* 55 (1997), pp. 7368–7388. DOI: 10.1103/PhysRevD.55.7368. arXiv: astro-ph/9611125.

[65] W. Hu and M. J. White. "A CMB polarization primer". In: *New Astron.* 2 (1997), p. 323. DOI: 10.1016/S1384-1076(97)00022-5. arXiv: astro-ph/9706147.

[66] U. Seljak and M. Zaldarriaga. "Signature of gravity waves in polarization of the microwave background". In: *Phys. Rev. Lett.* 78 (1997), pp. 2054–2057. DOI: 10.1103/PhysRevLett.78.2054. arXiv: astro-ph/9609169.

[67] A. Lue, L.-M. Wang, and M. Kamionkowski. "Cosmological signature of new parity-violating interactions". In: *Phys. Rev. Lett.* 83 (1999), pp. 1506–1509. DOI: 10.1103/PhysRevLett.83.1506. arXiv: astro-ph/9812088.

[68] B. Feng et al. "Searching for CPT violation with cosmic microwave background data from WMAP and BOOMERANG". In: *Phys. Rev. Lett.* 96 (2006), p. 221302. DOI: 10.1103/PhysRevLett.96.221302. arXiv: astro-ph/0601095.

[69] M. Zaldarriaga and U. Seljak. "Gravitational lensing effect on cosmic microwave background polarization". In: *Phys. Rev. D* 58 (1998), p. 023003. DOI: 10.1103/PhysRevD.58.023003. arXiv: astro-ph/9803150.

[70] P. A. R. Ade et al. "Improved constraints on primordial gravitational waves using Planck, WMAP, and BICEP/Keck observations through the 2018 observing season". In: *Phys. Rev. Lett.* 127 (2021), p. 151301. DOI: 10.1103/PhysRevLett.127.151301. arXiv: 2110.00483.

[71] Y. Akrami et al. "Planck 2018 results. XI. Polarized dust foregrounds". In: *Astron. Astrophys.* 641 (2020), A11. DOI: 10.1051/0004-6361/201832618. arXiv: 1801.04945.

[72] S. E. Clark, C.-G. Kim, J. C. Hill, and B. S. Hensley. "The origin of parity violation in polarized dust emission and implications for cosmic birefringence". In: *Astrophys. J.* 919 (2021), p. 53. DOI: 10.3847/1538-4357/ac0e35. arXiv: 2105.00120.

[73] A. Kosowsky and A. Loeb. "Faraday rotation of microwave background polarization by a primordial magnetic field". In: *Astrophys. J.* 469 (1996), pp. 1–6. DOI: 10.1086/177751. arXiv: astro-ph/9601055.

[74] W. Hu, M. M. Hedman, and M. Zaldarriaga. "Benchmark parameters for CMB polarization experiments". In: *Phys. Rev. D* 67 (2003), p. 043004. DOI: 10.1103/PhysRevD.67.043004. arXiv: astro-ph/0210096.

[75] N. Aghanim et al. "Planck 2018 results. V. CMB power spectra and likelihoods". In: *Astron. Astrophys.* 641 (2020), A5. DOI: 10.1051/0004-6361/201936386. arXiv: 1907.12875 [astro-ph.CO].

[76] Y. Minami et al. "Simultaneous determination of the cosmic birefringence and miscalibrated polarization angles from CMB experiments". In: *Prog. Theor. Exp. Phys.* 2019 (2019), 083E02. DOI: 10.1093/ptep/ptz079. arXiv: 1904.12440.

[77] S. M. Carroll, G. B. Field, and R. Jackiw. "Limits on a Lorentz and parity violating modification of electrodynamics". In: *Phys. Rev. D* 41 (1990), pp. 1231–1240. DOI: 10.1103/PhysRevD.41.1231.

[78] S. M. Carroll. "Quintessence and the rest of the world: suppressing long-range interactions". In: *Phys. Rev. Lett.* 81 (1998), pp. 3067–3070. DOI: 10.1103/PhysRevLett.81.3067. arXiv: astro-ph/9806099.

[79] P. Diego-Palazuelos et al. "Cosmic birefringence from the Planck Data Release 4". In: *Phys. Rev. Lett.* 128 (2022), p. 091302. DOI: 10.1103/PhysRevLett.128.091302. arXiv: 2201.07682.

[80] J. R. Eskilt and E. Komatsu. "Improved constraints on cosmic birefringence from the WMAP and Planck cosmic microwave background polarization data". In: *Phys. Rev. D* 106 (2022), p. 063503. DOI: 10.1103/PhysRevD.106.063503. arXiv: 2205.13962.

[81] M. Maggiore. *Gravitational Waves. Vol. 1: Theory and Experiments*. Oxford: Oxford University Press, 2007. DOI: 10.1093/acprof:oso/9780198570745.001.0001.

[82] K. Eda, Y. Itoh, S. Kuroyanagi, and J. Silk. "Gravitational waves as a probe of dark matter minispikes". In: *Phys. Rev. D* 91.4 (2015), p. 044045. DOI: 10.1103/PhysRevD.91.044045. arXiv: 1408.3534 [gr-qc].

[83] B. P. Abbott et al. "A guide to LIGO–Virgo detector noise and extraction of transient gravitational-wave signals". In: *Class. Quant. Grav.* 37.5 (2020), p. 055002. DOI: 10.1088/1361-6382/ab685e. arXiv: 1908.11170 [gr-qc].

[84] L. Barack and C. Cutler. "LISA capture sources: Approximate waveforms, signal-to-noise ratios, and parameter estimation accuracy". In: *Phys. Rev. D* 69 (2004), p. 082005. DOI: 10.1103/PhysRevD.69.082005. arXiv: gr-qc/0310125.

[85] T. Robson, N. J. Cornish, and C. Liu. "The construction and use of LISA sensitivity curves". In: *Class. Quant. Grav.* 36.10 (2019), p. 105011. DOI: 10.1088/1361-6382/ab1101. arXiv: 1803.01944 [astro-ph.HE].

[86] H. Kim, A. Lenoci, I. Stomberg, and X. Xue. "Adiabatically compressed wave dark matter halo and intermediate-mass-ratio inspirals". In: *Phys. Rev. D* 107.8 (2023), p. 083005. DOI: 10.1103/PhysRevD.107.083005. arXiv: 2212.07528 [astro-ph.GA].

[87] C. Cutler and E. E. Flanagan. "Gravitational waves from merging compact binaries: How accurately can one extract the binary's parameters from the inspiral wave form?" In: *Phys. Rev. D* 49 (1994), pp. 2658–2697. DOI: 10.1103/PhysRevD.49.2658. arXiv: gr-qc/9402014.

[88] N. Speeney, A. Antonelli, V. Baibhav, and E. Berti. "Impact of relativistic corrections on the detectability of dark-matter spikes with gravitational waves". In: *Phys. Rev. D* 106.4 (2022), p. 044027. DOI: 10.1103/PhysRevD.106.044027. arXiv: 2204.12508 [gr-qc].

[89] A. Coogan et al. "Measuring the dark matter environments of black hole binaries with gravitational waves". In: *Phys. Rev. D* 105.4 (2022), p. 043009. DOI: 10.1103/PhysRevD.105.043009. arXiv: 2108.04154 [gr-qc].

[90] L. Blanchet. "Gravitational Radiation from Post-Newtonian Sources and Inspiralling Compact Binaries". In: *Living Rev. Rel.* 17 (2014), p. 2. DOI: 10.12942/lrr-2014-2. arXiv: 1310.1528 [gr-qc].

[91] N. Yunes and F. Pretorius. "Fundamental Theoretical Bias in Gravitational Wave Astrophysics and the Parameterized Post-Einsteinian Framework". In: *Phys. Rev. D* 80 (2009), p. 122003. DOI: 10.1103/PhysRevD.80.122003. arXiv: 0909.3328 [gr-qc].

[92] E. Barausse, V. Cardoso, and P. Pani. "Can environmental effects spoil precision gravitational-wave astrophysics?" In: *Phys. Rev. D* 89.10 (2014), p. 104059. DOI: 10.1103/PhysRevD.89.104059. arXiv: 1404.7149 [gr-qc].

[93] J. F. Navarro, C. S. Frenk, and S. D. M. White. "The Structure of cold dark matter halos". In: *Astrophys. J.* 462 (1996), pp. 563–575. DOI: 10.1086/177173. arXiv: astro-ph/9508025.

[94] J. F. Navarro, C. S. Frenk, and S. D. M. White. "A Universal density profile from hierarchical clustering". In: *Astrophys. J.* 490 (1997), pp. 493–508. DOI: 10.1086/304888. arXiv: astro-ph/9611107.

[95] C. A. Correa, J. S. B. Wyithe, J. Schaye, and A. R. Duffy. "The accretion history of dark matter haloes – III. A physical model for the concentration–mass relation". In: *Mon. Not. Roy. Astron. Soc.* 452.2 (2015), pp. 1217–1232. DOI: 10.1093/mnras/stv1363. arXiv: 1502.00391 [astro-ph.CO].

[96] P. Gondolo and J. Silk. "Dark matter annihilation at the galactic center". In: *Phys. Rev. Lett.* 83 (1999), pp. 1719–1722. DOI: 10.1103/PhysRevLett.83.1719. arXiv: astro-ph/9906391.

[97] L. Sadeghian, F. Ferrer, and C. M. Will. "Dark matter distributions around massive black holes: A general relativistic analysis". In: *Phys. Rev. D* 88.6 (2013), p. 063522. DOI: 10.1103/PhysRevD.88.063522. arXiv: 1305.2619 [astro-ph.GA].

[98] L. Hui et al. "Black Hole Hair from Scalar Dark Matter". In: *JCAP* 06 (2019), p. 038. DOI: 10.1088/1475-7516/2019/06/038. arXiv: 1904.12803 [gr-qc].

[99] F. Hancock and H. Witek. "Black-hole hair from vector dark matter accretion". In: *Phys. Rev. D* 112.4 (2025), p. 044033. DOI: 10.1103/v425-4yvq. arXiv: 2506.06554 [gr-qc].

[100] T. F. Chase, D. López Nacir, and N. Yunes. "Gravitational waves of quasicircular, inspiraling black hole binaries in an ultralight vector-dark-matter environment". In: *Phys. Rev. D* 112.10 (2025), p. 103511. DOI: 10.1103/rg4j-8wr9. arXiv: 2505.21383 [astro-ph.CO].

[101] S. Chandrasekhar. "Dynamical Friction. I. General Considerations: the Coefficient of Dynamical Friction". In: *Astrophys. J.* 97 (1943), p. 255. DOI: 10.1086/144517.

[102] L. Hui, J. P. Ostriker, S. Tremaine, and E. Witten. "Ultralight scalars as cosmological dark matter". In: *Phys. Rev. D* 95.4 (2017), p. 043541. DOI: 10.1103/PhysRevD.95.043541. arXiv: 1610.08297 [astro-ph.CO].

[103] D. H. Weinberg et al. "Observational Probes of Cosmic Acceleration". In: *Phys. Rept.* 530 (2013), pp. 87–255. DOI: 10.1016/j.physrep.2013.05.001. arXiv: 1201.2434 [astro-ph.CO].

[104] D. W. Hogg. "Distance measures in cosmology". In: (May 1999). arXiv: astro-ph/9905116.

[105] E. Abdalla et al. "Cosmology intertwined: A review of the particle physics, astrophysics, and cosmology associated with the cosmological tensions and anomalies". In: *JHEAp* 34 (2022), pp. 49–211. DOI: 10.1016/j.jheap.2022.04.002. arXiv: 2203.06142 [astro-ph.CO].

[106] A. G. Adame et al. "DESI 2024 VI: cosmological constraints from the measurements of baryon acoustic oscillations". In: *JCAP* 02 (2025), p. 021. DOI: 10.1088/1475-7516/2025/02/021. arXiv: 2404.03002 [astro-ph.CO].

[107] A. Cooray and R. K. Sheth. "Halo Models of Large Scale Structure". In: *Phys. Rept.* 372 (2002), pp. 1–129. DOI: 10.1016/S0370-1573(02)00276-4. arXiv: astro-ph/0206508.

[108] N. Kaiser. "Clustering in real space and in redshift space". In: *Monthly Notices of the Royal Astronomical Society* 227.1 (July 1987), 1–21. ISSN: 1365-2966. DOI: 10.1093/mnras/227.1.1.

[109] D. J. Eisenstein and W. Hu. "Baryonic features in the matter transfer function". In: *Astrophys. J.* 496 (1998), p. 605. DOI: 10.1086/305424. arXiv: astro-ph/9709112.

[110] J. Schmalzing and K. M. Gorski. "Minkowski functionals used in the morphological analysis of cosmic microwave background anisotropy maps". In: *Mon. Not. Roy. Astron. Soc.* 297 (1998), p. 355. DOI: 10.1046/j.1365-8711.1998.01467.x. arXiv: astro-ph/9710185.

[111] J. M. Bardeen, J. R. Bond, N. Kaiser, and A. S. Szalay. "The statistics of peaks of Gaussian random fields". In: *The Astrophysical Journal* 304 (May 1986), p. 15. ISSN: 1538-4357. DOI: 10.1086/164143.

[112] Y. Akrami et al. "Planck 2018 results. IX. Constraints on primordial non-Gaussianity". In: *Astron. Astrophys.* 641 (2020), A9. DOI: 10.1051/0004-6361/201935891. arXiv: 1905.05697 [astro-ph.CO].

[113] T. Matsubara. "Statistical perturbation theory of cosmic fields. 1. Basic formalism and second order theory". In: (June 2000). arXiv: astro-ph/0006269.

[114] Y. Akrami et al. "Planck 2018 results. VII. Isotropy and Statistics of the CMB". In: *Astron. Astrophys.* 641 (2020), A7. DOI: 10.1051/0004-6361/201935201. arXiv: 1906.02552 [astro-ph.CO].

[115] A. J. S. Hamilton, I. Gott J. Richard, and D. Weinberg. "The topology of the large-scale structure of the universe". In: *The Astrophysical Journal* 309 (Oct. 1986), p. 1. ISSN: 1538-4357. DOI: 10.1086/164571.

[116] M. Biagetti, A. Cole, and G. Shiu. "The Persistence of Large Scale Structures I: Primordial non-Gaussianity". In: *JCAP* 04 (2021), p. 061. DOI: 10.1088/1475-7516/2021/04/061. arXiv: 2009.04819 [astro-ph.CO].

[117] G. Aad et al. "The ATLAS Experiment at the CERN Large Hadron Collider". In: *JINST* 3 (2008), S08003. DOI: 10.1088/1748-0221/3/08/S08003.

[118] G. Aad et al. "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". In: *Phys. Lett. B* 716 (2012), pp. 1–29. DOI: 10.1016/j.physletb.2012.08.020. arXiv: 1207.7214 [hep-ex].

[119] S. Navas et al. "Review of particle physics". In: *Phys. Rev. D* 110.3 (2024), p. 030001. DOI: 10.1103/PhysRevD.110.030001.

[120] M. Aaboud et al. "Measurement of the Inelastic Proton-Proton Cross Section at $\sqrt{s} = 13$ TeV with the ATLAS Detector at the LHC". In: *Phys. Rev. Lett.* 117.18 (2016), p. 182002. DOI: 10.1103/PhysRevLett.117.182002. arXiv: 1606.02625 [hep-ex].

[121] ATLAS Collaboration. *Review of the 13 TeV ATLAS Open Data release*. Tech. rep. ATL-OREACH-PUB-2020-001. CERN, 2020.

[122] L. Evans and P. Bryant. "LHC Machine". In: *JINST* 3 (2008). Ed. by L. Evans and P. Bryant, S08001. DOI: 10.1088/1748-0221/3/08/S08001.

[123] L. Serkin. "The release of the 13 TeV ATLAS Open Data: using open education resources effectively". In: *EPJ Web Conf.* 245 (2020), p. 08026. DOI: 10.1051/epjconf/202024508026.

[124] S. Chatrchyan et al. "Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC". In: *Phys. Lett. B* 716 (2012), pp. 30–61. DOI: 10.1016/j.physletb.2012.08.021. arXiv: 1207.7235 [hep-ex].

[125] K. Cranmer et al. *HistFactory: A tool for creating statistical models for use with RooFit and RooStats*. Tech. rep. CERN-OPEN-2012-016. CERN, 2012.

[126] B. Efron and C. Morris. "Data Analysis Using Stein's Estimator and its Generalizations". In: *Journal of the American Statistical Association* 70.350 (1975), pp. 311–319. DOI: 10.1080/01621459.1975.10479864.

[127] B. Efron and C. Morris. "Stein's Paradox in Statistics". In: *Scientific American* 236.5 (1977), pp. 119–127. DOI: 10.1038/scientificamerican0577-119.

[128] W. James and C. Stein. "Estimation with Quadratic Loss". In: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1 (1961); reprinted in Breakthroughs in Statistics, Springer (1992), pp. 443–460*. University of California Press, 1961, pp. 361–379. DOI: 10.1007/978-1-4612-0919-5_30.

[129] C. Stein. "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution". In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. University of California Press, 1956, pp. 197–206. DOI: 10.1525/9780520313880-018.

[130] B. Efron and C. Morris. "Stein's Estimation Rule and its Competitors—An Empirical Bayes Approach". In: *Journal of the American Statistical Association* 68.341 (1973), pp. 117–130. DOI: 10.1080/01621459.1973.10481350.

[131] F. R. Hampel. "The Influence Curve and its Role in Robust Estimation". In: *Journal of the American Statistical Association* 69.346 (1974), pp. 383–393. DOI: 10.1080/01621459.1974.10482962.

[132] P. J. Huber. "Robust Estimation of a Location Parameter". In: *The Annals of Mathematical Statistics* 35.1 (1964), pp. 73–101. DOI: 10.1214/aoms/1177703732.

[133] M. Tegmark. "How to Measure CMB Power Spectra without Losing Information". In: *Physical Review D* 55.10 (1997), pp. 5895–5907. DOI: 10.1103/PhysRevD.55.5895. arXiv: astro-ph/9611174.

[134] G. Efstathiou. "Myths and truths concerning estimation of power spectra". In: *Mon. Not. Roy. Astron. Soc.* 349 (2004), p. 603. DOI: 10.1111/j.1365-2966.2004.07530.x. arXiv: astro-ph/0307515.

[135] W. H. Richardson. "Bayesian-Based Iterative Method of Image Restoration". In: *Journal of the Optical Society of America* 62.1 (1972), pp. 55–59. DOI: 10.1364/JOSA.62.000055.

[136] L. B. Lucy. "An Iterative Technique for the Rectification of Observed Distributions". In: *The Astronomical Journal* 79 (1974), p. 745. DOI: 10.1086/111605.

[137] P. W. Holland and R. E. Welsch. "Robust Regression Using Iteratively Reweighted Least-Squares". In: *Communications in Statistics – Theory and Methods* 6.9 (1977), pp. 813–827. DOI: 10.1080/03610927708827533.

[138] P. J. Green. "Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and some Robust and Resistant Alternatives". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 46.2 (1984), pp. 149–170. DOI: 10.1111/j.2517-6161.1984.tb01288.x.

[139] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum Likelihood from Incomplete Data via the EM Algorithm". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1 (1977), pp. 1–22. DOI: 10.1111/j.2517-6161.1977.tb01600.x.

[140] C. F. J. Wu. "On the Convergence Properties of the EM Algorithm". In: *The Annals of Statistics* 11.1 (1983), pp. 95–103. DOI: 10.1214/aos/1176346060.

[141] G. D'Agostini. "A multidimensional unfolding method based on Bayes' theorem". In: *Nuclear Instruments and Methods in Physics Research Section A* 362.2–3 (1995), pp. 487–498. DOI: 10.1016/0168-9002(95)00274-X.

[142] M. H. Quenouille. "Notes on Bias in Estimation". In: *Biometrika* 43.3–4 (1956), pp. 353–360. DOI: 10.1093/biomet/43.3-4.353.

[143] J. W. Tukey. "Bias and Confidence in Not-quite Large Samples (abstract)". In: *The Annals of Mathematical Statistics* 29.2 (1958), p. 614. DOI: 10.1214/aoms/1177706647.

[144] P. Hall and M. A. Martin. "On Bootstrap Resampling and Iteration". In: *Biometrika* 75.4 (1988), pp. 661–671. DOI: 10.1093/biomet/75.4.661.

[145] C. M. Hirata and U. Seljak. "Analyzing Weak Lensing of the Cosmic Microwave Background Using the Likelihood Function". In: *Physical Review D* 67 (2003), p. 043001. DOI: 10.1103/PhysRevD.67.043001. eprint: astro-ph/0209489.

[146] C. M. Hirata and U. Seljak. "Reconstruction of Lensing from the Cosmic Microwave Background Polarization". In: *Physical Review D* 68 (2003), p. 083002. DOI: 10.1103/PhysRevD.68.083002. eprint: astro-ph/0306354.

[147] J. Carron and A. Lewis. "Maximum a Posteriori CMB Lensing Reconstruction". In: *Physical Review D* 96 (2017), p. 063510. DOI: 10.1103/PhysRevD.96.063510.

[148] R. E. Kalman. "A New Approach to Linear Filtering and Prediction Problems". In: *Journal of Basic Engineering* 82.1 (1960), pp. 35–45. DOI: 10.1115/1.3662552.

[149] B. P. Welford. "Note on a Method for Calculating Corrected Sums of Squares and Products". In: *Technometrics* 4.3 (1962), pp. 419–420. DOI: 10.1080/00401706.1962.10490022.

[150] P. L'Ecuyer. "Efficient and portable combined random number generators". In: *Communications of the ACM* 31.6 (1988), pp. 742–751. DOI: 10.1145/62959.62969.

[151] G. Marsaglia. "RANDOM NUMBERS FALL MAINLY IN THE PLANES". In: *Proceedings of the National Academy of Sciences* 61.1 (Sept. 1968), 25–28. ISSN: 1091-6490. DOI: 10.1073/pnas.61.1.25.

[152] P. L'Ecuyer. "Tables of linear congruential generators of different sizes and good lattice structure". In: *Mathematics of Computation* 68.225 (1999), pp. 249–260. DOI: 10.1090/S0025-5718-99-00996-5.

[153] G. Marsaglia. "Xorshift RNGs". In: *Journal of Statistical Software* 8.14 (2003), pp. 1–6. DOI: 10.18637/jss.v008.i14.

[154] M. Matsumoto and T. Nishimura. "Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator". In: *ACM Transactions on Modeling and Computer Simulation* 8.1 (Jan. 1998), 3–30. ISSN: 1558-1195. DOI: 10.1145/272991.272995.

[155] P. L'Ecuyer and R. Simard. "TestU01: A C library for empirical testing of random number generators". In: *ACM Transactions on Mathematical Software* 33.4 (2007), 22:1–22:40. DOI: 10.1145/1268776.1268777.

[156] M. E. O'Neill. *PCG: A Family of Simple Fast Space-Efficient Statistically Good Algorithms for Random Number Generation*. Tech. rep. HMC-CS-2014-0905. Harvey Mudd College, Computer Science Department, 2014.

[157] J. K. Salmon, M. A. Moraes, R. O. Dror, and D. E. Shaw. "Parallel random numbers: as easy as 1, 2, 3". In: *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis (SC11)*. 2011, 16:1–16:12. DOI: 10.1145/2063384.2063405.

[158] I. M. Sobol'. "On the distribution of points in a cube and the approximate evaluation of integrals". In: *USSR Computational Mathematics and Mathematical Physics* 7.4 (1967), pp. 86–112. DOI: 10.1016/0041-5553(67)90144-9.

[159] G. E. P. Box and M. E. Muller. "A Note on the Generation of Random Normal Deviates". In: *The Annals of Mathematical Statistics* 29.2 (1958), pp. 610–611. DOI: 10.1214/aoms/1177706645.

[160] G. Marsaglia and T. A. Bray. "A Convenient Method for Generating Normal Variables". In: *SIAM Review* 6.3 (1964), pp. 260–264. DOI: 10.1137/1006063.

[161] C. Bierlich et al. "A comprehensive guide to the physics and usage of PYTHIA 8.3". In: *SciPost Physics Codebases* (2022), p. 8. DOI: 10.21468/SciPostPhysCodeb.8. arXiv: 2203.11601.

[162] S. Agostinelli et al. "Geant4—a simulation toolkit". In: *Nuclear Instruments and Methods in Physics Research Section A* 506.3 (2003), pp. 250–303. DOI: 10.1016/S0168-9002(03)01368-8.

[163] A. Höcker and V. Kartvelishvili. "SVD approach to data unfolding". In: *Nuclear Instruments and Methods in Physics Research Section A* 372.3 (1996), pp. 469–481. DOI: 10.1016/0168-9002(95)01478-0. arXiv: hep-ph/9509307.

[164] J. A. Peacock. *Cosmological Physics*. Cambridge: Cambridge University Press, 1999. DOI: 10.1017/CBO9780511804533.

[165] A. Taylor, B. Joachimi, and T. Kitching. "Putting the Precision in Precision Cosmology: How accurate should your data covariance matrix be?" In: *Mon. Not. Roy. Astron. Soc.* 432 (2013), p. 1928. DOI: 10.1093/mnras/stt270. arXiv: 1212.4359 [astro-ph.CO].

[166] L. Knox. "Determination of inflationary observables by cosmic microwave background anisotropy experiments". In: *Phys. Rev. D* 52 (1995), pp. 4307–4318. DOI: 10.1103/PhysRevD.52.4307. arXiv: astro-ph/9504054.

[167] U. Seljak and M. Zaldarriaga. "A Line of sight integration approach to cosmic microwave background anisotropies". In: *Astrophys. J.* 469 (1996), pp. 437–444. DOI: 10.1086/177793. arXiv: astro-ph/9603033.

[168] E. Hivon et al. "Master of the cosmic microwave background anisotropy power spectrum: a fast method for statistical analysis of large and complex cosmic microwave background data sets". In: *Astrophys. J.* 567 (2002), p. 2. DOI: 10.1086/338126. arXiv: astro-ph/0105302.

[169] D. Alonso, J. Sanchez, and A. Slosar. "A unified pseudo-C_ℓ framework". In: *Mon. Not. Roy. Astron. Soc.* 484.3 (2019), pp. 4127–4151. DOI: 10.1093/mnras/stz093. arXiv: 1809.09603 [astro-ph.CO].

[170] W. K. Hastings. "Monte Carlo sampling methods using Markov chains and their applications". In: *Biometrika* 57.1 (Apr. 1970), 97–109. ISSN: 0006-3444. DOI: 10.1093/biomet/57.1.97.

[171] J. Dunkley et al. "Fast and reliable mcmc for cosmological parameter estimation". In: *Mon. Not. Roy. Astron. Soc.* 356 (2005), pp. 925–936. DOI: 10.1111/j.1365-2966.2004.08464.x. arXiv: astro-ph/0405462.

[172] P. H. Peskun. "Optimum Monte-Carlo sampling using Markov chains". In: *Biometrika* 60.3 (1973), pp. 607–612. DOI: 10.1093/biomet/60.3.607.

[173] G. O. Roberts, A. Gelman, and W. R. Gilks. "Weak convergence and optimal scaling of random walk Metropolis algorithms". In: *The Annals of Applied Probability* 7.1 (1997), pp. 110–120. DOI: 10.1214/aoap/1034625254.

[174] D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman. "emcee: The MCMC Hammer". In: *Publications of the Astronomical Society of the Pacific* 125 (2013), pp. 306–312. DOI: 10.1086/670067. arXiv: 1202.3665.

[175] A. Sokal. "Monte Carlo Methods in Statistical Mechanics: Foundations and New Algorithms". In: *Functional Integration*. Vol. 361. NATO ASI Series. Springer, 1997, pp. 131–192. DOI: 10.1007/978-1-4899-0319-8_6.

[176] S. P. Brooks and A. Gelman. "General Methods for Monitoring Convergence of Iterative Simulations". In: *Journal of Computational and Graphical Statistics* 7.4 (1998), pp. 434–455. DOI: 10.1080/10618600.1998.10474787.

[177] H. A. Kramers. "Brownian motion in a field of force and the diffusion model of chemical reactions". In: *Physica* 7.4 (1940), pp. 284–304. DOI: 10.1016/S0031-8914(40)90098-2.

[178] L. Onsager. "Crystal Statistics. I. A Two-Dimensional Model with an Order-Disorder Transition". In: *Physical Review* 65.3-4 (1944), pp. 117–149. DOI: 10.1103/PhysRev.65.117.

[179] C. N. Yang. "The Spontaneous Magnetization of a Two-Dimensional Ising Model". In: *Physical Review* 85.5 (1952), pp. 808–816. DOI: 10.1103/PhysRev.85.808.

[180] S. Geman and D. Geman. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-6.6 (1984), pp. 721–741. DOI: 10.1109/TPAMI.1984.4767596.

[181] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. "Hybrid Monte Carlo". In: *Physics Letters B* 195.2 (1987), pp. 216–222. DOI: 10.1016/0370-2693(87)91197-X.

[182] R. M. Neal. "MCMC using Hamiltonian dynamics". In: *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, 2011. arXiv: 1206.1901.

[183] J. Goodman and J. Weare. "Ensemble samplers with affine invariance". In: *Communications in Applied Mathematics and Computational Science* 5.1 (2010), pp. 65–80. DOI: 10.2140/camcos.2010.5.65.

[184] L. Verde, H. V. Peiris, D. N. Spergel, et al. "First Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Parameter Estimation Methodology". In: *The Astrophysical Journal Supplement Series* 148 (2003), pp. 195–211. DOI: 10.1086/377335. arXiv: astro-ph/0302218.

[185] M. Chu et al. "Cosmological Parameter Constraints as Derived from the Wilkinson Microwave Anisotropy Probe Data via Gibbs Sampling and the Blackwell-Rao Estimator". In: *Physical Review D* 71 (2005), p. 103002. DOI: 10.1103/PhysRevD.71.103002. arXiv: astro-ph/0411737.

[186] N. Aghanim et al. "Planck 2018 results. I. Overview and the cosmological legacy of Planck". In: *Astron. Astrophys.* 641 (2020), A1. DOI: 10.1051/0004-6361/201833880. arXiv: 1807.06205 [astro-ph.CO].

[187] A. Spurio Mancini et al. "COSMOPOWER: emulating cosmological power spectra for accelerated Bayesian inference from next-generation surveys". In: *Monthly Notices of the Royal Astronomical Society* 511.2 (2022), pp. 1771–1788. DOI: 10.1093/mnras/stac064. arXiv: 2106.03846.

[188] A. Lewis. "GetDist: a Python package for analysing Monte Carlo samples". In: *arXiv e-prints* (2019). arXiv: 1910.13970.

[189] J. Torrado and A. Lewis. "Cobaya: Code for Bayesian Analysis of hierarchical physical models". In: *Journal of Cosmology and Astroparticle Physics* 2021.05 (2021), p. 057. DOI: 10.1088/1475-7516/2021/05/057. arXiv: 2005.05290.

[190] B. Audren, J. Lesgourgues, K. Benabed, and S. Prunet. "Conservative Constraints on Early Cosmology: an illustration of the Monte Python cosmological parameter inference code". In: *Journal of Cosmology and Astroparticle Physics* 2013.02 (2013), p. 001. DOI: 10.1088/1475-7516/2013/02/001. arXiv: 1210.7183.

[191] J. R. Bond, A. H. Jaffe, and L. Knox. "Radical Compression of Cosmic Microwave Background Data". In: *The Astrophysical Journal* 533 (2000), pp. 19–37. DOI: 10.1086/308625. arXiv: astro-ph/9808264.

[192] M. Tegmark, A. Taylor, and A. Heavens. "Karhunen-Loeve eigenvalue problems in cosmology: How should we tackle large data sets?" In: *Astrophys. J.* 480 (1997), p. 22. DOI: 10.1086/303939. arXiv: astro-ph/9603021.

[193] M. Vallisneri. "Use and abuse of the Fisher information matrix in the assessment of gravitational-wave parameter-estimation prospects". In: *Phys. Rev. D* 77 (2008), p. 042001. DOI: 10.1103/PhysRevD.77.042001. arXiv: gr-qc/0703086.

[194] D. Coe. "Fisher Matrices and Confidence Ellipses: A Quick-Start Guide and Software". In: (June 2009). arXiv: 0906.4123 [astro-ph.IM].

[195] A. Heavens. "Statistical techniques in cosmology". In: (June 2009). arXiv: 0906.0664 [astro-ph.CO].

[196] G. Jungman, M. Kamionkowski, A. Kosowsky, and D. N. Spergel. "Cosmological-parameter determination with microwave background maps". In: *Phys. Rev. D* 54 (1996), p. 1332. DOI: 10.1103/PhysRevD.54.1332. arXiv: astro-ph/9512139.

[197] L. Verde. "Statistical methods in cosmology". In: *Lect. Notes Phys.* 800 (2010), pp. 147–177. DOI: 10.1007/978-3-642-10598-2_4. arXiv: 0911.3105 [astro-ph.CO].

[198] W. J. Percival, W. Sutherland, J. A. Peacock, et al. "Parameter constraints for flat cosmologies from CMB and 2dFGRS power spectra". In: *Mon. Not. Roy. Astron. Soc.* 337 (2002), p. 1068. DOI: 10.1046/j.1365-8711.2002.06001.x. arXiv: astro-ph/0206256.

[199] L. S. Finn. "Detection, measurement, and gravitational radiation". In: *Phys. Rev. D* 46 (1992), p. 5236. arXiv: gr-qc/9209010.

[200] B. Allen et al. "FINDCHIRP: an algorithm for detection of gravitational waves from inspiraling compact binaries". In: *Phys. Rev. D* 85 (2012), p. 122006. arXiv: gr-qc/0509116.

[201] C. Röver, R. Meyer, and N. Christensen. "Modelling coloured residual noise in gravitational-wave signal processing". In: *Class. Quantum Grav.* 28 (2011), p. 015010. arXiv: 0804.3853.

[202] N. J. Cornish and T. B. Littenberg. "BayesWave: Bayesian inference for gravitational wave bursts and instrument glitches". In: *Class. Quantum Grav.* 32 (2015), p. 135012. arXiv: 1410.3835.

[203] T. B. Littenberg and N. J. Cornish. "Prototype global analysis of LISA data with multiple source types". In: *Phys. Rev. D* 107 (2023), p. 063004. arXiv: 2301.03673.

[204] J. S. Speagle. "dynesty: a dynamic nested sampling package for estimating Bayesian posteriors and evidences". In: *Mon. Not. Roy. Astron. Soc.* 493.3 (2020), pp. 3132–3158. DOI: 10.1093/mnras/staa278. arXiv: 1904.02180 [astro-ph.IM].

[205] N. Karnesis et al. "Eryn: a multipurpose sampler for Bayesian inference". In: *Mon. Not. Roy. Astron. Soc.* 526.4 (2023), pp. 4814–4830. DOI: 10.1093/mnras/stad2939. arXiv: 2303.02164 [astro-ph.IM].

[206] M. Tinto and S. V. Dhurandhar. "Time-delay interferometry". In: *Living Rev. Relativ.* 8 (2005), p. 4. arXiv: gr-qc/0409034.

[207] Q. Baghi. "The LISA Data Challenges". In: *56th Rencontres de Moriond on Gravitation*. Apr. 2022. arXiv: 2204.12142 [gr-qc].

[208] M. L. Katz et al. "FastEMRIWaveforms: New tools for millihertz gravitational-wave data analysis". In: *Phys. Rev. D* 104 (2021), p. 064047. arXiv: 2104.04582.

[209] B. J. Kavanagh, D. A. Nichols, G. Bertone, and D. Gaggero. "Detecting dark matter around black holes with gravitational waves: Effects of dark-matter dynamics on the gravitational waveform". In: *Phys. Rev. D* 102.8 (2020), p. 083006. DOI: 10.1103/PhysRevD.102.083006. arXiv: 2002.12811 [gr-qc].

[210] C. Cutler and M. Vallisneri. "LISA detections of massive black hole inspirals: parameter extraction errors due to inaccurate template waveforms". In: *Phys. Rev. D* 76 (2007), p. 104018. arXiv: 0707.2982.

[211] S. A. Usman et al. "The PyCBC search for gravitational waves from compact binary coalescence". In: *Class. Quant. Grav.* 33.21 (2016), p. 215004. DOI: 10.1088/0264-9381/33/21/215004. arXiv: 1508.02357 [gr-qc].

[212] G. Ashton et al. "BILBY: A user-friendly Bayesian inference library for gravitational-wave astronomy". In: *Astrophys. J. Suppl.* 241.2 (2019), p. 27. DOI: 10.3847/1538-4365/ab06fc. arXiv: 1811.02042 [astro-ph.IM].

[213] B. Zackay, L. Dai, and T. Venumadhav. "Relative binning and fast likelihood evaluation for gravitational wave parameter estimation". In: (2018). arXiv: 1806.08792.

[214] Planck Collaboration, N. Aghanim, et al. "Planck intermediate results. XLIX. Parity-violation constraints from polarization data". In: *Astron. Astrophys.* 596 (2016), A110. DOI: 10.1051/0004-6361/201629018. arXiv: 1605.08633.

[215] K. M. Smith. "Pseudo-C_ℓ estimators which do not mix E and B modes". In: *Phys. Rev. D* 74 (2006), p. 083002. DOI: 10.1103/PhysRevD.74.083002. arXiv: astro-ph/0511629.

[216] P. Amaro-Seoane et al. "Laser Interferometer Space Antenna". In: (Feb. 2017). arXiv: 1702.00786 [astro-ph.IM].

[217] P. J. E. Peebles. *The Large-Scale Structure of the Universe*. Princeton: Princeton University Press, 1980. ISBN: 978-0-691-08240-0.

[218] P. J. E. Peebles and M. G. Hauser. "Statistical analysis of catalogs of extragalactic objects. III. The Shane–Wirtanen and Zwicky catalogs". In: *The Astrophysical Journal Supplement Series* 28 (1974), p. 19. DOI: 10.1086/190308.

[219] M. Davis and P. J. E. Peebles. "A survey of galaxy redshifts. V. The two-point position and velocity correlations". In: *The Astrophysical Journal* 267 (1983), p. 465. DOI: 10.1086/160884.

[220] A. J. S. Hamilton. "Toward better ways to measure the galaxy correlation function". In: *The Astrophysical Journal* 417 (1993), p. 19. DOI: 10.1086/173288.

[221] S. D. Landy and A. S. Szalay. "Bias and variance of angular correlation functions". In: *The Astrophysical Journal* 412 (1993), p. 64. DOI: 10.1086/172900.

[222] M. Kerscher, I. Szapudi, and A. S. Szalay. "A comparison of estimators for the two-point correlation function". In: *The Astrophysical Journal Letters* 535 (2000), pp. L13–L16. DOI: 10.1086/312702. arXiv: astro-ph/9912088.

[223] H. A. Feldman, N. Kaiser, and J. A. Peacock. "Power-spectrum analysis of three-dimensional redshift surveys". In: *The Astrophysical Journal* 426 (1994), p. 23. DOI: 10.1086/174036. arXiv: astro-ph/9304022.

[224] S. Alam et al. "The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: cosmological analysis of the DR12 galaxy sample". In: *Monthly Notices of the Royal Astronomical Society* 470 (2017), pp. 2617–2652. DOI: 10.1093/mnras/stx721. arXiv: 1607.03155.

[225] W. H. Press and P. Schechter. "Formation of Galaxies and Clusters of Galaxies by Self-Similar Gravitational Condensation". In: *The Astrophysical Journal* 187 (1974), p. 425. DOI: 10.1086/152650.

[226] J. R. Bond, S. Cole, G. Efstathiou, and N. Kaiser. "Excursion set mass functions for hierarchical Gaussian fluctuations". In: *The Astrophysical Journal* 379 (1991), p. 440. DOI: 10.1086/170520.

[227] R. K. Sheth and G. Tormen. "Large-scale bias and the peak background split". In: *Monthly Notices of the Royal Astronomical Society* 308 (1999), p. 119. DOI: 10.1046/j.1365-8711.1999.02692.x. arXiv: astro-ph/9901122.

[228] R. K. Sheth, H. J. Mo, and G. Tormen. "Ellipsoidal collapse and an improved model for the number and spatial distribution of dark matter haloes". In: *Monthly Notices of the Royal Astronomical Society* 323 (2001), p. 1. DOI: 10.1046/j.1365-8711.2001.04006.x. arXiv: astro-ph/9907024.

[229] H. J. Mo and S. D. M. White. "An analytic model for the spatial clustering of dark matter haloes". In: *Monthly Notices of the Royal Astronomical Society* 282 (1996), p. 347. DOI: 10.1093/mnras/282.2.347. arXiv: astro-ph/9512127.

[230] U. Seljak. "Analytic model for galaxy and dark matter clustering". In: *Monthly Notices of the Royal Astronomical Society* 318 (2000), p. 203. DOI: 10.1046/j.1365-8711.2000.03715.x. arXiv: astro-ph/0001493.

[231] A. Mead et al. "An accurate halo model for fitting non-linear cosmological power spectra and baryonic feedback models". In: *Monthly Notices of the Royal Astronomical Society* 454 (2015), p. 1958. DOI: 10.1093/mnras/stv2036. arXiv: 1505.07833.

[232] Z. Zheng, A. A. Berlind, D. H. Weinberg, A. J. Benson, et al. "Theoretical Models of the Halo Occupation Distribution: Separating Central and Satellite Galaxies". In: *The Astrophysical Journal* 633 (2005), p. 791. DOI: 10.1086/466510. arXiv: astro-ph/0408564.

[233] J. R. Bond, L. Kofman, and D. Pogosyan. "How filaments are woven into the cosmic web". In: *Nature* 380 (1996), pp. 603–606. DOI: 10.1038/380603a0. arXiv: astro-ph/9512141.

[234] E. Platen, R. van de Weygaert, and B. J. T. Jones. "A cosmic watershed: the WVF void detection technique". In: *Monthly Notices of the Royal Astronomical Society* 380 (2007), p. 551. DOI: 10.1111/j.1365-2966.2007.12125.x. arXiv: 0706.2788.

[235] M. C. Neyrinck. "ZOBOV: a parameter-free void-finding algorithm". In: *Monthly Notices of the Royal Astronomical Society* 386 (2008), p. 2101. DOI: 10.1111/j.1365-2966.2008.13180.x. arXiv: 0712.3049.

[236] E. Jennings, Y. Li, and W. Hu. "The abundance of voids and the excursion set formalism". In: *Monthly Notices of the Royal Astronomical Society* 434 (2013), p. 2167. DOI: 10.1093/mnras/stt1169. arXiv: 1304.6087.

[237] S. Contarini, A. Pisani, N. Hamaus, F. Marulli, et al. "Cosmological Constraints from the BOSS DR12 Void Size Function". In: *The Astrophysical Journal* 953 (2023), p. 46. DOI: 10.3847/1538-4357/acde54. arXiv: 2212.03873.

[238] N. Hamaus, P. M. Sutter, and B. D. Wandelt. "Universal Density Profile for Cosmic Voids". In: *Physical Review Letters* 112 (2014), p. 251302. DOI: 10.1103/PhysRevLett.112.251302. arXiv: 1403.5499.

[239] G. Lavaux and B. D. Wandelt. "Precision cosmography with stacked voids". In: *The Astrophysical Journal* 754 (2012), p. 109. DOI: 10.1088/0004-637X/754/2/109. arXiv: 1110.0345.

[240] N. Hamaus, A. Pisani, J.-A. Choi, G. Lavaux, et al. "Precision cosmology with voids in the final BOSS data". In: *Journal of Cosmology and Astroparticle Physics* 2020.12 (2020), p. 023. DOI: 10.1088/1475-7516/2020/12/023. arXiv: 2007.07895.

[241] E. Hubble. "A relation between distance and radial velocity among extra-galactic nebulae". In: *Proceedings of the National Academy of Sciences* 15 (1929), pp. 168–173. DOI: 10.1073/pnas.15.3.168.

[242] E. Hubble and M. L. Humason. "The Velocity-Distance Relation among Extra-Galactic Nebulae". In: *The Astrophysical Journal* 74 (1931), p. 43. DOI: 10.1086/143323.

[243] W. Baade. "The Period-Luminosity Relation of the Cepheids". In: *Publications of the Astronomical Society of the Pacific* 68 (1956), p. 5. DOI: 10.1086/126870.

[244] A. Sandage. "Current Problems in the Extragalactic Distance Scale". In: *The Astrophysical Journal* 127 (1958), p. 513. DOI: 10.1086/146483.

[245] P. J. E. Peebles. *Principles of Physical Cosmology*. Princeton: Princeton University Press, 1993. ISBN: 978-0-691-01933-8. DOI: 10.2307/j.ctvxrpxvb.

[246] K. Górski. "On the pattern of perturbations of the Hubble flow". In: *The Astrophysical Journal Letters* 332 (1988), p. L7. DOI: 10.1086/185255.

[247] P. L. Schechter. "Mass-to-light ratios for elliptical galaxies". In: *The Astronomical Journal* 85 (1980), p. 801. DOI: 10.1086/112742.

[248] D. Lynden-Bell et al. "Spectroscopy and photometry of elliptical galaxies. V. Galaxy streaming toward the new supergalactic center". In: *The Astrophysical Journal* 326 (1988), p. 19. DOI: 10.1086/166066.

[249] R. Wojtak et al. "Cosmic variance of the local Hubble flow in large-scale cosmological simulations". In: *Monthly Notices of the Royal Astronomical Society* 438 (2014), pp. 1805–1812. DOI: 10.1093/mnras/stt2321. arXiv: 1312.0276.

[250] V. Marra, L. Amendola, I. Sawicki, and W. Valkenburg. "Cosmic Variance and the Measurement of the Local Hubble Parameter". In: *Physical Review Letters* 110 (2013), p. 241305. DOI: 10.1103/PhysRevLett.110.241305. arXiv: 1303.3121.

[251] M. A. Strauss and J. A. Willick. "The density and peculiar velocity fields of nearby galaxies". In: *Physics Reports* 261 (1995), pp. 271–431. DOI: 10.1016/0370-1573(95)00013-7. arXiv: astro-ph/9502079.

[252] R. Watkins and H. A. Feldman. "An unbiased estimator of peculiar velocity with Gaussian distributed errors for precision cosmology". In: *Monthly Notices of the Royal Astronomical Society* 450 (2015), pp. 1868–1873. DOI: 10.1093/mnras/stv651.

[253] R. B. Tully et al. "Cosmicflows-4". In: *The Astrophysical Journal* 944 (2023), p. 94. DOI: 10.3847/1538-4357/ac94d8. arXiv: 2209.11238.

[254] R. Jimenez and A. Loeb. "Constraining Cosmological Parameters Based on Relative Galaxy Ages". In: *The Astrophysical Journal* 573 (2002), pp. 37–42. DOI: 10.1086/340549. arXiv: astro-ph/0106145.

[255] M. Moresco et al. "Unveiling the Universe with emerging cosmological probes". In: *Living Reviews in Relativity* 25 (2022), p. 6. DOI: 10.1007/s41114-022-00040-z. arXiv: 2201.07241.

[256] M. Moresco et al. "A 6% measurement of the Hubble parameter at $z \sim 0.45$: direct evidence of the epoch of cosmic re-acceleration". In: *Journal of Cosmology and Astroparticle Physics* 2016.05 (2016), p. 014. DOI: 10.1088/1475-7516/2016/05/014. arXiv: 1601.01701.

[257] M. Moresco et al. "Setting the Stage for Cosmic Chronometers. II. Impact of Stellar Population Synthesis Models Systematics and Full Covariance Matrix". In: *The Astrophysical Journal* 898 (2020), p. 82. DOI: 10.3847/1538-4357/ab9eb0. arXiv: 2003.07362.

[258] J. Carrick, S. J. Turnbull, G. Lavaux, and M. J. Hudson. "Cosmological parameters from the comparison of peculiar velocities with predictions from the 2M++ density field". In: *Monthly Notices of the Royal Astronomical Society* 450 (2015), pp. 317–332. DOI: 10.1093/mnras/stv547. arXiv: 1504.04627.

[259] E. R. Peterson, W. D. Kenworthy, D. Scolnic, A. G. Riess, et al. "The Pantheon+ Analysis: Evaluating Peculiar Velocity Corrections in Cosmological Analyses with Nearby Type Ia Supernovae". In: *The Astrophysical Journal* 938 (2022), p. 112. DOI: 10.3847/1538-4357/ac4698. arXiv: 2110.03487.

[260] H.-J. Seo and D. J. Eisenstein. "Probing dark energy with baryonic acoustic oscillations from future large galaxy redshift surveys". In: *The Astrophysical Journal* 598 (2003), pp. 720–740. DOI: 10.1086/379122. arXiv: astro-ph/0307460.

[261] W. J. Percival and M. White. "Testing cosmological structure formation using redshift-space distortions". In: *Monthly Notices of the Royal Astronomical Society* 393 (2009), pp. 297–308. DOI: 10.1111/j.1365-2966.2008.14211.x. arXiv: 0808.0003.

[262] L. Anderson et al. "The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: baryon acoustic oscillations in the Data Release 9 spectroscopic galaxy sample". In: *Monthly Notices of the Royal Astronomical Society* 427 (2012), pp. 3435–3467. DOI: 10.1111/j.1365-2966.2012.22066.x. arXiv: 1203.6594.

[263] F. Bernardeau, S. Colombi, E. Gaztanaga, and R. Scoccimarro. "Large scale structure of the universe and cosmological perturbation theory". In: *Phys. Rept.* 367 (2002), pp. 1–248. DOI: 10.1016/S0370-1573(02)00135-7. arXiv: astro-ph/0112551.

[264] S. O. Rice. "Mathematical Analysis of Random Noise". In: *Bell System Technical Journal* 23.3 (1944), pp. 282–332. DOI: 10.1002/j.1538-7305.1944.tb00874.x.

[265] R. J. Adler and J. E. Taylor. *Random Fields and Geometry*. Springer Monographs in Mathematics. New York: Springer, 2007. DOI: 10.1007/978-0-387-48116-6.

[266] R. J. Adler. *The Geometry of Random Fields*. Wiley, Chichester (1981); reprinted SIAM, Philadelphia, 2010. DOI: 10.1137/1.9780898718980.

[267] H. Tomita. "Curvature Invariants of Random Interface Generated by Gaussian Fields". In: *Progress of Theoretical Physics* 76.4 (1986), pp. 952–955. DOI: 10.1143/PTP.76.952.

[268] Edelsbrunner, Letscher, and Zomorodian. "Topological Persistence and Simplification". In: *Discrete & Computational Geometry* 28.4 (Nov. 2002), 511–533. ISSN: 1432-0444. DOI: 10.1007/s00454-002-2885-2.

[269] N. Otter et al. "A roadmap for the computation of persistent homology". In: *EPJ Data Science* 6 (2015). arXiv: {1506.08903} (cs.CG).

[270] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. "Stability of Persistence Diagrams". In: *Discrete & Computational Geometry* 37.1 (Dec. 2006), 103–120. ISSN: 1432-0444. DOI: 10.1007/s00454-006-1276-5.

[271] A. Cole and G. Shiu. "Persistent Homology and Non-Gaussianity". In: *JCAP* 03 (2018), p. 025. DOI: 10.1088/1475-7516/2018/03/025. arXiv: 1712.08159.

[272] E. Sellentin and A. F. Heavens. "Parameter inference with estimated covariance matrices". In: *Mon. Not. Roy. Astron. Soc.* 456 (2016), pp. L132–L136. DOI: 10.1093/mnrasl/slv190. arXiv: 1511.05969.

[273] J. Carron. "On the assumption of Gaussianity for cosmological two-point statistics and parameter dependent covariance matrices". In: *Astron. Astrophys.* 551 (2013), A88. DOI: 10.1051/0004-6361/201220538. arXiv: 1204.4724.

[274] J. Alsing, T. Charnock, S. Feeney, and B. Wandelt. "Fast likelihood-free cosmology with neural density estimators and active learning". In: *Mon. Not. Roy. Astron. Soc.* 488.3 (2019), pp. 4440–4458. DOI: 10.1093/mnras/stz1960. arXiv: 1903.00007 [astro-ph.CO].

[275] J. Akeret et al. "Approximate Bayesian computation for forward modeling in cosmology". In: *JCAP* 08 (2015), p. 043. DOI: 10.1088/1475-7516/2015/08/043. arXiv: 1504.07245.

[276] B. Reid, S. Ho, N. Padmanabhan, W. J. Percival, et al. "SDSS-III Baryon Oscillation Spectroscopic Survey Data Release 12: galaxy target selection and large-scale structure catalogues". In: *Mon. Not. R. Astron. Soc.* 455 (2016), pp. 1553–1573. DOI: 10.1093/mnras/stv2382. arXiv: 1509.06529.

[277] A. J. Ross, W. J. Percival, A. G. Sánchez, et al. "The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: analysis of potential systematics". In: *Mon. Not. R. Astron. Soc.* 424 (2012), pp. 564–590. DOI: 10.1111/j.1365-2966.2012.21235.x. arXiv: 1203.6499.

[278] A. J. Cuesta, M. Vargas-Magaña, F. Beutler, et al. "The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: baryon acoustic oscillations in the correlation function of LOWZ and CMASS galaxies in Data Release 12". In: *Mon. Not. R. Astron. Soc.* 457 (2016), pp. 1770–1785. DOI: 10.1093/mnras/stw066. arXiv: 1509.06371.

[279] B. Efron and C. Stein. "The jackknife estimate of variance". In: *Ann. Statist.* 9.3 (1981), pp. 586–596. DOI: 10.1214/aos/1176345462.

[280] P. Norberg, C. M. Baugh, E. Gaztañaga, and D. J. Croton. "Statistical analysis of galaxy surveys – I. Robust error estimation for two-point clustering statistics". In: *Mon. Not. R. Astron. Soc.* 396 (2009), pp. 19–38. DOI: 10.1111/j.1365-2966.2009.14389.x. arXiv: 0810.1885.

[281] S. Dodelson and M. D. Schneider. "The Effect of Covariance Estimator Error on Cosmological Parameter Constraints". In: *Phys. Rev. D* 88 (2013), p. 063537. DOI: 10.1103/PhysRevD.88.063537. arXiv: 1304.2593.